Análisis, validación y estudio poblacional de las inversiones entre dos genomas humanos

Tesis doctoral

David Vicente Salvador

Universitat Autònoma de Barcelona
Facultat de Biociències
Departament de Genètica i Microbiología
Bellaterra, 2014

Esta memoria de tesis es presentada por el Licenciado en Biología David Vicente Salvador para optar al título de Doctor.

David Vicente Salvador

El Doctor Mario Cáceres Aguilar, investigador ICREA del Institut de Biotecnologia i Biomedicina de la Universitat Autònoma de Barcelona,

CERTIFICA que David Vicente Salvador ha realizado la tesis doctoral titulada "Análisis, validación y estudio poblacional de las inversiones entre dos genomas humanos" bajo su dirección, al igual que el trabajo de investigación realizado en el Institut de Biotecnologia i Biomedicina de la Universitat Autònoma de Barcelona que se muestra en ella.

Para que quede constancia, a día 29 de Septiembre de 2014 firma el presente certificado en Bellaterra.

Dr. Mario Cáceres Aguilar

ÍNDICE

Resumen / Abstract ____ **7**

1. INTRODUCCIÓN ____ **11**

1.1 La importancia de la variación genética **13**
1.2 Variación estructural **15**
1.2.1 ¿Que es la variación estructural? **15**
1.2.2 Historia y métodos de detección **20**
1.2.3 Origen y mecanismos de formación **39**
1.3 Variación estructural interespecífica **53**
1.3.1 CNVs **53**
1.3.2 Inversiones **55**
1.4 Variación estructural intraespecífica humana **58**
1.4.1 CNVs **59**
1.4.2 Inversiones polimórficas **67**
1.5 ¿Cuantas inversiones polimórficas hay en el genoma humano? **77**
1.5.1 Bases de datos y redundancia **78**
1.5.2 Espectro de variación detectada **79**
1.5.3 Validación experimental y genotipación **80**
1.6 Objetivos **88**

2. MATERIALES Y MÉTODOS ____ **89**

2.1 Obtención de la secuencia, alineamiento, definición de los puntos de rotura y anotación de las inversiones **91**
2.2 Soporte por mapeo de extremos apareados (PEM) de fósmidos **94**
2.3 Muestras de ADN **96**
2.4 Diseño de cebadores, validación por PCR y PCR inversa (iPCR) y genotipación experimental **97**
2.5 Secuenciación **100**
2.6 Genotipación bioinformática in silico **101**

2.7 Análisis de la variación nucleotídica **103**
2.8 Análisis de la frecuencia y distribución poblacional **106**

3. RESULTADOS **109**

3.1 Análisis, validación y estudio poblacional de las inversiones entre dos genomas humanos **111**
3.2 Participación en otros estudios **173**

4. DISCUSIÓN **175**

4.1 Análisis de las diferencias en la definición de los puntos de rotura **177**
4.2 La repetitividad del genoma humano como fuente de errores **180**
4.3 Las características de los puntos de rotura dividen las inversiones **183**
4.3.1 Mecanismos de formación homólogos vs. no homólogos **187**
4.3.2 La variación estructural comparte los puntos de rotura: ecología genómica **188**
4.3.3 Implicaciones del origen recurrente de las inversiones **190**
4.4 La frecuencia de las inversiones indica diferencias entre poblaciones **194**
4.4.1 Diferentes factores afectan las frecuencias obtenidas de los distintos métodos de genotipación **194**
4.4.2 Explicación de la distribución poblacional actual de las inversiones a través de sus posibles efectos funcionales **197**

5. CONCLUSIONES **213**

Bibliografía **217**

Agradecimientos **231**

RESUMEN

Las inversiones fueron las primeras variantes estructurales detectadas y asociadas a efectos fenotípicos en varias especies. Sin embargo, la dificultad de su estudio las ha llevado a ser las peor caracterizadas en genomas complejos como el humano. En los últimos años, se ha predicho un elevado número de posibles inversiones polimórficas en humanos mediante técnicas a gran escala como el mapeo de extremos apareados de fósmidos o la secuenciación de genomas completos, pero pocas de estas predicciones han sido validadas o estudiadas en detalle. En este trabajo se han investigado las 90 inversiones que provienen de la comparación de dos genomas ensamblados de forma independiente: el genoma de Referencia *HG18* y el de J. Craig Venter (*HuRef*). El análisis detallado de su secuencia ha demostrado que 31 (34.4%) son errores en la comparación de ambos genomas. A continuación se han analizado experimentalmente 46 de las 59 regiones candidatas restantes (51.1%) mediante *PCR* y *PCR* inversa en el ADN de *HuRef* y un panel de 9 individuos de HapMap de origen Africano, Asiático y Europeo. De éstas, 18 han resultado contener inversiones polimórficas reales y 30 son errores de ensamblaje en uno de los genomas (25 errores en *HG18* y 5 en *HuRef*). Estos errores se han confirmado experimentalmente amplificando la región en los clones *BAC* del genoma de Referencia o en el ADN de *HuRef*, respectivamente. De esta manera se ha podido eliminar un gran número de predicciones falsas y se ha contribuido a definir un catálogo fiable de inversiones polimórficas en el genoma humano. Además, 17 de las inversiones validadas se han genotipado en 90 individuos de HapMap de origen Europeo y en dos especies de primates y 7 con puntos de rotura sencillos se han genotipado *in silico* en 1092 individuos de 14 poblaciones del proyecto de los 1000 Genomas, a través de la detección de secuencias que contienen los puntos de rotura. Los genotipos nos han permitido encontrar *SNP* marcador, establecer las frecuencias del alelo invertido en diferentes poblaciones y la orientación ancestral. Mediante el análisis de la variación nucleotídica y haplotípica se ha podido determinar también el origen único o recurrente de las inversiones en la población Europea, y se han encontrado tres inversiones que habrían ocurrido en haplotipos diferentes. El análisis de secuencia de los puntos de rotura nos ha permitido determinar su mecanismo de formación e identificar genes cuya expresión podría verse afectada. Como resultado, se ha visto que las inversiones forman dos grupos según las características de sus puntos de rotura: las inversiones con puntos de rotura no localizados en repeticiones invertidas (RIs) generalmente tienen un tamaño menor y están formadas por mecanismos no homólogos que determinan su origen único, mientras que las inversiones con puntos de rotura en RIs tienen un tamaño mayor y están formadas por mecanismos homólogos que determinan su posible origen recurrente. Por otra parte, las inversiones analizadas destacan por localizarse fuera de regiones codificantes, en regiones intergénicas o dentro de intrones, aunque algunas invierten

parcial o completamente genes duplicados. Finalmente, se han clasificado las inversiones según su posible implicación adaptativa mediante el análisis de las diferencias de frecuencia en las poblaciones, el estado ancestral, el índice de estructuración de la población *Fst*, y los posibles efectos sobre genes determinados por la posición de los puntos de rotura. En general no se esperan efectos drásticos de sus puntos de rotura, aunque las inversiones *HsInv0006* y *HsInv0030* son candidatas a tener efectos sobre los genes *DSTYK* y *CTRB2/CTRB1*, respectivamente, y en el caso de la inversión *HsInv0006* su distribución poblacional sugiere posibles efectos adaptativos.

ABSTRACT

Inversions were the first type of structural variants to be detected and associated to phenotypic effects in different species. However, studying them was difficult and they have become one of the less characterized variants in complex genomes such as the human. In the last years, a great number of putative polymorphic inversions have been predicted in humans by high-throughput techniques, like fosmid paired-end mapping or whole genome sequencing, but few of them have been validated or studied in detail. In this work we have investigated the 90 inversions coming from the comparison of two independently assembled genomes, the Reference genome *HG18* and J. Craig genome (*HuRef*). By analysing in detail its sequence, we have shown that 31 (34.4%) regions are errors in the comparison of both genomes. Next,we have experimentally analyzed 46 out of the 59 remaining candidate regions (51.1%) by PCR and inverse PCR using DNA from HuRef and 9 HapMap individuals from Africa, Asia and Europe. Of those, 18 have resulted to be polymorphic inversions and 30 are assembly errors of one of the genomes (25 HG18 errors and 5 HuRef errors). These errors have been experimentally confirmed by amplifying the region in the BAC clones from the Reference genome or in *HuRef's* DNA, respectively. Thus, we have been able to eliminate a high number of false predictions and contributed to the definition of a reliable catalog of polymorphic inversions in the human genome. In addition, for 17 of the validated inversions we have genotyped 90 European HapMap individuals and two primate species, and for 7 inversions with simple breakpoints we have also genotyped *in silico* 1092 individuals from 14 populations from the 1000 Genomes Project through the detection of sequences that contain the breakpoints. Genotypes have been used to find tag SNPs, establish the frequency of the inverted allele in the different populations and the ancestral orientation. Trough the analysis of the nucleotide and haplotype variation it has also been possible to determine the unique or recurrent origin of inversions, and three inversions generated in different haplotypes have been found. The analysis of the breakpoint sequence have allowed us to determine its formation mechanism and to identify genes whose expression could be affected. As a result, inversions can be classified in two groups depending on its breakpoint characteristics: inversions with simple breakpoints not located in inverted repeats are smaller and are formed by non-homologous mechanisms that imply an unique origin, whereas inversions with breakpoints located in IRs are longer and are formed by homologous mechanisms that are related to their potential recurrent origin. On the other hand, the analyzed inversions tend to be located out of coding regions, in intergenic regions or within introns, although few of them invert partially or completely duplicated genes. Finally, inversions have been classified according to its possible adaptive effects based on the analysis of frequency differences among populations, ancestral state, the *Fst* population structure index, and possible effects over genes determined by breakpoint

position. The inversions that we have analyzed are located out of coding regions, in intergenic regions or into introns, although few of them partial or completely invert duplicated genes. In general no drastic effects of its breakpoints are expected, but inversions *HsInv0006* and *HsInv0030* are candidates to affect the *DSTYK* and *CTRB2/CTRB1* genes, respectively, and in the case of inversion *HsInv0006*, its population distribution suggests possible adaptive effects.

1. INTRODUCCIÓN

1. INTRODUCCIÓN

1.1 La importancia de la variación genética

Los genes fueron tempranamente definidos por *Gregor Mendel*, un monje naturalista austríaco que publicó sus trabajos sobre la herencia de los caracteres de la planta del guisante, definiendo los factores responsables de la transmisión de esos caracteres a través de distintas generaciones. Actualmente la definición de gen hace referencia a la región de secuencia genómica que corresponde a una unidad de herencia, asociada con regiones reguladoras, regiones transcritas y otras regiones funcionales [Pearson et al. 2006]. El concepto va ligado a las diferentes formas en que se presentan los genes, los alelos. Conocemos como genotipo el conjunto de alelos que corresponden a un determinado gen y aunque un organismo diploide solo pueda tener dos alelos para ese gen, uno por cromosoma, en una población pueden existir diferentes alelos. El conjunto de alelos para un gen determinado en una población determina la variación alélica y por extensión define la variación genética. Es pues, la variación genética, el conjunto de formas distintas de uno o más genes. El genoma engloba el conjunto de genes de un organismo, y por lo tanto, podemos hablar de variación en el genoma, para referirnos a la variación genética de todos los genes de un organismo. Dependiendo de si nos referimos a la variación genética de una población o de la comparación de varias poblaciones, hablamos de variación genética intrapoblacional o interpoblacional, y lo mismo ocurre con la variación intraespecífica e interespecífica.

Mendel pudo definir los factores heredables responsables de los caracteres que observaba en las plantas de guisante a través de los estudios de entrecruzamiento que realizó. Estos caracteres observables se denominan fenotipo. Más allá de la rugosidad y el color de las vainas de guisante, el fenotipo engloba otras características moleculares no observables. No están sólo determinados por el genotipo sino que el ambiente interactúa con las variables genéticas. Hoy en día, podemos decir que el modelo escogido por *Mendel*, fue uno de los modelos de relación genotipo-fenotipo más sencillos posibles, con efecto ambiental negligible. Lamentablemente, la gran mayoría de caracteres de interés en la especie humana, como pueden ser los que definen las diferencias poblacionales, la susceptibilidad a determinadas enfermedades o incluso la respuesta personal que tenemos a los fármacos diseñados para tratarlas, tienen una relación entre genotipo y fenotipo mucho más compleja que la planta del guisante. Es por eso que el avance en el conocimiento de esta relación es uno de los objetivos principales de la investigación médica y biológica, y el camino hacia ese conocimiento parte del estudio de la variación en el genoma humano.

El genoma muestra diferentes tipos de variación que engloba desde los cambios de una sola base, *SNPs*, (del inglés Single Nucleotide Polymorphism), denominados variación nucleotídica, hasta la variación más grande, visible a través del microscopio, como los cambios en el número de cromosomas enteros denominados aneuploidias o las reorganizaciones de su estructura, pasando por la variación de tamaño medio que no es observable a través del microscopio (variación estructural). Se han hallado asociaciones importantes entre la variación genética mayoritariamente nucleotídica aunque también estructural, y algunas enfermedades y caracteres fenotípicos, pero estamos lejos de comprender realmente la contribución genética al fenotipo [The 1000 genomes project consortium, 2010]. Hasta ahora ha habido una focalización en el estudio de la variación nucleotídica, una buena muestra de ello es que los dos proyectos más importantes de estudio de la variación genética humana como son "The international HapMap project" [Gibbs et al. 2003] y "The 1000 Genomes Project" [The 1000 Genomes Project Consortium 2012] se han centrado en la variación nucleotídica. La búsqueda de *SNPs* causantes de enfermedades a partir del análisis de su ligamiento con *SNPs* comunes en individuos afectados, ha sido exitosa para enfermedades producidas por un solo gen, pero no lo ha sido para las enfermedades más comunes y complejas, como la enfermedad cardiovascular, el cáncer, la obesidad, la diabetes, y enfermedades psiquiátricas o inflamatorias, donde cada variante contribuye hasta cierto punto en el riesgo a padecer la enfermedad [Gibbs et al. 2003]. Además se han desarrollado estrategias indirectas en las que se comparan *SNPs* de un grupo de individuos afectados por una enfermedad en concreto con los *SNPs* de un grupo sano control. Este tipo de estudios analizan las variantes diferentes entre ambos grupos para encontrar las que tienen un papel en el riesgo a sufrir la enfermedad, por eso se denominan estudios de asociación a nivel del genoma completo, o *GWAS* (del inglés Genome-wide association studies) [Gibbs et al. 2003].

A pesar de esta focalización en los estudios de la variación genética nucleotídica, los resultados no han sido tan provechosos como se esperaba, debido a la complejidad genética de la gran mayoría de enfermedades. Este hecho junto con el descubrimiento de una gran variación estructural en el genoma humano [Iafrate et al. 2004, Sebat et al. 2004], ha hecho que hoy por hoy los objetivos no son sólo la variación nucleotídica sino también la variación estructural del genoma humano, que puede contener millones de nucleótidos heterogéneos en cada genoma, y que ha contribuido seguramente de manera importante a la diversidad humana y a la susceptibilidad a enfermedades.

Por último, la variación genética no sólo es importante para entender la relación genotipo-fenotipo o para encontrar la causa genética de las enfermedades sino que analizando sus patrones podemos determinar como los procesos evolutivos y demográficos han moldeado el genoma a lo largo de la historia. Hay procesos que han afectado a los patrones de variación del genoma entero, como es la historia demográfica de las poblaciones, con sus fluctuaciones de tamaño y estructura. Por otra parte, los

procesos evolutivos han afectado la variación de regiones específicas, como son la selección natural, mayor supervivencia de los individuos más adaptados a unas condiciones ambientales concretas; la mutación, cambio puntual en la secuencia nucleotídica, o la recombinación, intercambio genético que produce una nueva combinación alélica [Tishkoff and Verrelli, 2003]. Por lo tanto, el análisis de la evolución del genoma a través de los patrones de variación genética proporciona también claves para conocer las bases genéticas de los caracteres fenotípicos complejos.

1.2 Variación estructural

1.2.1 ¿Que es la variación estructural?

La variación estructural implica cambios en la secuencia que varían en tamaño desde pocos nucleótidos hasta las grandes reorganizaciones cromosómicas visibles al microscopio y el criterio de clasificación por tamaño es distinto según los autores. Estos cambios en la orientación, localización y ganancia o pérdida de fragmentos de ADN definen la variación estructural junto con las nuevas uniones de secuencia que se generan con las mutaciones y que se conocen como puntos de rotura. Los puntos de rotura definen los límites de las variantes estructurales y, en muchos casos, sirven para detectarlas. La mejora de las técnicas de detección ha permitido detectar variantes estructurales cada vez más pequeñas por lo que su definición ha ido cambiando a lo largo del tiempo. El descubrimiento de una gran cantidad de variantes estructurales con tamaños en el orden de las kilo bases, Kb, a partir de la secuenciación del genoma humano [Iafrate et al. 2004, Sebat et al. 2004] hizo que muchos autores tomaran como referencia los cambios mayores de 1 Kb, aunque en los últimos años se ha descubierto un gran número de variantes más pequeñas. Las variantes estructurales se pueden clasificar en dos grandes grupos según su tamaño. Aquellas que pueden ser identificadas usando un microscopio, habitualmente del orden de mega bases, Mb, se denominan variantes microscópicas y las variantes del orden de kilo bases, se denominan submicroscópicas y necesitan de técnicas con mayor resolución para ser detectadas.

Las variantes microscópicas más grandes destacan en el estudio de los cariotipos, es decir, el conjunto de cromosomas de un individuo o especie, gracias a la tinción de los cromosomas y fueron las primeras en ser descubiertas. Algunas de estas variantes son las aneuploidias, los heteromorfismos, los sitios frágiles, los cromosomas anillo, los isocromosomas, los cromosomas marcador, y otras reorganizaciones complejas.

Las aneuploidias son variaciones en el número de cromosomas y se engloban en la variación estructural ya que se pueden considerar duplicaciones, ganancia de un fragmento de ADN por copia, o deleciones, pérdida de un fragmento de ADN, de cromosomas enteros. Pueden afectar a los autosomas o a los cromosomas sexuales y se

clasifican según el número de copias que tenga el individuo afectado para un determinado cromosoma. Las nulisomías implican la falta de un par de cromosomas homólogos, las monosomías son pérdidas de un cromosoma y las disomías son duplicaciones de cromosomas. Las disomías normalmente afectan a los cromosomas sexuales en los individuos heterogaméticos (normalmente los individuos macho) cuando hay dos copias de uno de los cromosomas, aunque también se dan disomías uniparentales en los cromosomas autosómicos. Ocurren cuando un organismo con reproducción sexual, recibe dos copias de un cromosoma del padre o la madre y ninguna del otro. Las trisomías describen la ganancia de un cromosoma. Las tetrasomías se forman por la ganancia de dos cromosomas homólogos y por último las pentasomías por la ganancia de tres cromosomas [Griffiths et al. 2000]. Las aneuploidias fueron descubiertas en su gran mayoría gracias a las enfermedades y síndromes de los que son causantes. En humanos los síndromes más conocidos corresponden a trisomías, ya que no son tan nocivas como las monosomías (aunque una excepción es la monosomía del cromosoma *X* en mujeres, el síndrome de Turner*)*. Es esperable que la duplicación de un cromosoma tenga un efecto menos negativo que la deleción de un cromosoma, ya que en la primera sigue habiendo una dotación genética completa que realiza su función, aunque sean muchos los problemas derivados de tener una dosis génica (el número de copias de un gen determinado en un célula) duplicada; mientras que la deleción de un cromosoma implica que todos los genes de ese cromosoma dejen de realizar su función. Las aneuploidias de cromosomas autosómicos a su vez son más nocivas que las de cromosomas sexuales. Este hecho está relacionado con el mecanismo de compensación de dosis génica que tienen los cromosomas sexuales. Entre las trisomías se encuentran los síndromes de Klinefelter (hombres *XXY*), de Down (trisomía del cromosoma *21*), de Edwards (trisomía del cromosoma *18*), las trisomías de los cromosomas *8*, *9* y *16* y los síndromes del triple *X* (mujeres *XXX*) y doble *Y* (hombres *XYY*) [Griffiths et al. 2000].

Además de las aneuploidias, el análisis del cariotipo humano ha permitido encontrar otras variantes estructurales: por ejemplo el heteromorfismo se refiere a regiones visibles de los cromosomas que varían en tamaño, en morfología o bien en como se ven tras ser teñidas; los sitios frágiles son constricciones de los cromosomas o incluso pequeñas roturas; los isocromosomas son cromosomas metacéntricos que durante la meiosis o mitosis dividen su centrómero horizontalmente en vez de verticalmente y pierden un brazo, el brazo restante es copiado y el cromosoma tiene dos brazos idénticos en sentido inverso; los cromosomas marcador o cromosomas supernumerarios son cromosomas de estructura anormal que son detectados en experimentos de *FISH* (del inglés Fluorescence In Situ Hibridization, técnica que se explica con más detalle en los siguientes apartados), además del complemento normal de cromosomas; y los cromosomas en anillo se forman por la fusión de ambos brazos cromosómicos entre otros.

En general, las variantes estructurales microscópicas han sido descubiertas a raíz de su asociación a un síndrome o enfermedad o bien su estudio ha llevado a encontrarlos y es por eso que se denominan también anomalías estructurales. Más allá de estas variantes estructurales fácilmente detectables al microscopio con una tinción básica, se han detectado más variantes microscópicas que debido a su menor tamaño, requieren mayor resolución y en este caso, fueron descubiertas gracias a la mejora de las técnicas de tinción, técnicas de bandeo, y de elongación de los cromosomas metafásicos [Feuk et al. 2006]. Entre ellas están las translocaciones, duplicaciones, deleciones, inserciones e inversiones (**Figura 1.1**).

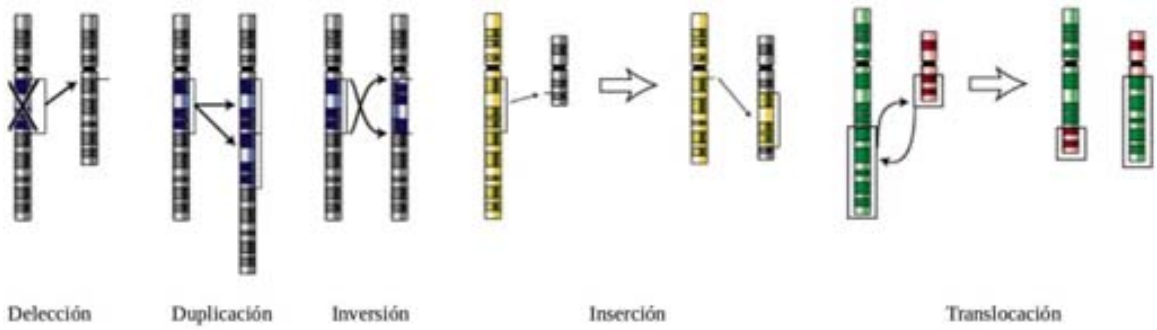

Figura 1.1: Ejemplo de variantes estructurales microscópicas. (De izquierda a derecha) Deleción intersticial, Duplicación en tándem, Inversión paracéntrica, Inserción o Translocación desequilibrada y Translocación recíproca.

Las translocaciones recíprocas engloban el intercambio mutuo de ADN entre los brazos de dos cromosomas. En otras ocasiones, se da la fusión de dos cromosomas acrocéntricos cerca de los centrómeros en los brazos cortos de manera que se genera un cromosoma dicéntrico. Cuando los cromosomas son telocéntricos se unen directamente por el centrómero. Estas variantes se denominan translocaciones robertsonianas. En los cromosomas generados se suele inactivar un centrómero de manera que acaban siendo viables. Por otro lado el material genético de los brazos cortos normalmente se pierde. Las duplicaciones son repeticiones de un fragmento del cromosoma, y pueden ser directas cuando mantienen la misma orientación que el fragmento original, en tándem cuando además el fragmento original y el nuevo se encuentran en la misma región e invertidas cuando la orientación del fragmento repetido es la opuesta al fragmento original. Las deleciones son pérdidas de un fragmento cromosómico, y pueden ser intersticiales cuando el fragmento está en medio de un brazo cromosómico y terminales cuando está en un extremo del brazo y se pierde el telómero. Si se encuentran en un cromosoma autosómico se denominan deleciones constitutivas y tienen efectos fenotípicos graves, como los síndromes de Wolf-Hirschhorn y cri du chat producidos por las deleciones en los brazos *p* o *q* del cromosoma *18* o *p* del cromosoma *4* y la deleción del brazo *p* del cromosoma *5*, respectivamente. Las inserciones o translocaciones desequilibradas hacen referencia a la deleción de un segmento intersticial de un cromosoma que es insertado o transferido a una nueva ubicación en otro cromosoma homólogo o no homólogo e incluso en el mismo

cromosoma de donde proviene el fragmento. Al igual que las duplicaciones, pueden ser directas si mantienen la orientación del fragmento original o invertidas si no la conservan. Por último, las inversiones son cambios en la orientación de un fragmento de cromosoma, en los que se rompe un fragmento y es reparado en la misma posición pero en orientación inversa. Se denominan paracéntricas cuando están localizadas en un brazo del cromosoma y pericéntricas cuando cada uno de los dos puntos de rotura se encuentra en un brazo cromosómico y por lo tanto incluyen el centrómero, que cambia de posición. A este nivel microscópico se detectan más fácilmente las inversiones pericéntricas que suelen tener un tamaño mayor [Huret et al. 2000].

Hoy en día, se puede detectar un mayor número de variantes microscópicas gracias al avance de las técnicas de *FISH*, que tienen una mayor precisión que las técnicas de tinción clásicas [Feuk et al. 2006]. Consecuentemente con el aumento progresivo de la resolución de las técnicas de detección, se ha descubierto un nuevo espectro de variación estructural, de tamaño menor de 3 Mb y que no puede ser descubierta usando un microscopio. Estamos hablando de la variación estructural submicroscópica. La variación submicroscópica al igual que la variación microscópica, incluye diferencias en la orientación, el número de copias y la localización de segmentos del genoma, simplemente a una escala de tamaño menor, cromosómica en vez de cariotípica. No obstante, existen diferencias entre ambas ya que la variación estructural microscópica implica procesos cromosómicos que no tienen sentido a nivel submicroscópico, como las translocaciones robertsonianas. En el punto opuesto están las inserciones que producen los elementos móviles transponibles del ADN. Estas secuencias denominadas transposones se “mueven” o cambian su posición en el genoma y se diferencian entre sí por su mecanismo de transposición. Los transposones con mecanismos de copia y pegado generan duplicaciones de secuencia en el genoma que se localizan en sitios distintos del fragmento original, de ahí que se denominen transposiciones.

Como ya hemos comentado anteriormente, las variantes microscópicas o anomalías estructurales se han considerado raras y han estado asociadas a síndromes y enfermedades a pesar de no conocerse los mecanismos moleculares a través de los que predisponen a sufrirlos en muchos casos [Hall and Quinlan. 2012] y prácticamente no se han relacionado con fenotipos “sanos” [Feuk et al. 2006]. Por lo tanto, se ha asumido durante años que la variación genética estaba localizada en las variantes de escala pequeña, como son los *SNPs*, los *VNTR* y otros cambios de tamaño del orden de pocos nucleótidos. Por eso el conocimiento de la variación entre estos dos extremos no ha aumentado hasta los últimos años. Se ha visto que esta variación de tamaño intermedio tiene una frecuencia mucho más alta en los individuos y en las poblaciones que la variación microscópica y por otro lado, afecta a muchas más bases que la variación más pequeña, la nucleotídica. Un ejemplo de ello se muestra en los resultados del análisis de la variación entre el genoma de J. Craig Venter y el genoma de Referencia, donde se calcula que ambos genomas difieren en un 1.5% debido a la variación estructural frente al 0.1% de variación

por SNPs [Pang et al. 2010]. La variación submicroscópica puede contener millones de bases de ADN, afectar a genes y regiones reguladoras o bien estar situada en regiones no funcionales del genoma y no tener consecuencias fenotípicas claras. Por lo tanto no sólo puede explicar los fenotipos asociados a enfermedad sino también los fenotipos que forman parte de la diversidad dentro de una misma población o especie. De ahí que hoy en día se entienda como variación estructural la variación submicroscópica.

Hasta ahora hemos clasificado las variantes estructurales según su tamaño pero también se pueden clasificar según impliquen un cambio en la cantidad de material genético o no. Las variantes no balanceadas implican ganancias o perdidas, mientras que las variantes balanceadas no implican cambio alguno en la cantidad de ADN. Las variantes no balanceadas incluyen las inserciones, deleciones, duplicaciones y transposiciones, mientras que las inversiones, translocaciones equilibradas y disomías uniparentales segmentales (de un fragmento del cromosoma) son variantes balanceadas, aunque las translocaciones pueden ser no balanceadas si se da un intercambio desigual de ADN. Precisamente este criterio de clasificación ha desembocado en un nuevo concepto, la variación en el número de copias o *CNV* (del inglés Copy Number Variant). Se define una variante estructural como *CNV* cuando comprende un fragmento de genoma que varía en el número de copias respecto al número de copias de otro genoma que ejerce de referencia.

Además se dan reorganizaciones estructurales complejas que no pueden clasificarse según este criterio y que suelen incluir más de una variante [Quinlan and Hall, 2012]. Una característica clave que define estas variantes es el agrupamiento de puntos de rotura obtenidos a partir de una sola mutación que no puede ser explicado por los mecanismos habituales relacionados con la reparación del ADN o la recombinación. La complejidad de estas variantes no es homogénea. Agrupan desde alteraciones en un solo locus, como pueden ser múltiples deleciones, duplicaciones y reorganizaciones o bien pequeñas inserciones y deleciones presentes en los puntos de rotura de una variante estructural más grande; hasta reorganizaciones muy complejas entre distintos loci en varios cromosomas que incluyen patrones complejos de *CNVs* en los puntos de rotura de otras variantes o cerca de ellos. Estos casos de mayor complejidad suelen estar asociados con enfermedades raras y con cáncer, aunque la mayoría de variantes estructurales complejas han sido identificadas en individuos sanos, por lo que forman parte de la variación estructural normal [Quinlan and Hall, 2012]. Algunos de estos patrones son difíciles de atribuir a una variante compleja ya que se pueden generar patrones similares de agrupación de puntos de rotura por mutaciones estructurales en sitios complejos generados a lo largo de la evolución, como pueden ser los formados por duplicaciones segmentales que han ocurrido repetidas veces a lo largo del tiempo. Este tipo de situaciones hacen que no sea fácil distinguir entre variantes estructurales simples y complejas.

1.2.2 Historia y métodos de detección

Los cromosomas fueron observados por primera vez a lo largo de la segunda mitad del siglo XIX, tras la incorporación de las técnicas de fijación y de tinción en las preparaciones citológicas. Las primeras observaciones se centraron en establecer el número de cromosomas normal de un organismo (cariotipo) y en las diferencias entre especies. Aunque aún no eran el foco de atención, se detectaron las primeras aberraciones cromosómicas y no fue hasta 1921 cuando Alfred Sturtevant publicó la primera evidencia de una inversión cromosómica [Sturtevant. 1921]. En la siguiente década, Theodosius Dobzhansky y colaboradores descubrieron inversiones cromosómicas en los cromosomas politénicos de las especies *Drosophila pseudoobscura* y *Drosophila persimilis* [Dobzhansky et al. 1938]. Se trata de cromosomas gigantes de las células de las glándulas salivares en el género *Drosophila*. Además demostraron la existencia de inversiones polimórficas, es decir que mantienen más de un alelo en la población. Siguiendo estos descubrimientos, en las décadas siguientes los genetistas se centraron en las reorganizaciones genómicas microscópicas, ya que eran las únicas variantes visibles gracias a los primeros métodos citogenéticos. Se creía que las variantes estructurales eran raras en otras especies fuera del género *Drosophila*, ya que las únicas que podían detectar, las microscópicas, estaban asociadas con síndromes o enfermedades como el cáncer. Más tarde, en 1953 James Watson y Francis Crick descubrieron la estructura del ADN [Watson and Crick. 1953], gracias a los datos de difracción de rayos X proporcionados por Rosalind Franklin y posteriormente en el año 1966 se completó la definición del código genético por Har Gobind Khorana y colaboradores [Khorana et al. 1966]. Estos descubrimientos aumentaron el interés por la variación más pequeña, la variación nucleotídica.

En las décadas siguientes se desarrollaron técnicas de clonaje molecular, la secuenciación del ADN y la reacción en cadena de la polimerasa, *PCR*, que permitieron a los investigadores analizar este tipo de variación cada vez a mayor escala. En la década de los 90 y los primeros años del siglo XXI, gracias a la información generada en el proyecto de secuenciación del genoma humano, se estableció que la estructura del genoma era relativamente estática y que una gran parte de la variación genética estaba en las diferencias nucleotídicas. Se llegó a establecer que las diferencias fenotípicas entre dos humanos se podían explicar por la presencia de 1 *SNP* cada 1000 pb (pares de bases), y entre un humano y un chimpancé por 1 cada 100 pb. Esta asunción caló muy hondo en la comunidad científica y fue repetida durante algunos años después. En el año 2001 se publicó la secuencia Referencia del genoma humano, aunque un año antes ya se trabajaba con la secuencia borrador. Evan Eichler y colaboradores demostraron el mismo año 2001 que la estructura del genoma humano es más dinámica de lo que se esperaba al alinear la secuencia del borrador del genoma consigo misma, ya que descubrieron que el 5% está repetido en duplicaciones segmentales o LCR (del inglés Low Copy Repeats) [Bailey et al. 2001]. Se consideran duplicaciones segmentales los fragmentos de más de 1 Kb que

tienen múltiples copias y que comparten más del 90% de sus nucleótidos. Estos datos evidenciaron que el genoma humano ha sufrido grandes reorganizaciones estructurales en su evolución reciente.

El poder acceder a la secuencia referencia del genoma humano supuso un avance en las técnicas de detección de variantes a escala genómica. Se construyeron *microarrays*, micro matrices que contenían cromosomas artificiales de bacteria, *BACs* (del inglés Bacterial Artificial Chromosome) o bien sondas de oligonucleótidos en representación de regiones a lo largo de todo el genoma. La hibridación genómica comparativa *aCGH* (del inglés array Comparative Genomic Hibridization) del genoma de otros individuos con estos *arrays*, reveló en individuos de fenotipo normal niveles de *CNVs* mucho más altos de lo esperado [Iafrate et al. 2004, Sebat et al. 2004]. Estos dos estudios sobre *CNVs* cambiaron nuestro conocimiento sobre la variación estructural en el genoma humano, hasta el punto que ahora mismo se considera más importante que la variación nucleotídica, tanto por los millones de nucleótidos que engloba en un solo individuo como por sus efectos sobre elementos funcionales del genoma [Feuk et al. 2006]. En los años siguientes se han realizado un gran número de estudios de mapeo a nivel genómico que han permitido descubrir todo tipo de variantes estructurales, no sólo *CNVs* sino también inversiones. La resolución de los *microarrays* de oligonucleótidos ha ido en continuo aumento permitiendo detectar *CNV*s en regiones cada vez más pequeñas, y las técnicas de secuenciación y mapeo de los extremos apareados de fragmentos de ADN, *PEM* (del inglés Paired-end Mapping); han permitido detectar todo tipo de variantes [Hall and Quinlan. 2012], aunque este auge se ha notado más en el número de *CNVs* ya que son más fáciles de detectar. En los siguientes apartados se explican más detalladamente estas técnicas de detección.

Finalmente, en los últimos años ha habido una mejora en las técnicas de secuenciación que ha permitido que el coste de este tipo de estudios sea cada vez más bajo y eso ha marcado un *boom* de detección de variación estructural. Además la secuenciación de genomas completos es ahora asequible en términos de tiempo y dinero, dando pie a un nuevo paso en la revelación de la mayor parte de la variación estructural genómica a través de la comparación entre genomas completos, siendo el ensamblaje de novo la clave en la detección de nuevas variantes no representadas en el genoma de Referencia.

1.2.2.1 Descubrimiento de variantes microscópicas y métodos clásicos de detección

En las primeras décadas del siglo XX se estaba estudiando el cariotipo de varias especies entre ellas por supuesto la especie humana, y aunque estos estudios habían comenzado en el siglo anterior, la controversia acerca del número de cromosomas normal de un individuo era aún grande. Para observar los cromosomas se usaba una técnica de tinción sencilla, la reacción de un colorante ante el ADN, normalmente el reactivo de Schiff. Esta

tinción se denomina Feulgen en honor a su descubridor [Griffiths et al. 2000]. Los cariotipos que se podían observar consistían en cromosomas condensados que se distinguían muy poco entre sí, pero se podían distinguir las aneuploidias, los cromosomas marcador, y las reorganizaciones más grandes. Se comenzaron a relacionar con enfermedades, como la trisomía *21* y el síndrome de Down, descubierta por Lejeune en 1959 [Lejeune et al. 1959]. Casi una década más tarde aparecieron las técnicas de bandeo, que permitieron diferenciar partes de los cromosomas, y subdividirlas en regiones más pequeñas o bandas de diferente tinción. Fueron usadas por primera vez en humanos por *Caspersson* en el año 1971 [Caspersson et al. 1971]. Con ellas se pudo descubrir la variación estructural microscópica de menor tamaño y refinar la localización de las variantes que ya habían sido descubiertas y asociadas a enfermedad. Además permitieron distinguir los cromosomas de masas aberrantes de ADN de tamaño similar. Los patrones de bandas son muy característicos de cada cromosoma dentro de la misma especie y también son diferentes según el colorante y la técnica usada. El más conocido es el producido por el reactivo Giemsa, un colorante que tiñe el ADN tras la digestión proteolítica suave de los cromosomas.

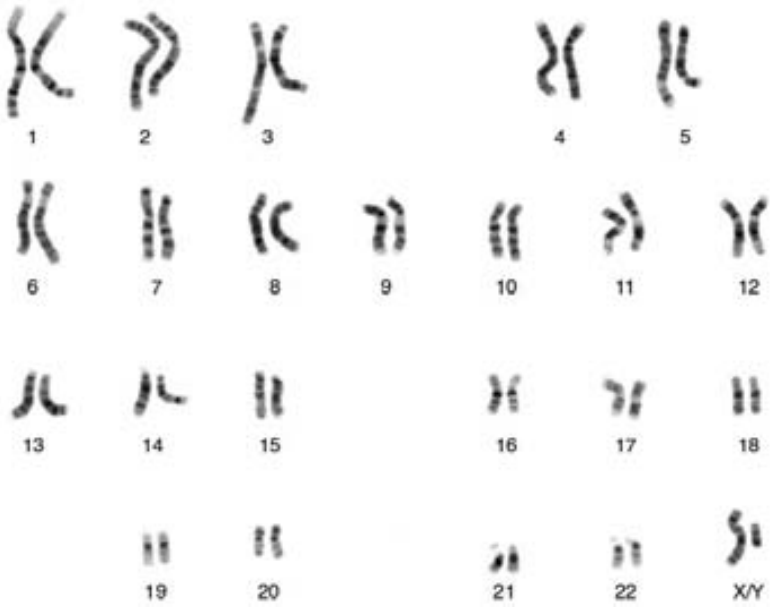

Figura 1.2: Ejemplo del patrón de bandas G en un cariotipo humano de varón. Se muestran los cromosomas por parejas y ordenados de mayor a menor tamaño, primero los autosomas y finalmente el par de cromosomas sexuales. Figura modificada a partir de *Talking Glossary of Genetic Terms* (www.genome.gov/glossary/).

Se generan las bandas G (**Figura 1.2**). Se piensa que están relacionadas con la densidad de empaquetamiento de la cromatina, donde las regiones G oscuras tendrían una densidad mayor de ADN y por lo tanto la tinción se ve más fuerte. El bandeo por Quinacrina o bandeo Q fue la primera técnica de tinción de bandas que se usó y requiere un microscopio de fluorescencia, por lo que ya no se usa [Caspersson et al. 1971]. El bandeo de inversión, bandeo R, necesita tratamiento por calor e invierte las bandas blancas y

negras habituales de los bandeos G y Q. Otras técnicas de tinción son el bandeo-C y la tinción de la zona del organizador nucleolar, o tinción *NOR*. Estas últimas técnicas tiñen porciones específicas del cromosoma [Griffiths et al. 2000]. El bandeo-C tiñe la heterocromatina estructural, que se encuentra normalmente cerca del centrómero, y la tinción *NOR* resalta los satélites y los brazos de los cromosomas acrocéntricos. También se usan la bandas T que tiñen las regiones teloméricas. [Griffiths et al. 2000].

Un sistema especial de bandas es el de los cromosomas politénicos de insectos dípteros (**Figura 1.3**). Fueron estudiadas en los trabajos de inversiones cromosómicas polimórficas en el género *Drosophila* por *Dobzhansky* y colaboradores antes de que apareciesen las técnicas de tinción de bandas en humanos. Los cromosomas politénicos son cromosomas que replican muchas veces su ADN sin que se separe en diferentes cromátidas, de manera que a más copias, más largos y gruesos se vuelven. A lo largo de estos cromosomas, se observan bandas transversales, y son más numerosas que las bandas G, Q o R. Estas bandas varían en tamaño y morfología y estas variaciones forman los patrones específicos de cada cromosoma. Además de las bandas, hay regiones características como los engrosamientos del cromosoma o los estrechamientos también conocidos como anillos de Balbiani, que hacen aún más característicos los patrones cromosómicos [Griffiths et al. 2000].

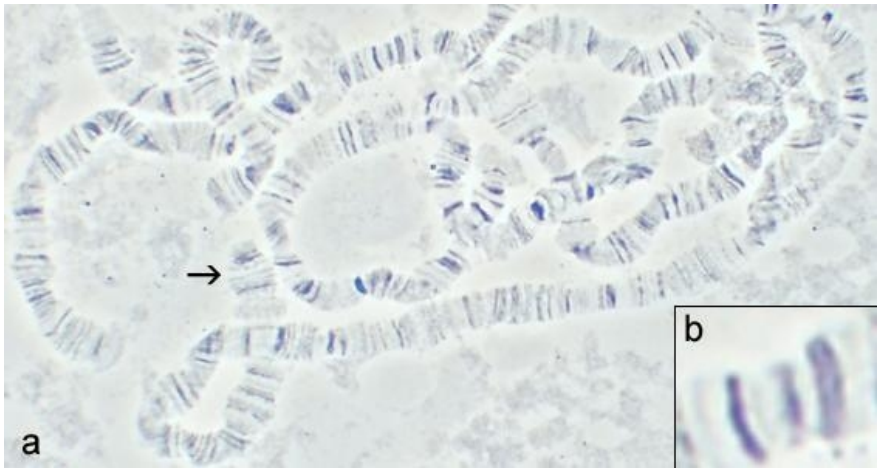

Figura 1.3: Fotografía de cromosomas politénicos de *Drosophila melanogaster*. (a) Se muestran la imagen de los cromosomas obtenida por contraste de fase. El cromocentro se encuentra en la parte superior derecha. La flecha indica el extremo del cromosoma *X*. (b) Detalle del patrón de bandas e interbandas. Figura modificada a partir de Griffiths et al. 2000.

La mejora de la tinción de los cromosomas no sólo incluyó el bandeo, sino que también una mejora de la resolución a partir del uso de cromosomas en profase o metafase temprana (prometafase), antes de que alcancen la condensación máxima. El número de bandas observables para todos los cromosomas aumentó, pasó de unas 300 hasta unas 800. Esto permitió detectar anomalías menos obvias que normalmente no se verían con los bandeos convencionales. Una década más tarde, en 1982, se publicó una técnica que

fue descrita como un método inmunológico para localizar genes en los cromosomas politénicos de *Drosophila* [Langer-Safer et al. 1982]. La técnica denominada *FISH*, se basa en la hibridación de sondas de ADN marcadas con moléculas fluorescentes en células en interfase, cromosomas metafásicos o fibras de ADN, para detectar la presencia o la orientación relativa de secuencias diana. Para cumplir ese objetivo, las sondas sólo hibrídan con las partes del cromosoma con las que tienen un alto grado de complementariedad. El uso de esta técnica aumentó de nuevo la resolución en la detección de variantes estructurales y permitió descubrir variantes estructurales microscópicas que no se podían detectar anteriormente (**Figura 1.4**). Se aplica en distintos campos como el consejo genético, la medicina, e incluso para distinguir individuos de especies diferentes. Además puede usarse con secuencias diana de ARN en células normales, tumorales y de distintos tejidos, de manera que permite conocer los patrones de expresión génica en células y tejidos y en distintos estados del desarrollo. Su principal limitación es que solo permite analizar una región específica del genoma, o mejor dicho, las sondas se generan de una manera individualizada para unas regiones diana, de manera que no es una técnica que permita analizar diversas regiones del genoma en un mismo experimento, sino una técnica dirigida.

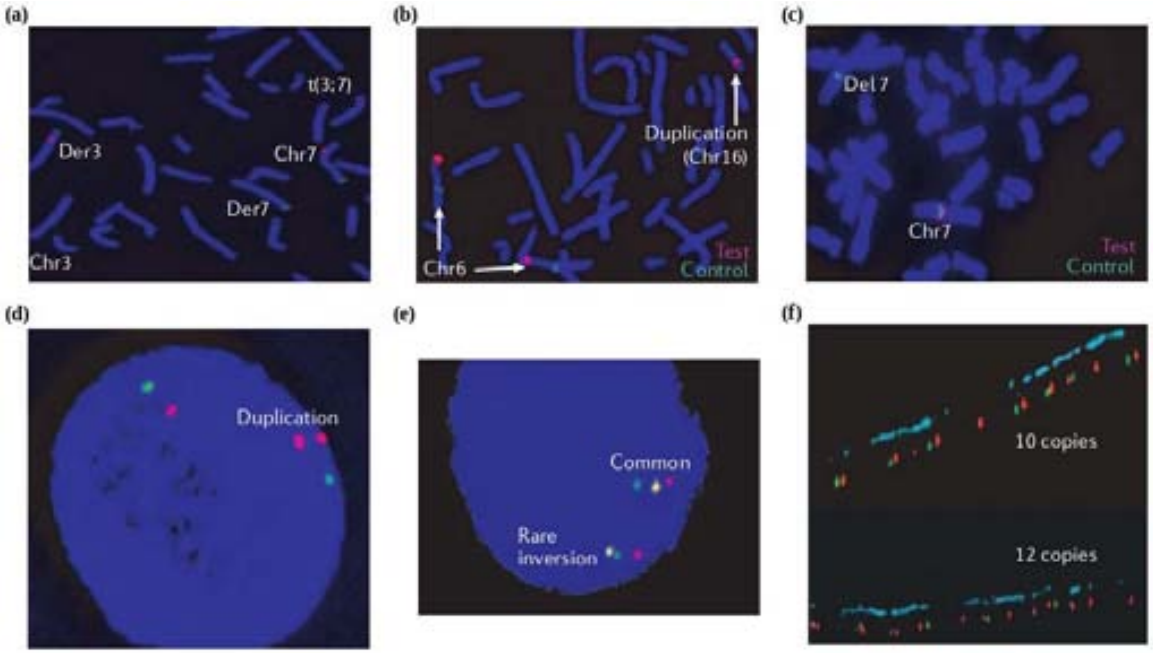

Figura 1.4: Detección de variantes estructurales mediante *FISH*. (a) Translocación críptica entre los cromosomas 3 y 7, que se encuentran en metafase. Los cromosomas derivados de la translocación se indican con (Der). (b) Duplicación en el cromosoma 16. (c) Deleción en el cromosoma 7. (d) Duplicación detectada en un núcleo en interfase. (e) Inversión. (f) *FISH* de alta resolución en cromosomas estirados para detección de *CNVs* en dos cromosomas diferentes. Figura modificada de Feuk et al. 2006.

Las sondas se fabrican a partir de clones de la secuencia diana, y luego se marcan con moléculas fluorescentes de distintos colores. Las variantes estructurales de detectan gracias al diseño de sondas específicas para la variante a detectar, por ejemplo las duplicaciones y deleciones se detectan por el número de señales fluorescentes que se distinguen en los cromosomas diferentes del número señales esperadas. Una deleción se detecta por la falta de señal fluorescente correspondiente a una secuencia y una duplicación por la presencia de una señal fluorescente extra. Las translocaciones se detectan porque la señal aparece en un cromosoma diferente a donde se localiza la secuencia diana. En el caso de las inversiones se diseñan sondas de distintos colores dentro y fuera de la secuencia invertida o bien a lo largo de un cromosoma, y es precisamente el orden de las señales fluorescentes el que desvela la orientación de la secuencia.

Una aplicación específica es el coloreado cromosómico. En este caso se usan un conjunto de fragmentos de ADN clonados que sabemos pertenecen a cromosomas o regiones cromosómicas concretas y se marcan con compuestos fluorescentes diferentes que los colorean permitiendo distinguir las reorganizaciones entre cromosomas en el microscopio de fluorescencia. En general las técnicas usadas para la detección de variantes estructurales han evolucionado de un planteamiento dirigido a un planteamiento más global o genómico a medida que la capacidad de detección de variantes de menor tamaño ha ido en aumento. Por ejemplo el coloreado de cromátidas por hibridación genómica direccional [Ray et al. 2013] es una técnica citogenética que se basa en la técnica de *FISH* para determinar la orientación de un fragmento de ADN y está orientada a la detección de inversiones cromosómicas ya sea a nivel genómico o dirigido (**Figura 1.5**).

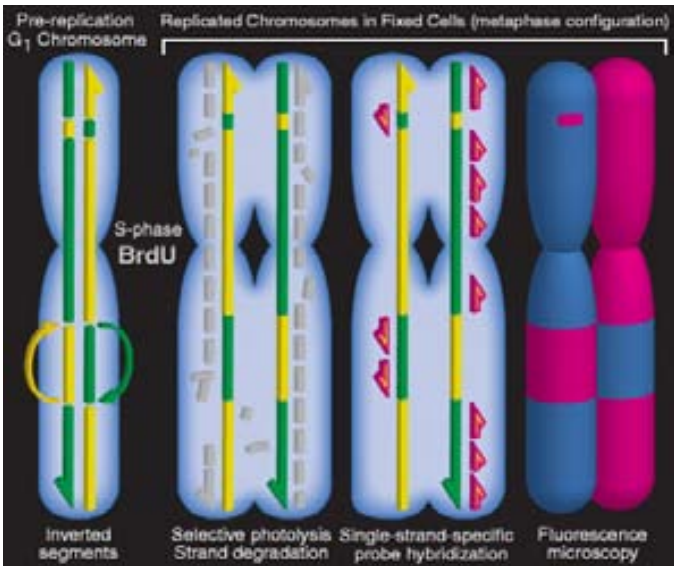

Figura 1.5: Técnica de coloreado de cromátidas por hibridación genómica direccional aplicada a la detección de inversiones cromosómicas. De izquierda a derecha están indicados la fase previa a la replicación de las cromátidas en la metafase y los tres pasos de la técnica. Antes de la metafase las cromátidas madre están formadas por dos cadenas complementarias. En la

metafase cada cadena sencilla da lugar a una nueva cadena complementaria que incorpora la 5'-Bromo-2'-deoxiuridina (*BrdU)* que las vuelve fotosensibles. Entonces se degradan las cadenas nuevas por fotolisis y se hibridan las sondas fluorescentes a las cromátidas hijas de cadena sencilla. Éstas revelan una señal fluorescente que indica los cambios de orientación ya que las zonas invertidas no tendrán el color asignado a la cadena madre sino el color de la cadena complementaria. Imagen obtenida de Ray et al. 2013.

La resolución de esta técnica es de 1 Mb, es decir, las sondas están separadas entre ellas por 1 Mb de distancia, de manera que se detectan las inversiones de 1 Mb o más. Las inversiones de menor tamaño sólo se detectarían en el caso de que la sonda hibridáse completamente dentro de la inversión. Se puede aumentar la resolución aumentando la densidad de las sondas pero hay un límite de densidad para poder detectar la señal fluorescente.

El *Southern Blot* es otra de las técnicas dirigidas que se puede usar para detectar variantes estructurales. Fue publicada en 1975 por *Southern* y también se basa en la hibridación de sondas para detectar la presencia de una secuencia diana en una mezcla de fragmentos de ADN [Southern. 1975]. Si el ADN no está fragmentado, se suele digerir con enzimas de restricción. La técnica utiliza la electroforesis para separarlos, que consiste en la migración del ADN en un campo eléctrico. Debido a la carga eléctrica negativa del ADN, que es proporcional a su tamaño, los fragmentos migran del polo negativo o cátodo hacia el polo positivo o ánodo, a través del gel, lo que permite separar los fragmentos de tamaños concretos. Se puede conocer el tamaño comparando sus posiciones con las de fragmentos de tamaño conocido. Los fragmentos del mismo tamaño se agrupan en bandas en el gel. Estas bandas pueden visualizarse gracias a la tinción del ADN con agentes intercalantes como el bromuro de etidio, que hace que el ADN se pueda ver cuando se expone a luz ultravioleta. La técnica de *Southern blot* se basa en transferir por capilaridad las bandas resultantes de la electroforesis a una membrana absorbente, donde mantienen la misma posición relativa. La membrana se sumerge en una solución con la sonda que está marcada para poder detectarse, de manera que la sonda se une sólo a las bandas que son homólogas. Se lava la membrana para eliminar la sonda que no se ha unido a ninguna banda y se revela para ver en que bandas se ha hibridado. Esta técnica permite detectar inversiones a partir del diseño de sondas específicas para la secuencia de sus puntos de rotura. Por ejemplo, se puede digerir el ADN con enzimas de restricción que tengan dianas dentro y fuera de la región invertida de manera que obtengamos un fragmento que contiene el punto de rotura. El diseño de una sonda específica para la secuencia que contiene parte de la inversión nos permitirá detectar su presencia. En el caso de los *CNVs* se trata de encontrar dianas de restricción que nos generen el fragmento a analizar y diseñar una sonda específica de la secuencia que puede estar duplicada o delecionada, de manera que la presencia o ausencia de señal nos permitirá detectar una deleción y la intensidad de la señal nos permitirá detectar una duplicación.

Muchos métodos incluyen la electroforesis ya que permite separar macromoléculas (ADN, ARN y proteínas) por su tamaño. Aunque no se sabe bien en qué año fue inventada, se tienen referencias del siglo XIX. En el año 1984, Schwartz y Cantor publicaron la electroforesis de campo pulsante o *PFGE* (del inglés Pulsed Field Gradient Gel Electrophoresis) [Schwartz and Cantor. 1984], una variación de la técnica que introduce varios campos eléctricos oscilantes que están orientados en varias direcciones. En la técnica convencional, los fragmentos pequeños de ADN atraviesan la matriz o porosidad del gel más fácilmente que los fragmentos de gran tamaño, que a partir de las 30-50 Kb migran por el gel a la misma velocidad formando una gran banda difusa. En el *PFGE* se consigue que las moléculas de gran tamaño reaccionen de forma diferente a los cambios de sentido del campo eléctrico, los fragmentos más grandes reaccionan más lentamente, y así migran a posiciones diferentes del gel según su tamaño. El rango de resolución para fragmentos de ADN aumenta hasta en 2 órdenes de magnitud. La aplicación en la detección de variantes estructurales es similar al *Southern blot*, con la diferencia de que usando *PFGE* podemos detectar variantes más grandes.

Otra técnica dirigida que sirve para detectar variantes estructurales es la *PCR* o reacción en cadena de la polimerasa. Fue publicada en el año 1986 y está diseñada para obtener un gran número de copias a partir de un fragmento de *ADN* molde [Mullis, 1986]. Se usa mucho en biología molecular ya que evita tener que clonar una secuencia para tener muchas copias. Se basa en la actividad de las enzimas polimerasas de ADN que replican las hebras. Se aplican pasos alternos de altas y bajas temperaturas para separar las hebras de ADN y dejar que vuelvan a unirse para poder duplicarlas nuevamente. Se repiten los ciclos para obtener múltiples copias del fragmento molde. La amplificación de la secuencia diana se produce gracias a unos cebadores que son específicos de la secuencia a amplificar y que hibrídan por complementariedad de bases.

Los constantes cambios de temperatura se realizaban en baños de agua al principio. Otro inconveniente era la desnaturalización de las polimerasas. La aparición de los termocicladores, aparatos capaces de cambiar la temperatura del tubo de ensayo muy rápidamente, y la incorporación de las polimerasas termoresistentes de organismos termófilos, fueron las dos grandes mejoras de la técnica que hicieron que sea ahora una de las más utilizadas por su rapidez y eficacia. Junto con la electroforesis permiten amplificar y separar secuencias específicas de ADN. Su aplicación en la detección de variantes estructurales se basa en el diseño de cebadores específicos de las secuencias a analizar, por ejemplo, en los puntos de rotura en el caso de las inversiones. De esta manera la amplificación del fragmento a analizar nos indica su presencia a través de la electroforesis del producto y la posterior tinción de las bandas.

En el caso de los *CNVs* se puede aplicar la *PCR* en tiempo real o *PCR* cuantitativa. Esta técnica cuantifica el ADN o ARN a la vez que lo amplifica, por lo que necesita termocicladores capaces de leer la fluorescencia que emite el producto amplificado. Si se

diseñan cebadores específicos para un *CNV* pueden analizarse el número de copias de un individuo comparando la cantidad de ADN con la de un individuo de referencia. La diferencia con la *PCR* normal es que se detecta el producto en tiempo real en vez de al final. En cuanto a la detección, se usan dos métodos diferentes, los colorantes fluorescentes que se intercalan en el ADN de manera inespecífica como el *SYBR Green* y las sondas de ADN específicas, que están formadas por oligonucleótidos marcados fluorescentemente. En el primer caso, el incremento de producto de *PCR* provoca un incremento en la fluorescencia, que se mide en cada ciclo, de manera que a partir de la fluorescencia conocemos la concentración de ADN. En el segundocaso, se detecta el ADN diana al que se une la sonda, por lo tanto la cuantificación es más específica y los productos inespecíficos de *PCR* no la afectan. Además permite detectar varias secuencias diana en la misma reacción si se usan sondas con diferentes marcadores. La sonda tiene un emisor de fluorescencia en un extremo y un inhibidor en el otro. Si ambos están juntos no se emite fluorescencia, pero la acción de la polimerasa rompe esta proximidad gracias a su actividad exonucleasa 5'→3' y la fluorescencia se emite. La rotura de la sonda en cada ciclo hace que la fluorescencia emitida sea proporcional al incremento de producto. En ambos casos la detección de *CNVs* se basa en el diseño de cebadores específicos que nos permitan amplificar la secuencia candidata que emitirá fluorescencia directa o indirectamente pero de manera proporcional a la cantidad de producto amplificado; y la comparación de la intensidad de la señal fluorescente con una secuencia control, nos permitirá detectar la presencia de una deleción en el caso de emitir menor intensidad de señal o una duplicación en el caso de emitir mayor intensidad.

Existen otras aplicaciones de la *PCR* que permiten la detección de variantes en condiciones donde la técnica normal no lo consigue, por ejemplo la *PCR* de largo alcance. Ésta se aprovecha de la capacidad de algunas Taq polimerasas para amplificar productos más grandes que los que se pueden amplificar con una Taq polimerasa normal. Pueden llegar a amplificar productos de hasta 10 Kb mientras que en una *PCR* normal el límite se sitúa alrededor de las 5 Kb. Esto puede ser útil para detectar variantes estructurales con puntos de rotura localizados en zonas repetidas, como pueden ser las inversiones mediadas por duplicaciones segmentales. En la misma situación se puede aplicar otra variación de la *PCR*, la *PCR* inversa [Ochman et al. 1988]. Ésta se basa en generar moléculas circulares de ADN con los fragmentos que contienen los puntos de rotura. El ADN se corta mediante enzimas de restricción y se ligan los fragmentos consigo mismos. Los cebadores están orientados en orientación inversa a la *PCR* normal, de manera que al circularizar los fragmentos quedan orientados uno frente al otro y se amplifican los extremos del fragmento, por lo que la técnica nos permite conocer que hay a un lado y a otro de la duplicación segmental, en este caso, y detectar la inversión (**Figura 1.6**) [Aguado et al. 2014].

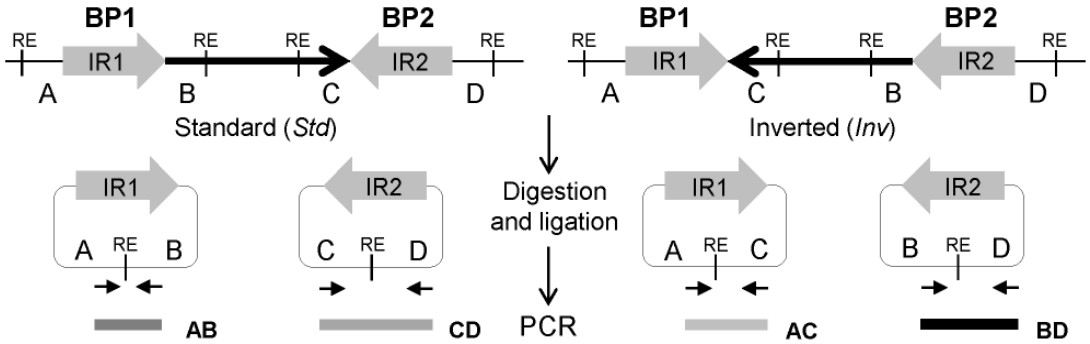

Figura 1.6: Esquema de la *PCR* inversa en la genotipación de inversiones. La orientación estándar a la izquierda y la orientación invertida a la derecha, están representadas por flechas gruesas de color negro. Los puntos de rotura están dentro de duplicaciones segmentales representadas en color gris, las regiones de hibridación de los cebadores por las letras A, B, C, D; y las dianas de las enzimas de restricción por RE. Los círculos formados a partir de los productos digeridos y ligados consigo mismos se encuentran justo debajo de la secuencia sin cortar. Finalmente los cebadores están representados por pequeñas flechas negras. Figura tomada de Aguado et al. 2014.

En ambos casos la detección de una inversión o un *CNV* se produce al igual que por el protocolo estándar de *PCR*, mediante el diseño de cebadores específicos de la variante a detectar y la amplificación de la secuencia.

Finalmente, la *PCR* de fusión de haplotipos (hfPCR) es aplicable a la misma situación [Turner et al. 2006]. Se basa en generar un fragmento de ADN que resulta de la fusión de dos fragmentos diferentes con el objetivo de generar genes de fusión, es decir, constructos donde se usan los elementos de regulación de un gen para expresar la parte codificante de otro gen [Yon and Fried. 1989]. Esto es útil por ejemplo para ver qué secuencias regulan la expresión del primer gen o bien para expresar más eficientemente el segundo gen. La técnica parte de los dos fragmentos de ADN a fusionar y de tres cebadores, dos de ellos son específicos para el primer fragmento y uno para el segundo fragmento. El segundo cebador del primer fragmento contiene una cola de nucleótidos que es complementaria al segundo fragmento de manera que en el primer ciclo de amplificación se usan los cebadores 1 y 2 y obtenemos el primer fragmento con la cola complementaria al segundo fragmento y en el segundo ciclo con los cebadores 1 y 3 obtenemos el producto de fusión que es amplificado en los siguientes ciclos [Yon and Fried. 1989]. En el caso de la detección de inversiones, los dos fragmentos corresponderían a secuencias dentro y fuera de la inversión que estarían separadas por la duplicación segmental. Precisamente fue la técnica que Turner y colaboradores desarrollaron para genotipar este tipo de inversiones con puntos de rotura localizados en LCRs [Turner et al. 2006].

1.2.2.2 Secuenciación del genoma humano y aparición de métodos basados en secuencia

En apartados anteriores hemos visto como la secuencia del genoma humano permitió diseñar sondas por todo el genoma y esto fue clave para el desarrollo de los *microarrays o* micro matrices, que permite comparar dos genomas en cuanto a su variación de *CNVs*. Estas técnicas provocaría una revolución en cuanto a la detección de variantes estructurales submicroscópicas y marcaría el paso de las técnicas dirigidas a las técnicas genómicas [Sharp et al 2006]. Estas técnicas se pueden dividir en la hibridación genómica comparada en matriz, *aCGH*, y las matrices de *SNPs*. En la hibridación genómica comparada se marcan fluorescentemente un genoma prueba y un genoma de referencia, con marcadores de distinto color. Entonces se les bloquean los elementos repetitivos para que no puedan hibridar y se hibridan ambos genomas completos con una matriz de fragmentos de ADN fijada en un sustrato vidrioso, que representan las regiones a testar y que pueden corresponder a ventanas de todo el genoma. Las diferencias en la señal fluorescente indican los cambios en el número de copias entre ambos genomas (**Figura 1.7**).

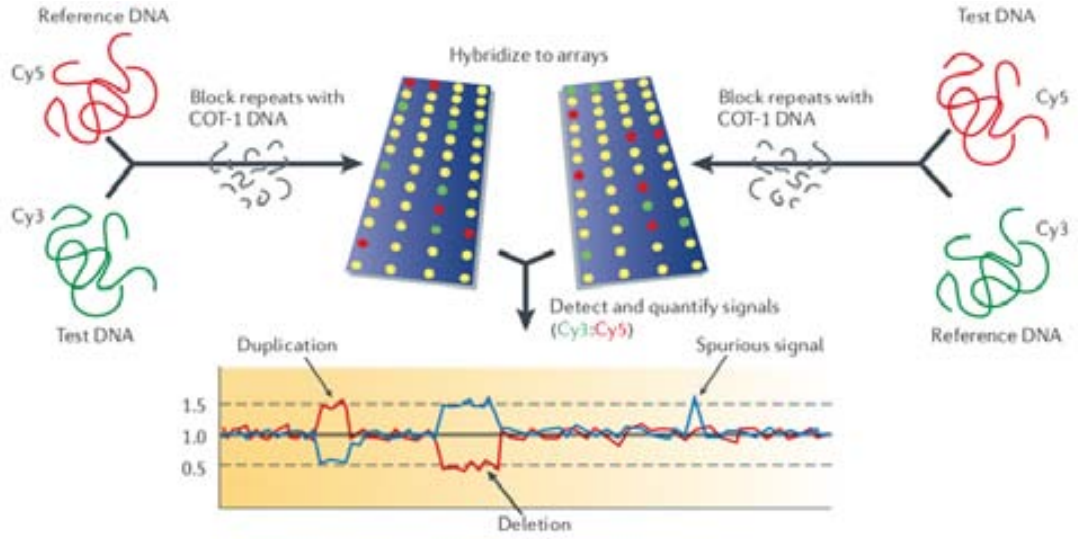

Figura 1.7: Esquema de la técnica de hibridación genómica comparada. El ADN de referencia y el de prueba están marcados con diferentes marcadores, *Cy5* y *Cy3* respectivamente. Se hibrídan a las matrices de ADN genómico después de bloquear los elementos repetitivos con ADN *COT-1*. Después se determinan las diferencias en la señal fluorescente para detectar los *CNVs*. Figura modificada de Feuk et al. 2006.

Una pérdida de ADN en la muestra de referencia y una ganancia en la muestra de prueba producen una misma señal fluorescente, de manera que es necesario tener un genoma de Referencia muy bien caracterizado para poder interpretar los datos del experimento [Alkan et al. 2011]. Gracias al proyecto de secuenciación del genoma humano se pudo generar una librería de fragmentos clonados, *BACs* de localización conocida que se

usaron en los primeros *microarrays*. Los *CNVs* que se podían detectar con estas matrices de *BACs* tenían un tamaño mayor de 100 Kb debido a la baja resolución y densidad de sondas [Iafrate et al. 2004, Sebat et al. 2004]. El uso de otros tipos de sondas como ADN complementarios clonados o productos de *PCR* de cadena única, permitió analizar los *CNVs* con mayor resolución y se usaron especialmente para ver efectos en exones [Sharp et al. 2006]. Finalmente el uso de oligonucleótidos, es decir, de sondas sintetizadas de 25 a 75 pb de longitud, marcó la diferencia en cuanto a mayor resolución de los *microarrays*. Su síntesis es rápida y da muy buenos resultados en cuanto a uniformidad y densidad de las sondas en la matriz [Sharp et al. 2006], por lo que se usan actualmente microarrays con alrededor de 2 millones de sondas de 25 pb y con 1 millón de oligonucleótidos de 75 pb. De esta manera un *CNV* está representado por entre 3 y 10 sondas y su localización es mucho más precisa evitándose la variación entre muestras [Alkan et al. 2011].

En el paso del uso de *BACs* a oligonucleótidos apareció una variación de la técnica llamada *ROMA*, análisis de *microarray* de oligonucleótidos representacionales [Sebat et al. 2004]. Se basa en reducir la complejidad del genoma para aumentar la señal obtenida respecto al ruido de fondo. Esto se consigue cortando el ADN con enzimas de restricción que tengan dianas repartidas uniformemente por el genoma y ligando los fragmentos a adaptadores. Estos adaptadores tienen cebadores para realizar una amplificación por *PCR*. Se usan las condiciones de la *PCR* para amplificar sólo los fragmentos de un determinado tamaño que serán los que se hibrídan a la matriz de sondas [Feuk et al. 2006]. Esta variante se desarrolló debido al alto coste de generar *arrays* de oligonucleótidos de alta densidad que cubrieran todo el genoma. Además del *ROMA*, se han aprovechado *arrays* alternativos que fueron diseñados para genotipar *SNPs* pero que se usan para detectar *CNVs* a baja resolución [Sharp et al. 2006]. La diferencia de los *microarrays* de *SNPs* respecto a *aCGH* es que la hibridación se realiza con una única muestra y se computan las intensidades de hibridación de varias muestras para detectar las ganancias o pérdidas de ADN, es decir, se calculan intensidades medias a partir de muestras control con las que se comparan las obtenidas de las muestras prueba. Las desviaciones de la media indican el cambio en el número de copias. Además, estos arrays dan información sobre genotipos, pudiéndose detectar por ejemplo la pérdida de heterocigosidad, falta de variación nucleotídica, y esto es una evidencia de una deleción en la zona o bien una disomía uniparental [Feuk et al. 2006].

Los *microarrays* son útiles para detectar *CNVs* a lo largo del genoma, pero tienen limitaciones, como su resolución insuficiente para la localización con precisión nucleotídica, o el coste de los *microarrays* de mayor resolución. Por estos motivos, se requiere un análisis más detallado de las secuencias. Las técnicas clásicas como la *PCR* cuantitativa tienen la solución pero su aplicación es muy dirigida, por lo que se necesitan técnicas que permitan genotipar variantes estructurales a un coste asequible y de manera múltiple. Con ese propósito se desarrollaron dos técnicas basadas en la hibridación múltiple de sondas específicas de secuencias diana: la hibridación y amplificación de

múltiples sondas, *MAPH*, [Armour et al. 2000] y la amplificación múltiple de sondas dependiente de ligación, *MLPA*, [Schouten et al. 2002]. Ambas utilizan la *PCR* para amplificar las sondas marcadas fluorescentemente, y la diferencia con la *PCR* cuantitativa reside en que en estas técnicas los cebadores son universales y se cuantifican hasta 40 regiones diana [Sellner et al. 2004].

En la técnica de *MAPH* el ADN se fija en una membrana y se hibrida con las sondas específicas de las secuencias a detectar [Armour et al. 2000]. Estas sondas se generan clonando las secuencias diana y amplificando con cebadores de la secuencia diana hacia el vector, de manera que todas tienen la misma secuencia flanqueante. Se lava la membrana para eliminar las sondas que no han hibridado, luego se separan las sondas específicas de la membrana y se amplifican con cebadores universales. Finalmente se comparan las intensidades de las bandas o la altura de los picos dependiendo de si se separan por electroforesis normal o capilar. La comparación con muestras control permiten detectar deleciones y duplicaciones [Sellner et al. 2004]. En el *MLPA*, el ADN genómico se hibrida con las sondas en solución que tienen dos partes separadas (**Figura 1.8**). Una parte de unos 20-30 nucleótidos es específica de la secuencia diana y contiene una secuencia de unión de cebador universal en el otro extremo. La otra mitad tiene una secuencia específica de la región adyacente de la secuencia diana de unos 25-43 nucleótidos y un sitio de unión de cebador universal en el otro, con la diferencia que incluye un fragmento de tamaño variable, de entre 19-370 nucleótidos, que genera las diferencias entre sondas para separarlas electroforéticamente [Schouten et al. 2002].

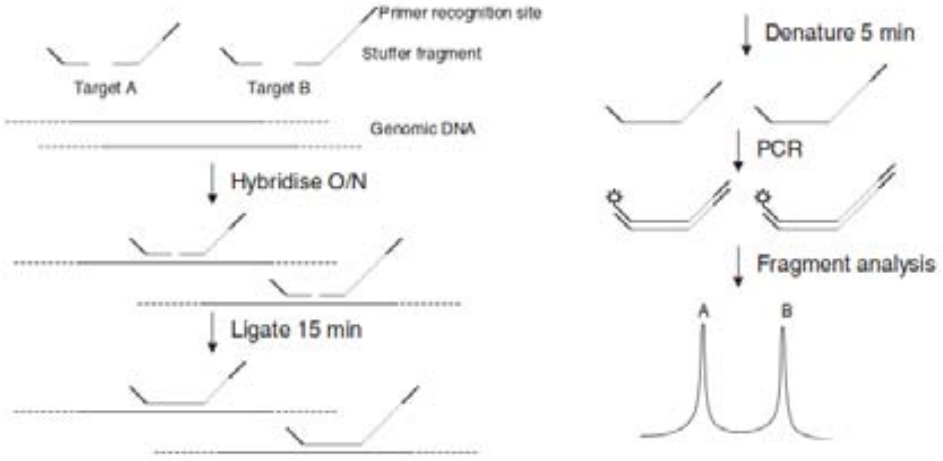

Figura 1.8: Esquema de la técnica de *MLPA*. De arriba a abajo y de izquierda a derecha se muestran los distintos pasos de la técnica, que incluyen la hibridación de las diferentes partes de las sondas, la ligación de éstas, la desnaturalización que permite que se separen del ADN genómico, su amplificación por *PCR* y su análisis. Figura modificada a partir de Sellner et al. 2004.

Las partes específicas de diana de ambas partes de la sonda se unen a las secuencias adyacentes a la región a analizar y entonces se unen por ligación. Se genera una sonda flanqueada por sitios de unión de cebadores universales, que puede ser amplificada por *PCR* mientras que las sondas que no han hibridado no pueden ser amplificadas, y por eso no se necesitan lavados para eliminarlas. La cantidad de sonda ligada es proporcional al *CNV* diana, y las alturas de los picos que se generan en la electroforesis capilar indican las deleciones o duplicaciones [Sellner et al. 2004].

Recientemente se están desarrollando otras técnicas que se basan en la detección de señales visuales y se denominan técnicas de análisis de una sola molécula de ADN. Persiguen la visualización de fragmentos de ADN a gran escala, donde estos fragmentos suelen estar estirados para poder observar su estructura [Alkan et al. 2011]. La técnica más desarrollada en este campo es el mapeo óptico [Teague et al. 2010]. Se basa en los mapas de restricción tradicionales pero realizados sobre ADN inmovilizado, de manera que se pueden identificar los tamaños de los productos de restricción y los cambios en su orden al compararlos con los patrones del genoma de Referencia digerido *in silico*, es decir, computacionalmente [Alkan et al. 2011]. A diferencia de los *microarrays* esta técnica permite detectar variantes balanceadas como inversiones y translocaciones. Se alargan y orientan las moléculas de ADN que se unen a una superficie de vidrio por interacción electrostática. Una vez fijado, el ADN se incuba con una enzima de restricción, luego se tiñe y es revelado por un microscopio de fluorescencia iluminado por un láser de iones de argón que está acoplado a una computadora [Teague et al. 2010]. Una cámara capta las imágenes en el momento que se ilumina la preparación. Todo este proceso es automático y permite analizar entre 50000 y 100000 moléculas al día, lo que equivale a secuenciar un genoma humano con una cobertura de 50 veces en un mes, usando un sólo microscopio. Esta es una de las ventajas de la técnica junto con la detección de todo tipo de variación estructural, incluidas las inversiones.

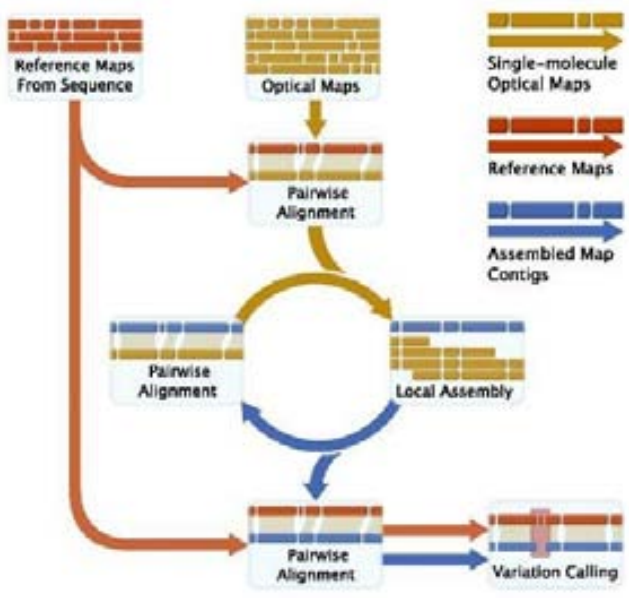

Figura 1.9: Comparación de mapas de restricción en la técnica de mapeo óptico. Se usa un mapa de restricción generado *in silico* con el genoma de Referencia, para comenzar el proceso iterativo de alineamiento de mapas similares por parejas. También inicia el ensamblaje local que genera un mapa consenso a partir de los mapas de una sola molécula de ADN. Después de varias iteraciones, se vuelven a alinear los mapas consenso al mapa de referencia y las diferencias indican variantes estructurales potenciales. Imagen tomada de Teague et al. 2010.

El ordenador almacena las imágenes de las moléculas de ADN digeridas y se comparan con un mapa de fragmentos de restricción de *ADN* generado *in silico* a partir del genoma de Referencia (**Figura 1.9**). Este mapa de referencia se usa para comenzar el proceso iterativo de alineamiento visual de fragmentos por pares, que agrupa juntos los fragmentos similares de una sola molécula; y de ensamblaje local, que genera un mapa óptico consenso a partir de un grupo de fragmentos provenientes de varias moléculas. Después de varias iteraciones, se alinean los mapas consenso generados con el mapa de referencia y se analiza en qué regiones hay diferencias, que son, las potenciales variantes estructurales [Teague et al. 2010].

1.2.2.3 Métodos de secuenciación y mapeo de fragmentos

Un nuevo hito en la historia de las técnicas de detección de variantes estructurales se produjo gracias a las tecnologías de secuenciación de alto rendimiento. Desde la secuenciación del genoma humano la ciencia y la tecnología han avanzado hacia la genómica personal, es decir, hacia lograr una secuenciación de alto rendimiento y bajo coste para que todo el mundo pueda tener la secuencia de su genoma. Si bien este objetivo aún no se ha conseguido, se han mejorado mucho las técnicas de secuenciación, que se encuentran ya en su tercera generación. Se han desarrollado nuevas técnicas que

usan la computación para el mapeo de los *reads* provenientes de la secuenciación de genomas a partir de las tecnologías de secuenciación de nueva generación, *NGS* (del inglés Next Generation Sequencing). Así en el año 2005, Tuzun y colaboradores publicaron el primer estudio relevante de variantes estructurales detectadas mediante la técnica de mapeo de extremos apareados (*PEM*) [Tuzun et al. 2005]. La diferencia de esta técnica respecto a las técnicas de análisis de una molécula de ADN es que no hay un alineamiento visual de fragmentos, sino que se usa la computación para alinear las secuencias de los *reads* que provienen de la secuenciación a un genoma de referencia, por lo que se trabaja sólo con secuencia y no se usa la fluorescencia. Las tecnologías de secuenciación pueden generar dos *reads* que se encuentran a una distancia conocida ya que provienen de los dos extremos de un fragmento de ADN o inserto. En resumen, el mapeo de extremos apareados consiste en generar una librería de fragmentos de ADN o insertos, de los que se secuencian los extremos y se alinean al genoma de Referencia. Como el tamaño de los insertos es conocido, cualquier cambio en la distancia de estas secuencias nos indicará un *CNV* y los cambios en la orientación nos detectarán inversiones. Los insertos que presentan estos cambios se denominan discordantes y los que tienen la distancia y orientación esperadas se denominan concordantes respecto al genoma de Referencia (**Figura 1.10**). Una gran ventaja de esta técnica es que permite detectar una variante con precisión de nucleótido, si se secuencian por completo los insertos.

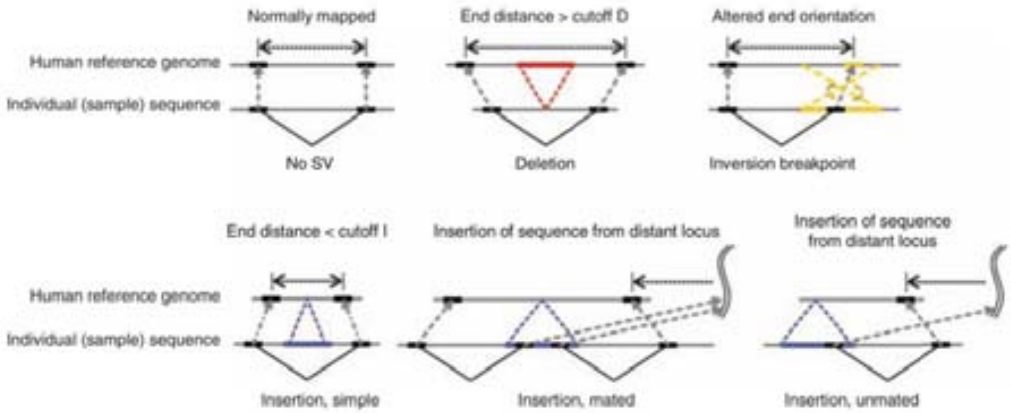

Figura 1.10: Detección de variantes estructurales mediante *PEM*. Están representados el genoma de Referencia arriba y el genoma prueba abajo, los *reads* en negro, las deleciones en rojo, inserciones en azul e inversiones en amarillo. En las deleciones se puede ver como los *reads* mapean a una distancia menor en el genoma prueba. En las inversiones uno de los *reads* mapea en orientación inversa a la esperada. En las inserciones los *reads* mapean a una distancia mayor en el genoma prueba, si proviene de otro lugar del genoma una pareja de *reads* pueden mapear en la secuencia original y en la secuencia adyacente de la inserción. Si dos parejas de reads detectan los puntos de rotura de una de estas inserciones se denominan apareadas. Imagen tomada de Korbel et al. 2007.

Las deleciones se detectan cuando los extremos apareados están más separados en el genoma de Referencia, mientras que las inserciones al contrario, se detectan porque están los extremos más juntos que en el genoma de Referencia. Finalmente las inversiones se detectan cuando una de las secuencias alinea en orientación inversa respecto al genoma de Referencia, porque está dentro de la región invertida, mientras que la otra secuencia alinea en la orientación esperada ya que está fuera de la región invertida [Korbel et al. 2007]. Esta técnica puede detectar variantes estructurales de diferentes tamaños en función del tamaño del inserto que se use. Por otra parte, las limitaciones de *PEM* son que el mapeo único y fiable de los extremos apareados de los insertos se ve dificultado por la repetitividad del genoma humano y que es difícil obtener unos puntos de rotura definidos con buena precisión, ya que requiere la construcción de librerías genómicas de distintos tamaños y que tengan una distribución de tamaños muy controlada, lo cual es muy costoso [Alkan et al. 2011]. Por lo tanto no funcionan de manera fiable para algunas variantes estructurales como las inversiones que tienen sus puntos de rotura en secuencias repetidas. *S*e han realizado estudios de simulación que analizan este problema y que concluyen que hasta un 80% de este tipo de inversiones pueden no haberse detectado [Lledó and Cáceres. 2013].

El potencial de esta técnica es muy grande pero tiene sus limitaciones, por ejemplo que es muy difícil alinear las secuencias apareadas en zonas duplicadas y en general en zonas donde la repetitividad del genoma no permite un alineamiento único, por lo que se pierde una gran parte de la variación estructural del genoma, además de generarse una parte importante de falsos positivos [Sharp et al. 2006] [Lledó and Cáceres. 2013]. Aún así, este método de detección ha revolucionado la detección de variantes estructurales y es muy importante para la detección de variantes balanceadas como las inversiones cromosómicas. Un ejemplo de ello es que en sólo tres estudios que afectan a una decena de individuos se han detectado más de 300 inversiones diferentes en el genoma humano [Tuzun et al. 2005] [Korbel et al. 2007] [Kidd et al. 2008]; y al igual que pasó con los *CNVs* y los *microarrays*, esta cantidad de variación ha hecho que las inversiones se tengan muy en cuenta como responsables de la variación del genoma humano.

Otro método de detección de variantes estructurales a partir de *reads* de secuenciación es el del *read* dividido o *split-read*. Consiste en definir el punto de rotura de una variante estructural a partir del alineamiento de la secuencia de un *read* en el genoma de Referencia y se denomina *read* dividido porque en el caso de detectar el punto de rotura de una variante estructural la secuencia del *read* alinea en dos sitios distintos y en el caso específico de las inversiones, en dos orientaciones distintas. [Mills et al. 2006]. Este método es capaz de definir los puntos de rotura con precisión de nucleótido y se aplicó por primera vez usando *reads* provenientes de secuenciación *Sanger* [Alkan et al. 2011]. Los *reads* provenientes de *NGS* son más cortos y requieren mayor potencia computacional para ser alineados correctamente y aún más sus partes. Se suelen buscar las partes de la secuencia una cerca de otra, reduciendo el coste computacional y en

algunos casos, se usan extremos apareados para limitar el espacio de búsqueda de las partes [Alkan et al. 2011]. El funcionamiento de la técnica es muy parecido al *PEM*, solo que en este caso en vez de tener dos secuencias con un tamaño de separación conocido, tenemos dos partes de una secuencia que alinean en sitios distintos. Si las dos partes están separadas en el genoma de Referencia y no en el de prueba, detectamos una deleción y lo contrario para las inserciones. En el caso de las inversiones, al igual que en el *PEM*, el *read* ha de alinear en el punto de rotura de manera que una parte alinee fuera de la región invertida y la otra dentro, al menos en uno de los dos genomas a comparar, para poder detectar la inversión. Evidentemente existe una gran limitación para detectar las inversiones flanqueadas por duplicaciones segmentales. Finalmente, la detección de duplicaciones en tándem requiere al igual que las inversiones, que los *reads* alineen en un punto de rotura de la parte duplicada para que una parte de la secuencia detecte una copia y la otra parte, la copia siguiente; en el caso de las duplicaciones intersticiales se requieren varios *reads* para detectarlas.

Otro método que utiliza las secuencias de *NGS* para detectar variantes estructurales es la cobertura de *reads*. Es un método indirecto en que se analiza el número de *reads* del genoma prueba en una región y se compara con la del un genoma control secuenciado de la misma forma y las diferencias entre ambos indican duplicaciones y deleciones en el genoma prueba. Las regiones duplicadas tienen una cobertura de secuencias significativamente mayor y las deleciones significativamente menor [Alkan et al. 2011]. En este caso, las variantes balanceadas no pueden ser detectadas. Este método se aplicó por primera vez para detectar reorganizaciones en cáncer [Campbell et al. 2008].

Existen otros métodos de detección que tampoco utilizan directamente las secuencias de *NGS* sino la información derivada. Es el caso del análisis de genotipos de *SNPs*. Se usan los genotipos de un gran número de individuos derivados de proyectos a gran escala como el proyecto HapMap [Gibbs et al. 2003] para detectar por ejemplo deleciones [Conrad et al. 2006] [McCarrol et al. 2006]. En este caso, se analizó la transmisión de los genotipos de *SNPs* entre padres e hijos en tríos familiares. Se buscaron los genotipos no concordantes con una herencia mendeliana, y su genotipo erróneo se relacionó con regiones delecionadas. Además, McCarrol y colaboradores [McCarrol et al. 2006] analizaron las desviaciones del equilibrio de Hardy-Weingberg para determinar las regiones delecionadas y las agrupaciones de genotipos nulos para determinar las regiones delecionadas homocigotas. El equilibrio de Hardy-Weingberg fue definido en 1908 por los autores que le dan nombre. Establece que la composición genética de una población permanece en equilibrio mientras no actúe la selección natural ni ningún otro factor como la mutación. En este estudio se dedujo que, cuando los genotipos de los individuos de una población para un *SNP* no cumplen esta ley, se debe a una deleción ya que se disponía de información de genotipos nulos para estos *SNPs* en algunos individuos. Sus limitaciones son que pueden detectar sólo deleciones y que su resolución depende de la densidad de *SNPs* [Sharp et al. 2006].

Otros métodos de detección de variantes estructurales a partir de los genotipos de *SNPs* se basan en el desequilibrio de ligamiento que se da cuando dos locus segregan juntos, es decir, están ligados. Está relacionado con la falta de recombinación entre ambos bien sea por azar o bien por otros elementos que la bloquean, como pueden ser las inversiones cromosómicas. Usando estas premisas, se desarrolló un método estadístico para detectar inversiones respecto al genoma de Referencia, de un tamaño superior a 200 Kb, presentes en la mayoría de individuos de una población, usando los genotipos de *SNPs* provenientes del proyecto HapMap [Bansal et al. 2007]. Este tipo de métodos tienen como limitación que cualquier región con alto desequilibrio de ligamiento simplemente se detectará como inversión cuando no lo es, simplemente porque no ha recombinado y esto genera una alta tasa de falsos positivos. Un método similar se usó para predecir la frecuencia de inversiones y detectar cuales corresponden al alelo menos frecuente en una población [Sindi and Raphael. 2010], a partir de los haplotipos generados con los genotipos de *SNPs* de individuos Europeos, Africanos y Asiáticos [Sharp et al. 2006]. Actualmente se han desarrollado algunos métodos similares que utilizan datos de *GWAs* [Cáceres et al. 2012].

1.2.2.4 Ensamblaje de novo y comparación genómica

Hasta ahora hemos visto los distintos métodos de detección de variantes estructurales que se basan en la comparación de los productos de secuenciación respecto a un genoma de Referencia, pero un paso más allá, se encuentra la comparación de genomas completos, es decir, la comparación entre genomas ensamblados independientemente. Se pueden detectar todo tipo de variantes con una precisión de nucleótido por alineamiento de ambos ensamblajes [Feuk et al. 2006]. El problema reside en el coste de generar genomas completos y realizar un ensamblaje de calidad sin usar el genoma de Referencia. Además la repetitividad del genoma humano hace que el ensamblaje *de novo* sea más complicado, en comparación con otras especies donde ya se utiliza más ampliamente, como por ejemplo en bacterias. La generación de librerías genómicas, clonaje de *BACs* y secuenciación de extremos es muy cara para aplicarla ampliamente, y es por eso que el camino hacia el ensamblaje de novo pasa por las técnicas de secuenciación de última generación y las mejoras de los algoritmos de ensamblaje [Alkan et al. 2011]. El ensamblaje de novo requiere *reads* grandes para que sea factible y las técnicas de secuenciación de última generación intentan avanzar en esa dirección.

Muchos son los individuos secuenciados pero no ensamblados, como por ejemplo los 1092 individuos secuenciados en el proyecto de los 1000 Genomas [1000 Genomes Project Consortium, 2012]. Varios genomas se han secuenciado mediante *NGS* y ensamblado con la ayuda del genoma de Referencia para ordenar los *contigs* o alinear las secuencias directamente [Bentley et al. 2008] [Wheeler et al. 2008] [Wang et al. 2008] [Ahn et al. 2009] [Fujimoto et al. 2010] [Gupta et al. 2012] [Lilleoja et al. 2012] [Azim et

al. 2013] [Shen et al. 2013]. Sin embargo, ha habido pocos intentos de ensamblar genomas *de novo* y algunos no han tenido la calidad necesaria para que se usen en la detección de variantes [Li et al. 2010]. Sólo se dispone de un genoma ensamblado *de novo* de alta calidad, es el caso del genoma de J. Craig Venter (HuRef) publicado en el año 2007 y secuenciado por secuenciación tradicional *Sanger* con la estrategia de perdigonazo o *shotgun* al igual que el genoma de Referencia de Celera genomics [Levy et al. 2007]. Esta estrategia se basa en la fragmentación del ADN y la secuenciación de los fragmentos, mediante la repetición de este proceso se generan fragmentos secuenciados con distintas terminaciones, y se utiliza un ordenador para ensamblarlos. La secuenciación por *Sanger* genera secuencias mucho más grandes que las obtenidas por *NGS* y esto permite el ensamblaje *de novo*.

La detección de variantes estructurales por comparación de genomas ensamblados *de novo* permite detectar variación estructural con la máxima precisión, es decir, con precisión de nucleótido. Además la comparación genómica es un método no sesgado, es decir que detecta con la misma probabilidad un alelo u otro, debido a que es un método directo. Ambos factores hacen que la detección de variantes por comparación genómica sea mucho más eficaz, es decir, que se detecten una mayor proporción de las variantes existentes entre ambos genomas. Existen algunas aproximaciones menos costosas para descubrir variantes estructurales como las combinaciones entre ensamblaje *de novo* y alineamiento de *contigs* con el genoma de Referencia [Alkan et al. 2011]. También se han hecho ensamblajes locales a partir de fósmidos discordantes secuenciados para descubrir variantes estructurales en 17 genomas humanos [Kidd et al. 2008] [Kidd et al. 2010]. Finalmente, se han aprovechado los ensamblajes existentes para descubrir variación estructural en el genoma humano. Por ejemplo, se comparó el borrador del genoma de chimpancé con el genoma de Referencia humano [Feuk et al. 2005] en busca de inversiones cromosómicas fijadas y se descubrieron 3 inversiones polimórficas en humanos. También se aprovechó el ensamblaje del genoma humano realizado por la empresa Celera Genomics en la carrera por la publicación del genoma humano y se comparó con el genoma de Referencia detectándose numerosas variantes estructurales [Khaja et al. 2006]. Evidentemente también se descubrió variación estructural entre el genoma ensamblado de novo de J. Craig Venter y el genoma de Referencia [Levy et al. 2007], variación que se muestra con mayor detalle en los siguientes apartados.

1.*2.3 Origen y mecanismos de formación*

En el apartado anterior hemos visto cómo detectar la variación estructural pero, ¿cómo se genera? Tenemos una idea estable del genoma pero éste no es inmune a recibir perturbaciones ya sean externas, como las mutaciones provocadas por la radiación ionizante, o bien internas, en forma de errores en la replicación. Para evitar que se acumulen mutaciones, las células tienen una maquinaria que mantiene la integridad del

ADN mediante vías de reparación. Pero los mecanismos de reparación no son perfectos y se dan errores, que forman variantes estructurales en el genoma [Onishi-Seebacher and Korbel. 2011]. También se generan errores durante la recombinación cromosómica que se da en la meiosis y mitosis celular. Además, también pueden generarse variantes estructurales por inserción de secuencia mediada a través de la transposición de elementos móviles.

Los mecanismos de formación de las variantes estructurales pueden clasificarse en los que utilizan secuencias homólogas, que luego encontramos en los puntos de rotura de las variantes, o bien en mecanismos no homólogos o mediados por micro-homología, que no usan secuencias homólogas o utilizan secuencias homólogas muy pequeñas, de alrededor de 10 nucleótidos. Entre los mecanismos por recombinación homóloga encontramos el apareamiento de cadena sencilla, *SSA* (del inglés Single Strand Annealing), y la recombinación homóloga no alélica, *NAHR* (del inglés Non-Allelic Homologous Recombination); y entre los mecanismos no homólogos encontramos distintos mecanismos de reparación que usan secuencias micro-homólogas o no, como la unión de extremos no homólogos, *NHEJ* (del inglés Non Homologous End Joining), la unión de extremos alternativa, *alt-EJ* (del inglés Alternative End Joining), o la unión de extremos mediada por micro-homología, *MMEJ* (del inglés Microhomology-Mediated End Joining), además de mecanismos que dan lugar a reorganizaciones o variantes complejas como son la reparación inducida por rotura mediada por micro-homología, *MMBIR* (del inglés Microhomology-Mediated Break-Induced Repair), el colapso de la horquilla de replicación y cambio de cadena molde, *FoSTeS* (del inglés Fork Stalling and Template Switching), y la cromotrípsis [Onishi-Seebacher and Korbel, 2011]. Los mecanismos homólogos agrupan mecanismos de reparación de las roturas de doble cadena, *DSB* (del inglés Double Strand Break), y de recombinación homóloga, mientras que los mecanismos no homólogos agrupan mecanismos de reparación de *DSB* y replicación del ADN [Lam et al. 2010]. Cada vía de reparación utiliza diferentes proteínas y tiene una eficacia de reparación diferente, de manera que su capacidad para formar mutaciones también es diferente [Conrad et al. 2010]. En los siguientes apartados se muestra el funcionamiento de los mecanismos de formación con mayor detalle.

Una vez que ya conocemos qué mecanismos producen la variación estructural nos podemos preguntar ¿cuándo ocurre? es decir, ¿en qué especie o población se generó una variante estructural determinada? Sólo se conoce el origen de algunas de las variantes estructurales que se han descubierto en el genoma humano, aunque como idea general, podemos entender que hubo un aumento de la complejidad en la estructura del genoma principalmente a raíz de un aumento de las regiones duplicadas del ancestro común de los grandes primates Africanos y se puede relacionar la localización de la variación en el número de copias en los genomas de chimpancé, gorila, macaco y humano con la localización de estas duplicaciones segmentales [Gazave et al. 2011]. Por lo tanto existen variantes estructurales que se encuentran en varias especies y variantes estructurales

específicas de una especie. *Feuk* y colaboradores identificaron 1576 regiones putativamente invertidas entre los genomas de humano y chimpancé, gracias a la comparación de los ensamblajes de ambos genomas [Feuk et al. 2005]. Validaron por *FISH* 5 inversiones en los genomas de humano, chimpancé y gorila, y encontraron que 4 de las 5 eran inversiones entre humano y chimpancé; además de 22 inversiones por *PCR* en humanos y chimpancés resultando en 19 inversiones entre ambos. Además, en tres casos, la orientación del genoma de gorila coincidió con la de chimpancé, de manera que la inversión es específica del genoma humano [Feuk et al. 2005]. Se han realizado otros análisis del estado ancestral de variantes estructurales a gran escala. Por ejemplo se asignó el estado ancestral de 1281 variantes estructurales descubiertas en el genoma humano mediante análisis bioinformático [Lam et al. 2010]. Este análisis se basó en la comparación de las secuencias adyacentes a los puntos de rotura en humanos con las correspondientes en los genomas de chimpancé, orangután y macaco. Para 1141 se asignó la orientación ancestral al estado en el genoma de chimpancé y para las 139 restantes fue asignado su estado ancestral en base a los genomas de macaco y orangután debido a regiones de baja calidad del ensamblaje del genoma de chimpancé [Lam et al. 2010], por lo que representan variantes específicas del genoma humano.

En el mismo estudio se determinaron los posibles mecanismos de formación, a partir del análisis bioinformático de las secuencias flanqueantes de casi 2000 variantes estructurales en el genoma humano. Se clasificó el 45% como resultantes de mecanismos no homólogos, el 21% de mecanismos homólogos, el 21% de inserciones de elementos móviles, el 5% como variación en el número de repeticiones en tándem producidas por el deslizamiento de la horquilla de replicación durante la replicación del ADN, mientras que para el resto de variantes no pudo ser determinado su mecanismo de formación [Lam et al. 2010]. En otro estudio se analizaron 1054 variantes estructurales y el 52.1% se clasificaron como mecanismos no homólogos o micro-homólogos, el 29% como mecanismos homólogos, el 18.9% como transposición de elementos móviles [Kidd et al 2010]. Por lo tanto los mecanismos no homólogos son más comunes, pero no quiere decir que hayan tenido un impacto más grande en la estructura del genoma, ya que la distribución de los mecanismos de formación no es uniforme [Lam et al. 2011].

Una de las observaciones de los estudios que han intentado analizar la importancia de los mecanismos de formación a partir de el porcentaje de variantes que han originado es que los mecanismos homólogos tienden a ser responsables de las variantes de mayor tamaño mientras que los mecanismos no homólogos lo son de las de menor tamaño. Este sesgo se produce porque los mecanismos homólogos están inducidos por la recombinación y los no homólogos por la reparación de errores y ambos procesos tienen tasas diferentes de error [Lam et al. 2010]. Los resultados obtenidos por Conrad y colaboradores [Conrad et al. 2010] en el análisis de *CNVs* dan soporte a esa idea. Concluyeron que la contribución de los mecanismos *NAHR* y *VNTR* es dependiente del tamaño del *CNV*, *NAHR* es más frecuente que *VNTR* entre las variantes más grandes, mientras que *VNTR* es más frecuente

entre las variantes pequeñas [Conrad et al. 2010]. Además en este estudio también se descubrió que las duplicaciones son dependientes de la secuencia flanqueante para su formación mientras que las deleciones no son tan dependientes y por tanto suelen estar formadas por mecanismos no homólogos. Resultados similares se obtuvieron por Mills y colaboradores [Mills et al. 2011], que establecieron los mecanismos no homólogos como dominantes entre las deleciones y la inserción por transposición de elementos móviles como el mecanismo dominante para las inserciones, aunque también detectaron secuencias micro-homólogas correspondientes a mecanismos no homólogos [Mills et al. 2011]. Además concluyeron que el mecanismo *VNTR* solo es responsable de las duplicaciones más pequeñas.

Es importante resaltar que estos resultados no parten de conjuntos de variantes estructurales donde cada clase está igualmente representada y existen sesgos en las técnicas de detección que afectan a los porcentajes finales. Por ese motivo, recientemente se ha usado un conjunto de variantes estructurales que tienen menos sesgos debidos a la detección, porque provienen de la comparación de los ensamblajes del genoma de Referencia y del genoma de J. Craig Venter, HuRef [Levy et al. 2013]. Se asignó el mecanismo de formación para 407365 ganancias, 382510 pérdidas y 117 regiones invertidas. Entre las variantes pequeñas de menos de 1 Kb, el 72.6% estaban producidas por mecanismos no homólogos y el 24.9% por eventos de microsatélites. Las variantes de menos de 10 Kb estaban producidas en un 25.8% por minisatélites y en un 24 % por retrotransposones, mientras que *NAHR* es el mecanismo más frecuente entre las de más de 10 Kb, con un 46.2% [Pang et al. 2013]. En la clasificación por tipos de variantes, los *CNVs* estaban mayoritariamente producidos por mecanismos no homólogos mientras que las inversiones estaban producidas en un 54.7% por mecanismos homólogos. En global, según los resultados de este estudio, los mecanismos no homólogos son responsables de la mayoría de variantes, aunque las variantes de mayor tamaño suelen estar formadas por mecanismos homólogos. Por lo tanto, el sesgo en la distribución de los mecanismos es claro, y el impacto de unos mecanismos u otros sobre la estructura del genoma se entiende mejor dividiendo las variantes por tamaño.

1.2.3.1 Mecanismos dependientes de homología

La recombinación homóloga, es decir, el intercambio de fragmentos entre dos cadenas similares o idénticas de ADN, está detrás de muchos procesos de reparación de roturas y *gaps*, y también es responsable de la correcta segregación de los cromosomas y de la creación de variación alélica en la meiosis mediante la generación de entrecruzamientos. Los mecanismos que dependen de la recombinación homóloga requieren regiones de alta identidad en la secuencia, que van desde unas 50 pb en *Escherichia coli* hasta 300 nucleótidos en mamíferos [Onishi-Seebacher and Korbel. 2011]. Además, la mayoría de mecanismos de recombinación homóloga necesitan una proteína que se encargue del

intercambio de cadena, en procariotas esta proteína es RecA y en eucariotas Rad51. Es necesaria porque uno de los primeros pasos de estas vías es la invasión de la secuencia de doble cadena homóloga por el extremo 3' de la cadena sencilla, y la lleva a cabo esta proteína. En este paso, el extremo 3' de la secuencia de cadena simple substituye al homólogo de doble cadena, es decir, al que correspondería en una replicación homóloga sin errores [Hastings et al. 2009b]. El intercambio entre secuencias que comparten fragmentos significativos de homología puede ocurrir por *NAHR* durante la reparación de *DSB* o en la meiosis, por apareamiento de cadena sencilla *SSA* en la reparación de *DSB* [Quinlan and Hall. 2012], o por deslizamiento de la horquilla de replicación durante la replicación del ADN en el caso de los *VNTR* [Onishi-Seebacher and Korbel, 2011].

Hay que diferenciar tres procesos que implican secuencias homólogas. En primer lugar la recombinación homóloga inducida por *DSB* incluye tres modelos diferentes. Si se repara a partir de ambos extremos adyacentes a la rotura, la reparación puede ser con entrecruzamiento, *dHJ* (del inglés Double Holliday Junction), o sin entrecruzamiento, *SDSA* (del inglés Synthesis-Dependent Strand Annealing), y si se repara a partir de un extremo, suele implicar la reparación de horquillas colapsadas o rotas, denominada replicación inducida por rotura, *BIR* (del inglés Break-Induced Repair). En segundo lugar está la reparación que actúa sobre secuencias repetidas directas y que no usa la recombinación, *SSA*, y por último, el deslizamiento de la horquilla de replicación, *SSM* (del inglés Slipped Strand Mispairing), que se produce gracias a secuencias homólogas. Todos estos procesos usan secuencias homólogas pero no todos la recombinación entre éstas [Hastings et al. 2009b].

NAHR

La recombinación homóloga es la base de los mecanismos que reparan el ADN, y la usan para copiar la parte dañada a partir de una secuencia idéntica. En esta reparación no habrá cambio en la estructura si se usa la secuencia homóloga de la misma posición cromosómica en la cromátida hermana o en el cromosoma homólogo, pero este proceso de reparación comete errores al usar otras secuencias homólogas en posiciones diferentes. A estos errores se les conoce como recombinación homóloga no alélica [Hastings et al. 2009b]. Hay varios mecanismos, de los cuales el más estudiado es el inducido por *DSB*. Existen tres modelos (**Figura 1.11**): la doble unión Holliday, *dHJ*, que puede dar lugar a conversión génica y entrecruzamiento, y que es responsable de la formación de *CNVs* e inversiones; el apareamiento de cadena dependiente de síntesis, *SDSA*, que serviría tanto para evitar los entrecruzamientos como la pérdida de heterocigosidad, *LOH* (del inglés Loss Of Heterozygosity), y que también es responsable de la formación de *CNVs* cuando repara secuencias con repeticiones directas; y por último la replicación inducida por rotura, *BIR* que puede generar *CNVs* si implica secuencias homólogas en posiciones diferentes del cromosoma.

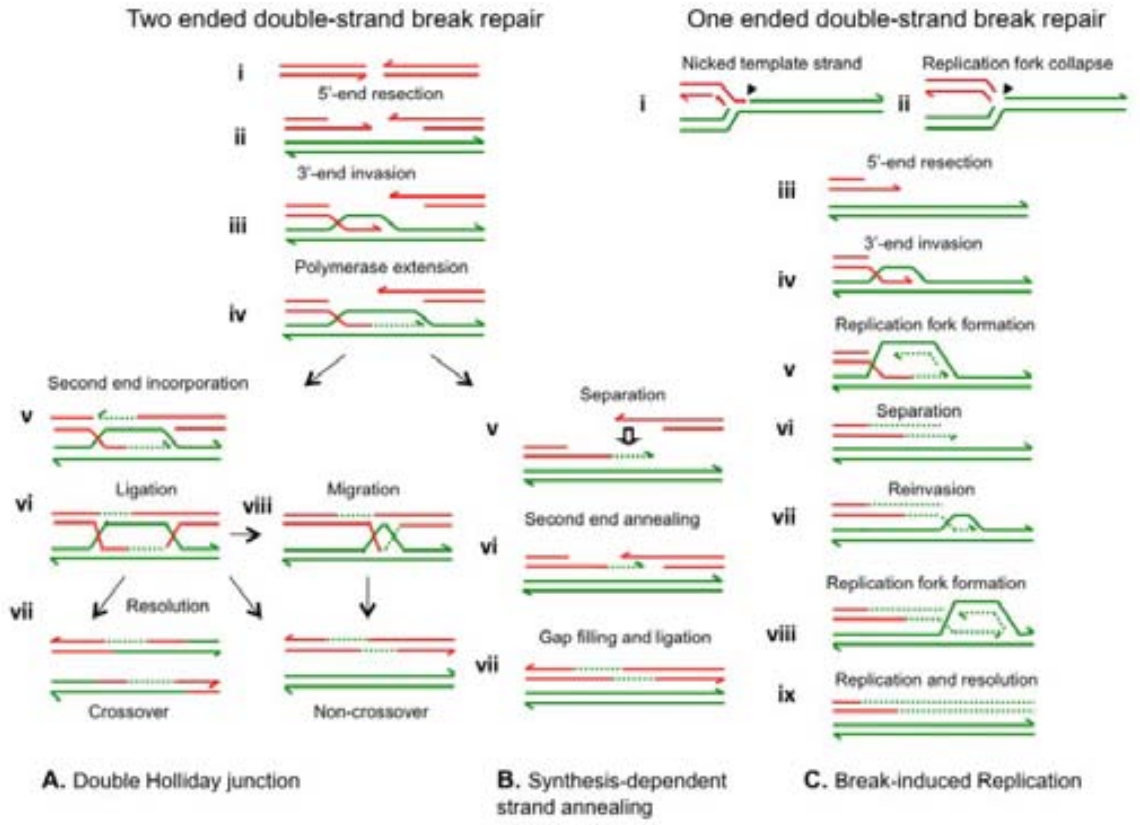

Figura 1.11: Mecanismos de reparación por recombinación homóloga. Se muestran los distintos pasos de las tres vías de reparación mediante recombinación homóloga. Dos de ellas se basan en la reparación a partir de ambos extremos libres generados por un *DSB*, son la doble unión Holliday (A) que puede resolverse con entrecruzamiento o sin él; y el apareamiento de cadena dependiente de síntesis (B) que evita los entrecruzamientos. Por otro lado la vía de replicación inducida por rotura (C) repara el *DSB* a partir de un sólo extremo, también sin entrecruzamientos. Imagen tomada de Hastings et al. 2009b.

Se ha descubierto gracias a las variantes estructurales asociadas a enfermedades, que la recombinación homóloga no alélica entre repeticiones de pocas copias, *LCRs* (del inglés Low Copy Repeats), es decir, duplicaciones segmentales, depende del tamaño de la *LCR*, del grado de identidad, de la distancia entre *LCRs* y de la orientación entre ellas [Stankiewicz and Lupski. 2002]. Además existe una correlación directa entre el tamaño de las *LCR* y la longitud de la variante estructural. Se ha visto que la *NAHR* ocurre en condiciones de más del 97% de identidad entre *LCRs* y una distancia no superior a las 10 Mb. Suponiendo la resolución mediante entrecruzamiento desigual de *dHJ*, podemos deducir cómo se forman las diferentes variantes estructurales. Dependiendo de donde estén localizadas las *LCR* implicadas y en qué orientación, *NAHR* genera un tipo de variantes u otras (**Figura 1.12**). Si están en orientación directa en cromosomas homólogos se producen duplicaciones en tándem y deleciones; por el contrario, si están en orientación invertida se produce una inversión [Stankiewicz and Lupski. 2002].

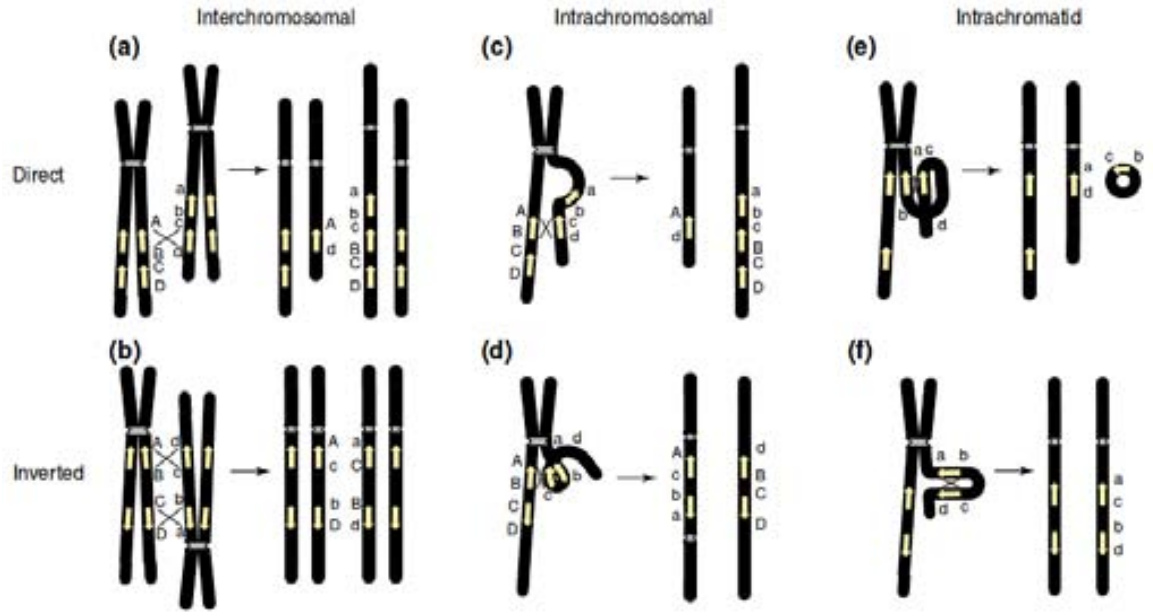

Figura 1.12: Variantes estructurales formadas por *NAHR* entre *LCRs*. Los cromosomas se muestran en negro, sus centrómeros en gris y las *LCRs* en amarillo. Se indican los productos de *NAHR* clasificados según la orientación de las *LCRs* y su localización entre cromosomas, dentro del mismo cromosoma o en la misma cromátida. Si las *LCRs* tienen la misma orientación, *NAHR* dará lugar a duplicaciones y deleciones si están localizadas en diferentes cromosomas o cromátidas (a) y (c), o deleciones y fragmentos acéntricos si están en la misma cromátida (e). Si las *LCRs* tienen diferente orientación, dará lugar a inversiones si están en diferentes cromosomas (b) o en la misma cromátida (f), y cromosomas dicéntricos y acéntricos con duplicaciones y deleciones si están en diferentes cromátidas (d). Figura modificada a partir de Stankiewicz and Lupski. 2002.

Ocurre lo mismo con los entrecruzamientos intracromosómicos e intracromátida. De esta manera *NAHR* es el mecanismo responsable de la creación de la mayoría de las variantes de gran tamaño, duplicaciones, deleciones e inversiones [Stankiewicz and Lupski 2010]. Los mecanismos de reparación por recombinación homóloga aunque son muy importantes para mantener la integridad del ADN, también cometen errores y entra en juego el azar. Existen medidas para evitarlos, como impedir los entrecruzamientos, regular las secuencias que se eligen para la reparación y usar sólo fragmentos grandes de máxima identidad, pero no siempre se pueden llevar a cabo, por ejemplo durante la meiosis los entrecruzamientos son importantes para la correcta segregación de los cromosomas y es en ese momento donde se forman la mayoría de variantes estructurales.

SSA

El mecanismo de apareamiento de cadena sencilla no requiere recombinación entre secuencias homólogas pero se ayuda de ellas para reparar las roturas de doble cadena. Contribuye a la generación de deleciones entre secuencias repetidas que quedan expuestas en la horquilla de replicación (**Figura 1.13**) [Onishi-Seebacher and Korbel. 2011].

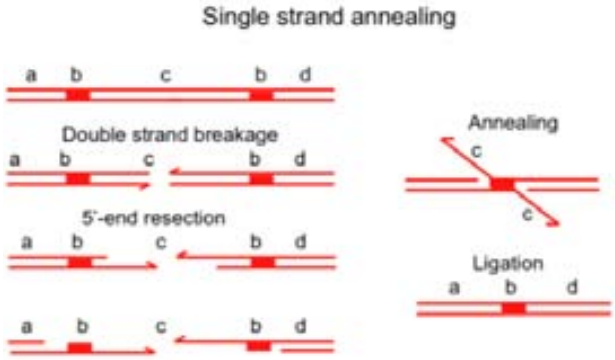

Figura 1.13: El mecanismo de apareamiento de cadena sencilla y cómo genera deleciones. De arriba a abajo y de izquierda a derecha se muestran los distintos pasos que sigue esta vía para reparar un *DSB* y cómo dos secuencias homólogas pueden inducir un error que acaba generando una deleción. Figura modificada a partir de Hastings et al. 2009b.

El apareamiento de cadena sencilla ocurre cuando no hay invasión de la secuencia homóloga por parte de ningún extremo generado en el *DSB*. Se da una erosión de los extremos 5' denominada resección y se generan extremos 3' de cadena sencilla de tamaño considerable. Si en este punto se exponen secuencias complementarias en dos extremos de cadena sencilla, se puede dar el apareamiento. Se eliminan entonces las secuencias de cadena sencilla no apareadas y se da la ligación. En este proceso se deleciona la secuencia localizada entre las secuencias homólogas y también una de ellas. Este mecanismo no invade ningún fragmento de doble cadena, por eso no requiere la proteína *RecA*/*Rad51*, pero sí que usa la proteína *Rad52* para el apareamiento de fragmentos de cadena sencilla. Las deleciones generadas por este mecanismo en humanos suelen tener un tamaño menor a la kilo base, ya que a mayor distancia entre las secuencias homólogas es más difícil que la resección llegue a ambas, por lo que la rotura se repararía por otras vías [Hastings et al. 2009b].

VNTR

La variación en el número de repeticiones en tándem, *VNTR*, se forma debido a un deslizamiento de la horquilla de replicación durante la replicación del ADN [Onishi-Seebacher and Korbel. 2011]. El deslizamiento lo puede provocar una primera repetición en tándem producida por mutaciones puntuales [Tan et al. 2010]. Esto es válido para repeticiones en tándem donde la unidad de repetición es muy pequeña, pero a medida que la unidad de repetición es más grande, es mucho menos probable encontrar repeticiones en tándem formadas al azar que inicien la expansión de las copias. En las repeticiones en tándem con una unidad mayor, las *LCRs* sirven para iniciar este proceso mediante el deslizamiento por mal apareamiento de cadena, *SSM*. Las *LCRs* facilitan el mal apareamiento cuando las cadenas están separadas causando una duplicación o deleción de la secuencia entre ellas (**Figura 1.14**). Por lo tanto en este caso el mecanismo

de generación es el *SSM* y como consecuencia se inicia un proceso de *VNTR*.

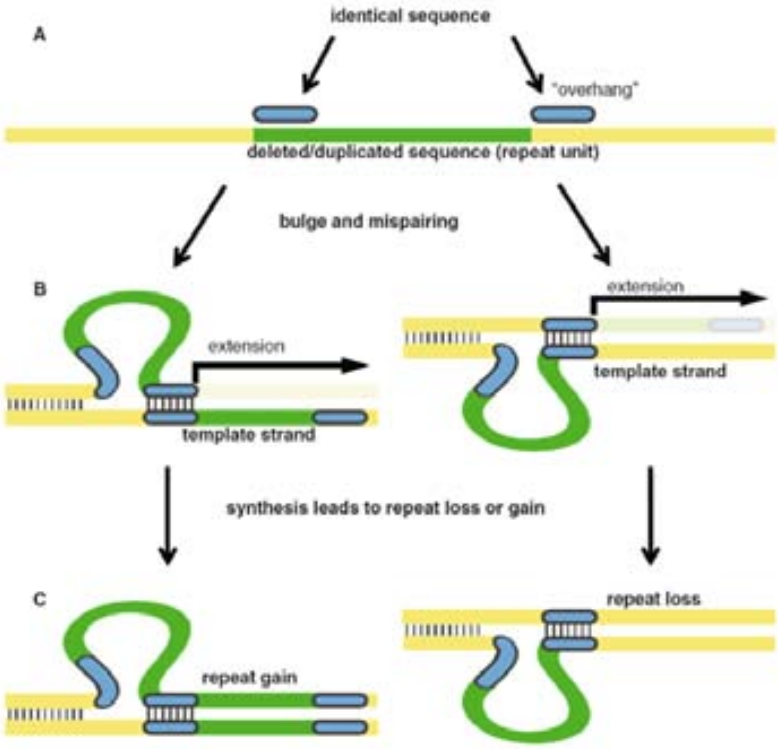

Figura 1.14: Deslizamiento de la horquilla de replicación por mal apareamiento de cadena y generación de *VNTR*. Se muestra cómo el mecanismo *SSM* puede generar *VNTR* a partir de una secuencia que está fuera de la repetición pero que es homóloga de una parte de la secuencia de dentro de la repetición. (a) Se muestran en azul las secuencias homólogas, la repetición no se ha generado todavía. (b) La secuencia azul que está fuera se puede aparear con su homóloga dentro de la unidad a repetir y se genera un lazo. (c) La síntesis de ADN da lugar a una duplicación o deleción de la secuencia. Si se duplica puede dar lugar a un *VNTR*. Imagen tomada de Tan et al. 2010.

Mediante este mecanismo, una secuencia que no estaba repetida en tándem puede constituir un *VNTR* o ser delecionada. El deslizamiento y mal apareamiento puede darse entre *LCRs* de kilo bases de tamaño, y una vez generada la primera copia en tándem ya no se necesitan éstas para inducir más ciclos de repetición [Tan et al. 2010].

1.2.3.2 Mecanismos no homólogos o micro-homólogos

Estos mecanismos, como su nombre indica, no usan secuencias homólogas o bien usan secuencias homólogas muy pequeñas de menos de 10 pb, que se conocen como secuencias micro-homólogas. Se clasifican los procesos en replicativos y no replicativos [Hastings et al. 2009b]. Los procesos no replicativos se basan en la reparación de *DSB*, como la unión de extremos no homólogos, *NHEJ*, o su versión alternativa, la unión de extremos mediada por micro-homología, *MMEJ*, y son responsables de la formación de

CNVs, inversiones y translocaciones. Los procesos replicativos son más complejos, normalmente están relacionados con la horquilla de replicación, y son los responsables de la formación de variantes complejas. Entre ellos se encuentran la reparación mediada por micro-homología e inducida por rotura, *MMBIR*, el colapso de la horquilla y el cambio de cadena molde, *FoSTeS*, y la cromotrípsis [Onishi-Seebacher and Korbel. 2011]. En los mecanismos no homólogos o micro-homólogos es muy difícil asignar un mecanismo de formación a partir de las secuencias flanqueantes de una variante estructural. Si bien es cierto que los múltiples puntos de rotura de las variantes complejas nos indican procesos replicativos, en presencia de un solo par de puntos de rotura cualquiera de los mecanismos no homólogos o micro-homólogos puede ser el responsable, exista o no micro-homología en las secuencias flanqueantes [Onishi-Seebacher and Korbel, 2011].

NHEJ

El mecanismo de unión de extremos no homólogos se basa en la reparación de *DSB* sin la utilización de secuencias homólogas, o mediante la utilización de secuencias micro-homólogas de hasta 10 nucleótidos. Hasta el descubrimiento de los procesos replicativos, ha sido el mecanismo más importante para explicar la variación estructural sin presencia de secuencias homólogas mayores de 10 pb alrededor. En esta vía de reparación se detectan los *DSB*, se genera un puente molecular entre los extremos sueltos, se modifican estos extremos para que sean compatibles y por último se ligan (**Figura 1.15**). En este proceso se suelen añadir nucleótidos adicionales y esto deja una huella genómica [Stankiewicz and Lupski 2010].

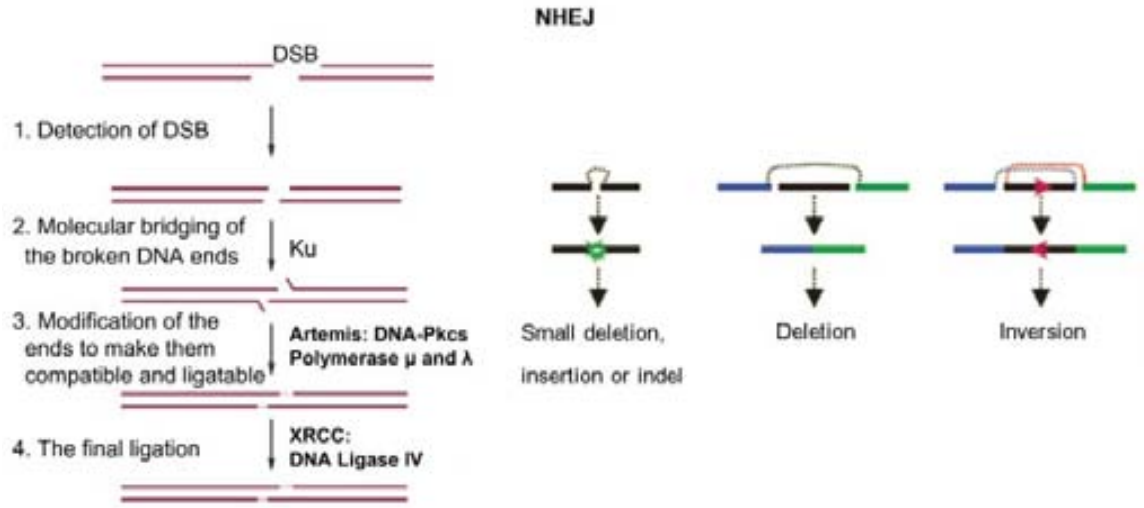

Figura 1.15: Esquema del mecanismo *NHEJ* y de la generación de variantes estructurales. En la parte izquierda está representado el proceso de *NHEJ*. Se muestra el *DSB* y las dos cadenas de ADN de color morado. Se detecta el *DSB*, se aparean los extremos de las secuencias complementarias y se añaden nucleótidos a los extremos para que se puedan ligar, generándose una huella del proceso. Finalmente se ligan [Gu et al. 2008]. En la parte derecha se muestra cómo este proceso genera variantes estructurales al reparar uno o varios *DSBs*. Figura modificada a

partir de Chen et al. 2011.

NHEJ puede unir los extremos del *DSB* de manera limpia, pero también puede dar lugar a deleciones, inserciones e inversiones. Si se produce un único *DSB*, se puede reparar correctamente o generar pequeñas deleciones de 1 a 4 nucleótidos y en algunos casos pequeñas inserciones. Si existe más de un *DSB*, entonces se pueden generar deleciones e inversiones de un tamaño mayor, del orden de cientos de pares de bases o kilo bases, dependiendo de lo que le ocurra al fragmento localizado entre las dos roturas de doble cadena [Chen et al. 2011].

MMEJ/altEJ

Es una vía alternativa de reparación de extremos no homólogos que utiliza secuencias de micro-homología de entre 5 y 25 pb. La reparación está dirigida por la micro-homología, cosa que puede prestar a confusión ya que el *NHEJ* también usa secuencias micro-homólogas, aunque de menor tamaño. En el *MMEJ* estas secuencias micro-homólogas aparean en los extremos del *DSB* y dan lugar a deleciones de la secuencia localizada entre ellas [Hastings et al. 2009]. Básicamente no hay diferencias en cuanto a las variantes que se pueden formar [Conrad et al. 2010]. La diferencia está en la longitud de las secuencias micro-homólogas, que en el *MMEJ* son más largas y por eso aparean, de manera que no se usan las proteínas de unión a los extremos Ku70 y Ku80, que sí necesita el *NHEJ*. Estas diferencias en el uso de la maquinaria proteica sugieren que se trata de una vía alternativa. El uso de proteínas también diferencia al *MMEJ* del mecanismo homólogo *SSA* ya que el primero no usa la proteína Rad52 de apareamiento de cadenas sencillas mientras que para *SSA* es vital.

MMBIR/FoSTeS

Los modelos replicativos que usan secuencias micro-homólogas incluyen dos modelos que funcionan de la misma manera o hacen referencia al mismo proceso, la reparación mediada por micro-homología e inducida por rotura, *MMBIR*, y el colapso de la horquilla y el cambio de cadena molde, *FoSTeS*. En ellos, una horquilla de replicación se detiene o se colapsa y esto lleva a cambios de cadena molde guiados por secuencias micro-homólogas (**Figura 1.16**). Estos cambios se dan entre el extremo 3’ de la cadena de nueva síntesis y secuencias micro-homólogas en otros lugares del cromosoma. Si el cambio de cadena molde se da por delante o por detrás de la horquilla colapsada se formarán deleciones o duplicaciones, respectivamente, y si se da en orientación invertida, inversiones. Se producen variantes estructurales complejas cuando se dan múltiples cambios de cadena molde a partir de una única horquilla de replicación. Si se implican otros cromosomas entonces se producen translocaciones. Además pueden darse entre secuencias muy lejanas, incluyendo por ejemplo, brazos enteros de cromosomas, ya que ocurren en el núcleo, dónde partes lejanas del mismo u otros cromosomas están próximas

[Lee et al. 2007] [Quinlan and Hall. 2012].

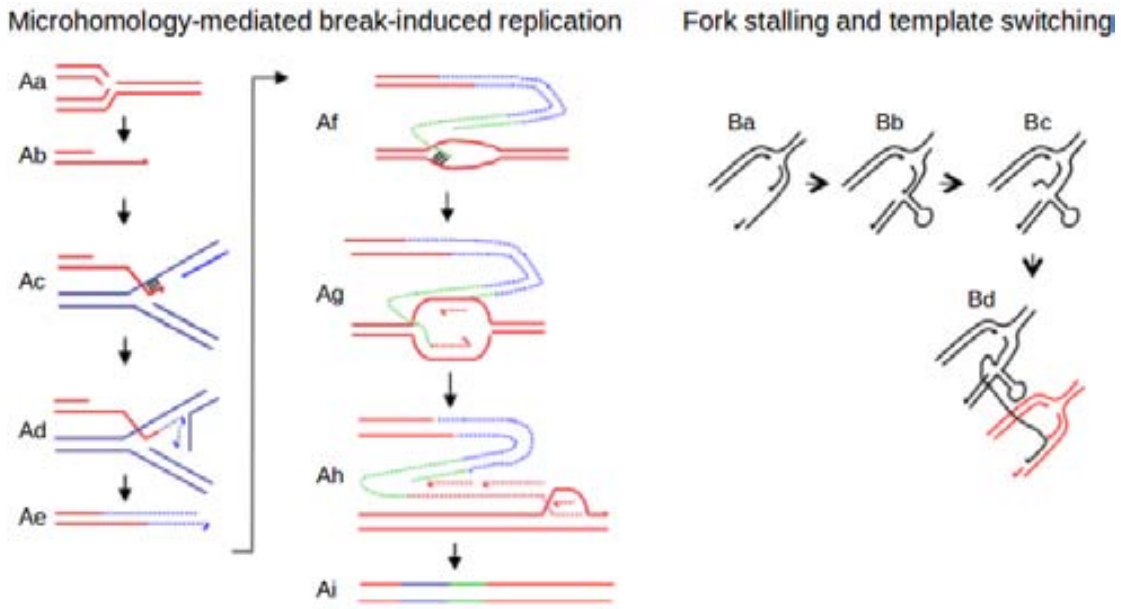

Figura 1.16: Esquema del mecanismo MMBIR/FoSTeS. En la parte izquierda está *MMBIR*. (Aa) Se da un colapso de la horquilla por una rotura de cadena simple, por ejemplo. Uno de los brazos se separa de la horquilla. (Ab) Hay resección y se expone un extremo 3'. Éste puede aparear con una secuencia micro-homóloga de otra horquilla. (Ac) Finalmente estos extremos 3' pueden invadir varias horquillas formando todo tipos de variantes. (Ad) El extremo libre 3' se extiende (secuencia azul). (Ae) Se separa de la secuencia molde. (Af) Se convierte de nuevo en un extremo 3' que puede aparear con otra secuencia micro-homóloga de cadena sencilla. (Ag) En este caso, la horquilla de replicación progresa. (Ah) Continúa hasta el final del cromosoma y la molécula final tiene secuencias de otras localizaciones genómicas. (Ai) Dependiendo de la posición en la que la síntesis vuelve al cromosoma original (en rojo) relativa a la posición donde se inició el colapso de la horquilla original, se formarán duplicaciones, deleciones o inversiones. En la parte derecha está representado *FoSTeS*. (Ba) La cadena simple queda expuesta y adquiere una estructura secundaria. (Bb) Colapsa la horquilla de replicación, quedando libres los extremos 3'. (Bc) Éstos pueden aparear con otras cadenas simples expuestas en otras horquillas que contengan secuencias micro-homólogas. (Bd) La migración de los extremos 3' genera duplicaciones, deleciones e inversiones dependiendo de la posición relativa de la segunda horquilla de replicación. Imagen modificada a partir de Hastings et al. 2009b.

Cromotrípsis

La cromotrípsis hace referencia a la gran cantidad de variantes estructurales complejas que se han detectado en muchos cánceres. Se postula como una catástrofe cromosómica, ya que incluye muchos *CNVs* y agrupaciones de puntos de rotura concentrados en un sólo cromosoma [Stephens et al. 2011]. Se han detectado estas agrupaciones de variantes complejas en el 2% de todos los cánceres y hasta en un 25% de los cánceres de hueso. Además, estos acúmulos de variantes complejas se dan en poco tiempo, descartando su producción por acumulación de mutaciones independientes. Stephens y colaboradores

postularon en el año 2011 que estas variantes complejas se generan en un único evento catastrófico que denominaron cromotrípsis [Stephens et al. 2011]. El modelo propone dos pasos sencillos, la disgregación en un solo evento de un cromosoma en fragmentos y su reparación errónea posterior. Las diferencias entre las variantes complejas que formarían parte de la cromotrípsis y las de individuos sanos son básicamente el mayor tamaño y complejidad de las variantes de procesos cancerígenos, pero los patrones son muy similares [Quinlan and Hall, 2012]. Este hecho, junto con la falta de mecanismos moleculares para explicar la cromotrípsis, hace pensar que en esencia la cromotrípsis no es más que el resultado de la generación de variantes complejas de manera masiva que se da en los procesos cancerígenos, por parte de los procesos replicativos *FoSTeS*/*MMBIR* [Liu et al. 2011].

1.2.3.3 Variantes únicas y de origen recurrente

Hemos visto cómo se forman las variantes estructurales y de manera innata podemos pensar que ocurren una vez en la historia evolutiva del genoma. Por ejemplo, las duplicaciones presentes en el genoma de primates se dieron en un punto de la evolución, en una especie y todas las especies que surgieron de ésta, contienen esas duplicaciones modificadas en menor o mayor grado. En este caso se trata de variantes únicas. Pueden ser monofiléticas, cuando afectan a una rama evolutiva a partir de un ancestro común, o polifiléticas, cuando especies de ramas diferentes del árbol evolutivo contienen la variante, que se ha mantenido en polimorfismo. Por otro lado, las variantes estructurales pueden tener un origen múltiple, es decir, haberse originado más de una vez. Se denominan variantes de origen recurrente. Siguiendo con la situación anterior, encontraríamos las duplicaciones de primates en dos ramas diferentes del árbol evolutivo pero no en su ancestro común, porque habrían aparecido de manera independiente en especies de ambas ramas. Este sería un ejemplo de variantes recurrentes polifiléticas. Una variante estructural recurrente monofilética se daría cuando detectamos una variante estructural que se ha revertido y generado varias veces en la historia evolutiva de una determinada especie. Un ejemplo de inversión cromosómica recurrente y polifilética es la inversión *Xq28* presente en la mayoría de mamíferos. En el año 2007, Cáceres y colaboradores demostraron que esta inversión ha ocurrido de manera independiente en diferentes linajes de euterios (mamíferos placentarios) [Cáceres et al. 2007]. En humanos, se conoce la recurrencia de inversiones asociadas a enfermedades como la hemofilia [Bagnall et al. 2002], síndrome RCAD y otros síndromes de microdeleción en varios cromosomas [Antonacci et al. 2009]. Además varios estudios han demostrado la recurrencia de inversiones polimórficas en todos los cromosomas [Flores et al. 2007] [Aguado et al. 2014].

La recurrencia de las variantes estructurales está directamente relacionada con las características de las secuencias flanqueantes y a su vez con los mecanismos de formación. Si pensamos en la ocurrencia de una misma variante de manera independiente en dos especies de linajes diferentes (variante polifilética) o en especies diferentes de un mismo linaje, la probabilidad de originarse a partir de mutaciones puntuales es muy muy pequeña, por eso es difícil que se hayan dado por errores en las vías de reparación. Una vía capaz de explicar este proceso es la recombinación homóloga no alélica. Se ha demostrado la recurrencia en origen de variantes estructurales que muestran secuencias homólogas alrededor de sus puntos de rotura [Aguado et al. 2014] y están formadas por *NAHR*. Los entrecruzamientos que se dan durante la recombinación entre *LCRs* de gran tamaño e identidad pueden generar variantes estructurales en el mismo lugar del genoma de especies diferentes, siempre y cuando se mantengan los *LCRs* [Cáceres et al. 2007] [Zody et al. 2008].

También se denominan variantes recurrentes las variantes que tienen puntos de rotura iguales entre diferentes individuos con enfermedades genéticas [Gu et al. 2008]. Debido a la variación de los puntos de rotura que se encuentra en estas variantes relacionadas con síndromes concretos, se definen como recurrentes las variantes que muestran puntos de rotura agrupados y que han ocurrido de forma independiente en muchos pacientes; por el contrario, las variantes no recurrentes tienen tamaños diferentes entre los pacientes y comparten pequeñas regiones de solapamiento. Además esta variación también afecta a *CNVs* y se ha relacionado con características clínicas de los pacientes [Gu et al. 2008].

Las variantes recurrentes se pueden detectar mediante el genotipado de individuos de diferentes especies del mismo linaje y también a partir del análisis de la variación nucleotídica. Cuando se genera una variante estructural, captura una serie de alelos de los *SNPs* que hay en la secuencia afectada y si esta variante evita que se dé recombinación en la secuencia, mantiene el haplotipo que se ha generado. Éste se irá diferenciando con el tiempo del haplotipo original, ya que ambos acumularán cambios puntuales diferentes entre ellos. Este efecto es claro en las inversiones cromosómicas, ya que evitan la recombinación entre ambas orientaciones en los individuos heterocigotos. Por lo tanto, si una variante aparece una única vez, podremos ver diferencias entre el haplotipo de la variante y el haplotipo de la secuencia sin variante e incluso pueden tener mutaciones propias como los *SNPs* marcador. Por otro lado, si la variante es recurrente, al originarse y revertirse, se romperá la diferenciación de los haplotipos. Otra consecuencia de la recurrencia es que las variantes no pueden ser genotipadas a partir de *SNPs* ya que su asociación se rompe cada vez que la variante revierte o se vuelve a generar. Además sus efectos no se tienen en cuenta cuando se realizan estudios de asociación de enfermedades y *SNPs* a nivel del genoma completo, *GWAS*, por lo que su estudio experimental es necesario para determinar su papel en las causas genéticas de enfermedades y síndromes.

En el mayor estudio sobre inversiones cromosómicas recurrentes hecho hasta el momento [Aguado et al. 2014], se detectaron 4 inversiones cromosómicas recurrentes, todas ellas sin cambios nucleotídicos propios del haplotipo de la secuencia invertida o estándar y con muchos cambios compartidos entre haplotipos, además de 6 inversiones en las que la recurrencia se demostró por genotipación de primates. Todas las inversiones recurrentes de este estudio están flanqueadas por repeticiones invertidas (RIs), sugiriendo la recombinación no alélica como mecanismo de formación.

1.3 Variación estructural interespecífica

Se denomina variación estructural interespecífica la variación estructural que diferencia genomas de especies distintas. Se ha estudiado mayoritariamente en el género *Drosophila*, y el primer ejemplo lo descubrieron Dobzhansky y colaboradores, en 3 inversiones paracéntricas que diferencian las especies *D. pseudoobscura* y *D. persimilis* que son morfológicamente idénticas [Dobzhansky et al. 1944]. Además se han estudiado también en detalle las diferencias entre el genoma humano y los genomas de grandes simios, principalmente nuestro pariente vivo más próximo, el chimpancé. Las primeras variantes microscópicas descubiertas fueron derivadas de la comparación de los cariotipos de ambas especies y se corresponden con 6 inversiones pericéntricas, una fusión telomérica, 4 deleciones / inserciones dentro del cromosoma *16*, deleciones o inserciones terminales y la variación de una parte importante de la heterocromatina [Lejeune et al. 1973]. También se compararon los cromosomas humanos con los de otras especies de grandes simios como el gorila o el orangután y se estableció que el orangután es la especie más similar y cercana al ancestro común de grandes simios y humanos [Dutrillaux, 1979]. Con la mejora de las técnicas de tinción, las comparaciones entre los genomas de grandes simios y humanos resultaron en más variantes estructurales y mejor definidas: por ejemplo se refinaron las diferencias entre humanos y chimpancés a 9 inversiones pericéntricas, la adición de heterocromatina telomérica en el cromosoma *18* de chimpancé y diferencias en la cantidad y localización de la heterocromatina [Yunis et al, 1980]. Con la llegada de la secuencia del genoma humano y la secuencia borrador del genoma de chimpancé, las comparaciones de la estructura abordaron la variación estructural submicroscópica y sus implicaciones en la especiación humana y de los grandes simios.

1.3.1 CNVs

Anteriormente se ha comentado el auge en el descubrimiento de variantes estructurales entre las especies del género *Drosophila*, especialmente inversiones, no obstante también hemos visto que la llegada de la secuencia del genoma humano y de las principales especies de primates promovieron el análisis de la variación entre primates y humanos.

Este paso coincidió con la consideración por parte de la comunidad científica que la duplicación génica es el mecanismo principal que usa la evolución para generar nuevos genes y procesos biológicos. Por eso se intentaron relacionar las diferencias fenotípicas existentes entre humanos y primates con la presencia de genes duplicados, de ahí la cantidad de estudios comparativos entre el genoma humano y el de chimpancé.

Se comenzó por intentar detectar genes duplicados en la especie humana que no lo están en primates, para explicar mediante variantes estructurales las características fenotípicas que nos hacen humanos. Fortna y colaboradores en el año 2004 usaron la técnica *aCGH* con ADN complementario como sonda representando 29,619 genes humanos sobre el genoma de chimpancés, gorilas, orangutanes y bonobos, para identificar genes que se han duplicado específicamente en cada linaje [Fortna et al, 2004]. Se identificaron 1005 genes no duplicados al menos en una de las especies y se determinó que existe una tendencia a la duplicación de genes en humanos y que varios de los genes involucrados están relacionados con la estructura y la función del cerebro [Fortna et al. 2004]. Otro estudio usó la técnica *array-CGH* esta vez con *BACs* representando todo el genoma humano, contra el genoma de chimpancé y de gorila [Wilson et al. 2006]. Se identificaron 63 regiones genómicas con mayor señal en humanos que en chimpancé y gorila, por lo que se trata de duplicaciones específicas de la especie humana. Entre las regiones duplicadas se encontraron genes relacionados con inmunoglobulinas e histonas mayoritariamente [Wilson et al. 2006]. En un tercer estudio los investigadores usaron la misma técnica que Fortna y colaboradores en el 2004 para extender la comparación a 10 especies de primates y analizar así las ganancias y pérdidas de ADN a lo largo de este linaje, donde nos encontramos los humanos [Dumas et al. 2007]. De los 24473 genes humanos que se analizaron, 4159 están afectados por *CNVs* específicos de linaje; y casi un tercio de los genes presentan *CNVs* en alguno de los genomas de primate [Dumas et al. 2007]. En otro estudio similar se encontraron 23 genes humanos más, afectados por *CNVs* específicos de humanos, además de *CNVs* específicos de chimpancés, gorilas y orangutanes [Armengol et al. 2010]. Se analizaron por *PCR* cuantitativa 11 de estos genes en las especies comparadas, además de 6 especies más de primates. Como resultado se confirmaron 4 genes duplicados en humanos (*ABCB10*, *E2F6*, *CDH12*, y *TDG*) como candidatos a determinar características fenotípicas humanas [Armengol et al. 2010].

Por otro lado también se pusieron como objetivo los genes duplicados en varias especies, para buscar las características comunes que compartimos con los grandes simios y primates. Por ejemplo Perry y colaboradores en el año 2006 analizaron los *CNVs* presentes en el genoma de 20 chimpancés no relacionados y los compararon con los *CNVs* en el genoma humano. De esta manera encontraron *CNVs* compartidos entre humanos y chimpancés, que son frecuentes en las poblaciones de ambas especies. Los autores propusieron que existen puntos calientes de generación de *CNVs* que se comparten entre especies y que el proceso está dirigido por *NAHR* entre duplicaciones segmentales ancestrales [Perry et al. 2006].

Aunque mayoritaria con chimpancé, la comparación también se extendió a otros grupos de primates. Se realizó también un estudio para identificar *CNVs* de manera intraespecífica en los genomas de grandes simios (chimpancés, gorilas, orangutanes y bonobos), para luego comparar la localización y frecuencia de los *CNVs* en los grandes simios y humanos [Gazave et al. 2011]. En este estudio, a diferencia de los anteriores, se analizó el polimorfismo de *CNVs* en cada especie para luego realizar la comparación, y se obtuvieron frecuencias de *CNVs* en grandes simios. Se detectaron mediante *aCGH*, en una primera fase se usaron *BACs* como sondas para descubrir los *CNVs* y en un segundo paso se refinaron sus puntos de rotura usando sondas de oligonucleótidos. Se concluyó que la mayoría de *CNVs* no son específicos de especie sino que están compartidos en dos o más especies. Serían resultado de una alta actividad de duplicación de segmentos por *NAHR*, que facilitaría la pérdida o creación de nuevas copias. Se observó que en bonobos, chimpancés y gorilas hay un enriquecimiento de *CNVs* en lugares con duplicaciones segmentales conocidas, que tienen su origen en el ancestro de los grandes simios [Marques-Bonet et al. 2009]. Es un claro ejemplo de cómo la variación estructural de un genoma ancestral ha determinado la de las especies derivadas.

Por otra parte se está avanzando en el análisis de los *CNVs* a partir de el uso de las nuevas plataformas de secuenciación. Aunque se ha comenzado por analizar la variación entre humanos, un paso a seguir es analizar la variación entre especies [Alkan et al. 2009].
En resumen, actualmente se siguen realizando estudios comparativos entre la especie humana y especies de primates para detectar los *CNVs* que pueden determinar las características fenotípicas que nos hacen humanos, bien sea para encontrar los genes duplicados en la especie humana o bien para detectar que características comunes con los primates podemos relacionar con los *CNVs* que encontramos en su genoma y en el nuestro. Evidentemente la gran cantidad de variantes estructurales que se han descubierto en estos estudios comparativos hace pensar que han tenido un papel importante en la evolución de la especie humana. Se han identificado variantes estructurales específicas de humanos que afectan a genes y parece poco probable que todas ellas sean cambios neutros, por lo que son candidatas a haber contribuido a la evolución de nuestro genoma [Kehrer-Sawartzki et al. 2007].

1.3.2 Inversiones

Las inversiones cromosómicas fueron las primeras variantes estructurales que se estudiaron en detalle [Dobzhansky et al. 1944]. Se han estudiado de manera intensiva en el género *Drosophila* y se han detectado un gran número de inversiones tanto fijadas como polimórficas.

En primates y humanos, se descubrieron mediante técnicas citogenéticas 9 inversiones diferenciando los genomas de humanos y chimpancés [Yunis et al. 1980]. De éstas, las de los cromosomas *1* y *18* son específicas del genoma humano y por lo tanto candidatas a haber producido cambios evolutivos solamente en el linaje humano. Hasta ese momento la contribución a la especiación por parte de las inversiones se atribuía a la esterilidad de híbridos o al mantenimiento de complejos de genes coadaptados [Lowry and Willis. 2010], y aunque se sabía de su potencial adaptativo a partir de la inhibición de la recombinación, no se les dio la misma importancia en primates que en el género *Drosophila*. El modelo de duplicación génica fue considerado el mecanismo central del cambio evolutivo [Dumas et al. 2007] y por eso, los estudios sobre especiación se centraron en los *CNVs*. También contribuyó a ello la dificultad en la detección de las variantes balanceadas, el descubrimiento de la gran cantidad de duplicaciones en el genoma del ancestro de los grandes simios y la repetitividad del genoma humano.

Algunos investigadores continuaron con los trabajos sobre el aislamiento de las especies a partir de las inversiones, esta vez entre humanos y chimpancés, y generaron un modelo para explicar la especiación humana a partir del aislamiento reproductivo y la selección natural [Navarro et al. 2003]. En este modelo cambios seleccionados positivamente se acumularían en los cromosomas con diferencias estructurales fijadas. Se generaría con ello una barrera genética por la acumulación de cambios incompatibles que producirían el aislamiento reproductivo y por lo tanto la especiación [Navarro et al. 2003]. En el caso de las inversiones, una vez se generadas entre poblaciones, se espera que acumulen mutaciones que afecten a genes que causan incompatibilidad entre especies. Esta es una de las hipótesis para explicar el aislamiento reproductivo en moscas, genes involucrados en el aislamiento precigótico y postcigótico se encontrarían en inversiones que diferencian especies cercanas [Kirkpatrick et al. 2010].

Después que se descubriese la gran cantidad de *CNVs* en el genoma humano, un estudio comparado entre el genoma humano y el genoma de chimpancé resultó en el descubrimiento de 1576 potenciales inversiones cromosómicas entre ambos genomas [Feuk et al. 2005]. Se validaron experimentalmente 23 de ellas mediante experimentos de *FISH* y *PCR*. Éstas se genotiparon en un panel de individuos humanos y 3 de ellas resultaron polimórficas. En resumen este estudio demostró la gran cantidad de inversiones que se han dado desde la separación de ambas especies hace 6 millones de años y la implicación de las inversiones en el cambio evolutivo, debido su capacidad para afectar genes específicos, más allá de la inhibición de la recombinación producida por las grandes inversiones microscópicas conocidas hasta el momento.

En otro estudio comparado se utilizó *PEM* para evitar los falsos positivos generados por la baja calidad del genoma de chimpancé y se detectaron 174 inversiones más entre ambos genomas [Newman et al. 2005]. Este auge de las inversiones hizo que se diseñasen nuevos métodos bioinformáticos para detectarlas e incluso que se usasen como caracteres

evolutivos para reconstruir la filogenia de las especies de mamíferos y primates, como ya se había hecho con el género *Drosophila*. [Chaisson et al. 2006]. Las evidencias de la participación de las inversiones en la especiación fueron creciendo y las hipótesis de cómo se produce el aislamiento reproductivo fueron diversificando. Actualmente, se considera que las inversiones han tenido y tienen un papel muy importante en la evolución de las especies al igual que los *CNVs*.

Existen varios ejemplos de inversiones que han mediado el aislamiento reproductivo de poblaciones participando en su proceso de especiación. Uno de estos ejemplos es el de una inversión cromosómica implicada en la especiación de una planta, la flor mono amarilla, *Mimulus guttatus* [Lowry and Willis. 2010]. Esta especie vive en el oeste de Norte América y su distribución geográfica es muy amplia. Además puede presentar dos formas diferentes en cuanto a su ecología, la forma anual y la forma perenne, por lo que ocupa dos ecotipos diferentes. La forma anual está adaptada al clima seco interior mientras que la forma perenne está adaptada al clima húmedo de la costa. Estas dos formas tienen diferencias genéticas, pero una diferencia clave es la época de floración; la forma anual florece antes del verano que es muy seco y la forma perenne florece más tarde y esto le permite crecer más. Esto produce un aislamiento reproductivo prezigótico, y además existe el aislamiento postzigótico porqué los híbridos sólo sobreviven en lugares costeros de clima húmedo. El entrecruzamiento entre plantas de los dos ecotipos permitió descubrir, gracias al estudio con marcadores moleculares, que los híbridos no recombinan en una parte de uno de sus cromosomas debido a la presencia de una inversión cromosómica [Lowry and Willis, 2010]**.** Además se pudo determinar que mucha de la variación fenotípica que diferencia a los dos ecotipos segrega junto con la inversión, por ejemplo caracteres relacionados con la morfología. Entre estos caracteres se descubrieron 3 que contribuyen al aislamiento reproductivo entre ambas formas. En resumen, la inversión es polimórfica en muchos de los individuos a lo largo de su distribución geográfica y actúa como un supergen que contribuye a la adaptación local y el aislamiento reproductivo entre las formas anuales y perennes [Lowry and Willis. 2010]. Es por lo tanto este ejemplo, una instantánea sobre el proceso de especiación de estas dos poblaciones de flor mono amarilla y concuerda con las observaciones, modelos e hipótesis previas sobre cómo las inversiones son responsables del cambio evolutivo a través de su efecto de inhibición de la recombinación.

Recientemente se han aportado varios ejemplos más sobre la implicación de las inversiones en la especiación. En el caso del pez espinoso de tres espinas, se detectaron las inversiones a partir de la secuenciación de los genomas de individuos de agua dulce y agua salada [Jones et al. 2012]. Mediante el análisis de los puntos de divergencia entre individuos de agua dulce y salada encontrados en todo el genoma, se determinó que la evolución reusa la variación genética para mantener el aislamiento reproductivo y las inversiones cromosómicas forman parte de esa variación [Jones et al. 2012].

En el caso del mosquito *Anopheles funestus*, también una inversión cromosómica genera aislamiento reproductivo entre los individuos portadores y los no portadores [Ayala et al. 2012]. En este estudio se estimó de manera cuantitativa el aislamiento reproductivo producido por la inversión *3Ra*, tanto el prezigótico como el postzigótico, además de la viabilidad y la adaptación local. Los autores del estudio muestrearon mosquitos a lo largo de un transecto en Camerún que atraviesa diversos hábitats. Después, diseñaron modelos genéticos y los aplicaron a los datos. Resultó que existe una adaptación local fuerte y la viabilidad de los homocigotos varía del 25% al 130%, en comparación con los heterocigotos. En la sabana, la inversión presenta subdominancia, es decir, menor eficacia biológica de los individuos portadores mientras que en las tierras más elevadas es sobredominante. Además está implicada en una fuerte selección del emparejamiento, de forma que en la sabana, las dos clases de individuos homocigotos presentan un 92% de aislamiento reproductivo [Ayala et al. 2012]. Este ejemplo se suma a los anteriores y nos muestra la capacidad de las inversiones cromosómicas de generar barreras genéticas implicadas en la generación de nuevas especies a partir de poblaciones de una misma especie.

1.4 Variación estructural intraespecífica humana

El polimorfismo se puede definir como la existencia en una población de dos o más alelos para una variante estructural, pero ha de darse la condición de que la frecuencia del alelo menos frecuente supere el 1%. Desde el año 2004 se han descubierto muchas variantes estructurales polimórficas en humanos. No obstante, esta gran variación estructural polimórfica en el genoma humano no se ha relacionado aún con la variación fenotípica existente, y sólo una pequeña parte de fenotipos comunes se han asociado con variantes estructurales. De todas maneras, los genes en los que se han descubierto variantes estructurales hacen pensar que muchos de estos polimorfismos tienen un papel importante en la variación de fenotipos y en muchas de las enfermedades comunes [Sharp et al. 2006].

Por el momento estamos lejos de saber cuánto varían fenotípicamente dos personas, pero los avances en la detección de variantes estructurales y los muchos estudios de análisis de *SNPs* en las poblaciones humanas, nos podrían dar una idea de cuánto varían los genomas de dos personas al azar, aunque evidentemente la relación con los caracteres fenotípicos sea mucho más compleja. Tal como se ha comentado anteriormente, la estima más cercana de la proporción de variación estructural entre dos genomas humanos proviene de la secuenciación y ensamblaje de novo del genoma de J. Craig Venter [Levy et al. 2007]. En la comparación con el genoma de Referencia, se detectó un total de 48.8Mb de secuencia diferente correspondiente a variación estructural [Levy et al. 2007] [Pang et al. 2010]. El resultado de estos estudios indica que el genoma de J. Craig Venter difiere del genoma de Referencia en un 1.2% debido a *CNVs* y en un 0.3% debido a inversiones.

Además la variación estructural detectada afecta a 4867 genes [Pang et al. 2010]. Hay que aclarar en primer lugar que el genoma de Referencia está formado por la información genética de varios individuos de distintas poblaciones, por lo que no se ha comparando la estructura de dos genomas estrictamente. En segundo lugar, tanto el genoma de Referencia como el genoma de J. Craig Venter seguramente contienen errores de ensamblaje que pueden dar lugar a errores en la comparación que se mezclan con las variantes estructurales reales. Además, a pesar de que la detección de variantes estructurales por comparación genómica no está expuesta a sesgos hacia un tipo de variantes u otras, ni tampoco relacionados con el tamaño; es difícil pensar que se ha detectado toda la variación entre ambos genomas. Por último, aunque no existiesen los problemas anteriores, el porcentaje de diferencia estructural entre dos genomas no puede ser estimado a partir de las diferencias entre dos individuos, ya que la variación no tiene por que ser igual entre individuos de diferentes poblaciones e incluso entre individuos de una misma población. Es necesario conocer la variación estructural global de varios individuos de diferentes poblaciones humanas para tener una estima aproximada. A pesar de estas limitaciones, tenemos una primera estima de un 1.5% de variación estructural entre dos genomas humanos que afecta una parte mucho más grande del genoma que la variación nucleotídica.

1.4.1 CNVs

Uno de los fenotipos más estudiados son los niveles de expresión génica, que son caracteres celulares heredables y permiten una cuantificación a gran escala a lo largo de todo el genoma. Las primeras variantes que se relacionaron con efectos sobre la expresión génica fueron los *CNVs*, principalmente en estudios centrados en uno o unos pocos genes [McCarroll et al. 2006]. Uno de los modelos más simples para explicar el impacto funcional de los *CNVs* es el cambio en los niveles de expresión de los genes que están en la zona afectada por el *CNV* o bien en los alrededores. Un incremento del número de copias de un gen determinado dará lugar a un aumento de la expresión de ese gen y una disminución en el número de copias, un descenso de la expresión [Hurles et al. 2008]. Entre estos primeros ejemplos sencillos está el caso del citocromo *P450 CYP2D6.* Concretamente el número de copias activas del gen del citocromo *P450 CYP2D6* es proporcional a la metabolización del sustrato CYP2D6, y por lo tanto se correlacionan [Johansson et al. 1993]. A pesar de este ejemplo, la realidad no se ajusta siempre a este modelo sencillo sino que las deleciones, inserciones y duplicaciones dan lugar a efectos variados. Además las variantes también pueden afectar otros elementos funcionales y reguladores del genoma.

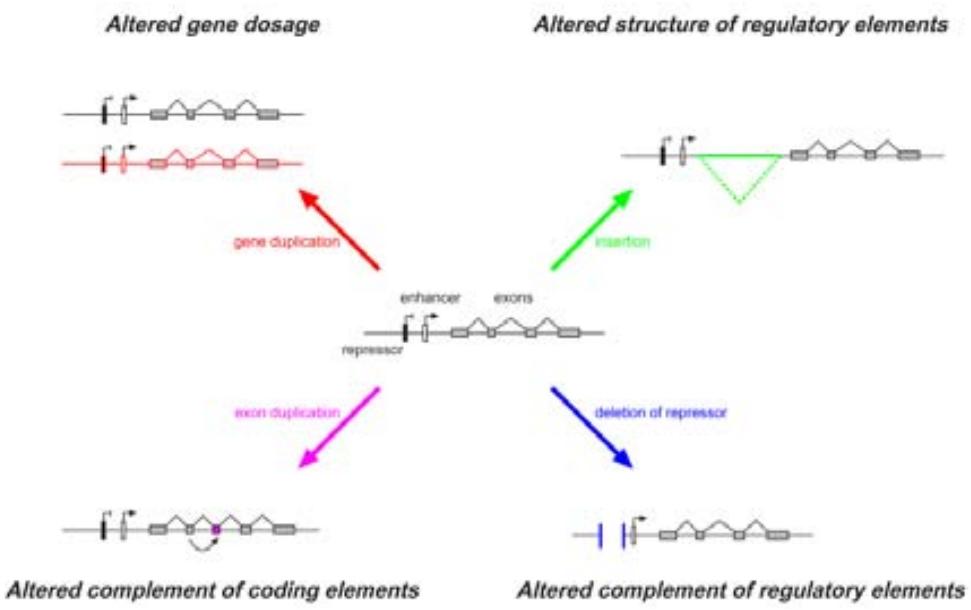

Figura 1.17: Ejemplos de los efectos de los *CNVs* sobre expresión génica. Los genes están representados por cajas de color gris que corresponden a los exones, unidas por líneas que forman una v invertida que representan los intrones, los potenciadores están representados por cajas de color blanco y los represores de color negro. Imagen tomada de Hurles et al. 2008.

Los *CNVs* pueden influenciar la expresión génica de varias maneras (**Figura 1.17**). Inserciones y deleciones o repeticiones en tándem pueden añadir o eliminar copias enteras de genes y cambiar así la dosis génica. Existen genes que codifican para proteínas que realizan funciones insensibles a la dosis génica, y por lo tanto a la cantidad de transcrito o la cantidad de proteína, pero otras funciones sí que se ven afectadas. Para los genes implicados en estas funciones, un cambio en el número de copias implica un efecto sobre el fenotipo, por eso, los constreñimientos selectivos evitan que se formen *CNVs* sobre los genes más sensibles a dosis génica, que realizan funciones importantes [Sharp et al. 2006]. Las inserciones que incluyen sólo parte de un gen pueden resultar en la formación de nuevas proteínas por mezcla de los exones, de nuevas isoformas por efectos en el *splicing* y en el *splicing* alternativo en aquellas que afectan intrones, o incluso nuevos genes producto de la fusión de genes afectados; aunque lo más probable es que todos estos productos no sean funcionales a no ser que mantengan un marco de lectura abierto.

Los *CNVs* también pueden afectar a los niveles de expresión génica estando fuera de las regiones codificadoras a través de su acción sobre los elementos reguladores. Un *CNV* que afecte a la localización o elimine un elemento regulador como puede ser un potenciador o represor afectará a la expresión del gen o genes regulados por ese elemento. Las primeras variantes en las que se detectó un efecto sobre la regulación génica fueron translocaciones de las regiones colindantes a genes del desarrollo cuya desregulación dio como resultado algunos trastornos. Es el caso por ejemplo de *SOX9* en la displasia campomélica, *SHH* en la holoprosencefalia y *PAX6* en la aniridia.

En el caso especial de las deleciones, la eliminación de una parte de la secuencia puede conllevar que mutaciones recesivas tengan efecto al eliminarse uno de los alelos y actúen en hemicigosis en el cromosoma no delecionado [Sharp et al. 2006]. Un ejemplo es la distorsión o ceguera del color rojo o verde. Las variantes estructurales que afectan a la familia génica de las opsinas rojas o verdes en el cromosoma *X* pueden dar lugar a genes mosaico y un efecto fenotípico de distorsión de los colores rojo y verde, que conocemos como daltonismo. Si se eliminan las opsinas de uno u otro color se genera un ceguera para ese color en particular [Deeb et al. 2006].

Otro ejemplo es el de las proteínas alfa-defensinas. Se trata de una familia de proteínas secretadas con función antimicrobiana y que forman parte del mecanismo de defensa del huésped, en este caso, la especie humana. Los genes que las codifican están localizados en el cromosoma *8*. En los años 90 se detectaron duplicaciones de la zona *8p23.1* en algunos individuos, pero no se les diagnosticó ningún síndrome o enfermedad. Se descubrieron de 7 a 8 copias de la secuencia que corresponden con la cromatina adicional que pudo detectarse mediante técnicas citogenéticas en los individuos afectados [Mars et al. 1995]. El análisis de la secuencia permitió encontrar los *CNVs* que afectan a los genes codificantes de las alfa-defensinas pero también de un grupo de genes vecinos codificantes de las beta-defensinas. En este caso la variación estructural afecta a genes relacionados directamente con la respuesta inmune y por lo tanto estos *CNVs* determinan el fenotipo inmunológico que posiblemente corresponde a una mejor respuesta antimicrobiana [Sharp et al. 2006].

Estos son ejemplos muy concretos, pero también es interesante ver de forma global que genes están más afectados por *CNVs*. Éstos se localizan en genes relacionados con la percepción sensorial y la interacción con el ambiente. En general también se relaciona los *CNVs* con enfermedades neurológicas y del desarrollo y caracteres normales como la altura, el índice de masa corporal y la fertilidad en hembras. [Haraksingh and Snyder. 2013]. No obstante los *CNVs* muestran una tendencia a localizarse en zonas de duplicación segmental con pocos o sin genes y los genes que ocupan sitios importantes en las redes biológicas no suelen estar afectados por *CNVs* [Haraksingh and Snyder. 2013].

1.4.1.1 Asociación con enfermedades

Históricamente el descubrimiento de la variación estructural ha estado ligado a enfermedades y síndromes, ya que las primeras variantes se descubrieron a raíz del análisis de los cariotipos de individuos afectados. En muchos trastornos se ha identificado una sola variante como responsable de la enfermedad, bien sea de nueva aparición o heredada, como por ejemplo la trisomía *21* que es relativamente común en la población. En el caso de los *CNVs*, aparte de los estudios dirigidos se han podido identificar síndromes genéticos con el cribado de las zonas duplicadas del genoma [Hurles et al.

2008]. Una parte muy importante de los *CNVs* se forman entre bloques homólogos, por lo que la búsqueda de variantes en regiones que son propensas a su formación en cohortes de pacientes ha sido muy fructífera para identificar trastornos genéticos [Hurles et al. 2008].

Las variantes individuales patogénicas suelen ser raras y demostrar que causan una enfermedad es muy complicado. Para poder relacionar *CNVs* y trastornos genéticos es necesario encontrar individuos con variantes que se solapan y con un fenotipo similar. Además no todos los tipos de *CNV* tienen los mismos efectos sobre genes, y las grandes deleciones *de novo* tienen más posibilidades de causar un trastorno que las duplicaciones pequeñas heredadas. No obstante, hay que ser muy cuidadoso en categorizarlos en variantes que causan enfermedad y en variantes normales ya que en ambos casos el abanico es muy grande y puede superponerse [Feuk et al. 2006]. Un ejemplo de esta dificultad se da en las enfermedades multifactoriales, dónde más de una variante estructural es la causante de la enfermedad o síndrome y cada una explica tan sólo un porcentaje de los individuos afectados.

A medida que aumenta nuestra capacidad de detectar variantes estructurales con mayor resolución, podemos diferenciar mejor las variantes causantes de trastornos de las que no lo son; por ejemplo a través de la comparación con mapas de *CNVs* de cohortes de individuos aparentemente sanos [Hurles et al. 2008]. Para las enfermedades mendelianas el mapeo de genes rotos por variantes estructurales ha revelado en ocasiones qué variante es la predominante. Son ejemplos la enfermedad Charcot-Marie-Tooth tipo I A y la nefronoptisis juvenil. Aún así el cribado de los pacientes sin la variante causal suele detectar otras variantes con una contribución menor al trastorno, normalmente en el mismo gen, como por ejemplo deleciones de exones o duplicaciones que rompen el marco de lectura [Hurles et al. 2008]. Últimamente las causas genéticas de las enfermedades más comunes han sido investigadas a partir de estudios de asociación con *SNPs*, y en cuanto a las variantes estructurales, se han limitado a investigar las candidatas a tener efectos directos sobre los genes implicados, como por ejemplo los *CNVs* que afectan a las alfa-globinas, los receptores Fcγ *3B* [Thabet et al. 2009], las beta-defensinas mencionadas anteriormente [Groth et al. 2008] o el componente 4 del complemento [Wu et al. 2007]. Mediante este análisis se han asociado *CNVs* con la susceptibilidad a contraer enfermedades infecciosas y enfermedades inmunológicas, y las relaciones causales van en aumento a medida que conocemos mejor la variación estructural del genoma humano. El hecho de que los estudios de asociación con *SNPs* no detecten todos los *SNPs* causales debido a las variantes estructurales recurrentes y a la problemática en la genotipación de *SNPs* en regiones de duplicación segmental, hace que se espere encontrar las causas a distintas enfermedades a partir de investigar la relación directa de genes y variantes estructurales [Hurles et al. 2008].

Por otro lado, en el 2010 se realizó el primer estudio de asociación entre *CNVs* y 8 enfermedades humanas comunes [Craddock et al. 2010]. Se usó un *array* para genotipar 3432 *CNVs* polimórficos en 19000 individuos. Se encontraron tres genes afectados por *CNVs* y relacionados con enfermedades, *IRGM* con la enfermedad de Crohn, *HLA* con la enfermedad de Crohn, artritis reumatoide y diabetes tipo 1; y *TSPAN8* con diabetes tipo 2. Estos tres genes ya habían sido identificados en estudios de asociación con *SNPs* previos [Craddock et al. 2010], por lo que se concluyó que los estudios de asociación con *CNVs* comunes no contribuyen al conocimiento de las bases genéticas de enfermedades ya analizadas por asociación con *SNPs*.

En general los *CNVs* han sido relacionados con la susceptibilidad a la infección por VIH, enfermedades autoinmunes y enfermedades genéticas como el síndrome Williams-Beuren [Merla et al. 2010] o el velocardiofacial [Scambler et al. 1992]. Son comunes las deleciones, como por ejemplo la microdeleción de *15q11.2q12* que fue identificada como causa del síndrome de Prader-Willi. Los pacientes de α-talasemias y de distrofia muscular de Duchenne también son portadores de deleciones con puntos de rotura variables [Stankiewicz and Lupski. 2010]. Los pacientes de deficiencia para la sulfatasa responsable de los esteroides son portadores de deleciones flanqueadas por duplicaciones segmentales, al igual que los pacientes del trastorno de Charcot-Marie-Tooth tipo 1A. Ambos trastornos están determinados por el cambio del número de copias del gen *PMP22*. Una duplicación del gen provoca el *CMT1A* y una deleción provoca el *HNPP*. Esta última deleción fue la primera variante submicroscópica responsable de un trastorno genético, que se caracteriza por la neuropatía hereditaria y el riesgo de parálisis por presión *HNPP* [Stankiewicz and Lupski. 2010]. Algunos de estos trastornos y síndromes se recogen en la **Tabla 1.1**. Una deleción de 3.7 Mb que contiene el gen *RAI1* es la responsable del síndrome Smith-Magenis, *SMS*, que también se conoce como síndrome de Potocki-Lupski *PTLS*. La deleción ocurre en el cromosoma *17p11.2*, y está mediada por duplicaciones segmentales. Un ejemplo de síndrome asociado a una microduplicación se da en el cromosoma *22q11.2*. Los pacientes del síndrome de DiGeorge y velocardiofacial son portadores de esta microduplicación a la vez que una deleción. En el cromosoma *17q21.31* se da otro síndrome por microdeleción. Se caracteriza por retraso en el desarrollo, hipotonía, dismorfismo facial, comportamiento anormalmente amistoso, epilepsia, defectos cardíacos y anomalías urinarias. En algunos casos se encuentran variantes estructurales comunes no flanqueadas por duplicaciones segmentales, lo que hace pensar en una predisposición a partir de las características del genoma en la región. Por ejemplo, se dan micro-duplicaciones que afectan al gen *MECP2* que codifica para la proteína 2 de unión metil-CpG ligada al cromosoma *X* [del Gaudio et al. 2006]. Éstas duplicaciones subteloméricas patogénicas del cromosoma *Xq28* son las más comunes no mediadas por duplicaciones segmentales. Se han asociado al síndrome de Rett, un trastorno del desarrollo neurológico que afecta a 1 de cada 10000 chicas.

Tabla 1.1: Trastornos genéticos de transmisión mendeliana. Tabla modificada a partir de Stankiewicz et al. 2002.

Síndrome o enfermedad	Localización cromosómica	Genes	Tipo de variante	Tamaño (Kb)
Síndrome de Bartter tipo III	*1p36*	*CLCNKA/B*	Deleción	11
Enfermedad de Gaucher	*1q21*	*GBA*	Deleción	16
Nefronoptisis juvenil familiar	*2q13*	*NPHP1*	Deleción	290
Distrofia muscular fascioescapulohumeral	*4q35*	*FRG1*	Deleción	25-222
Hiperplasia adrenal congenita III	*6p21.3*	*CYP21*	Deleción	30
β-talasemia	*11p15.5*	*β-globin*	Deleción	4
α-talasemia	*16p13.3*	*α-globin*	Deleción	3.7 o 4.2
Charcot-Marie-Tooth (CMT1A)	*17p12*	*PMP22*	Duplicación	1400
Neuropatía hereditaria (HNPP)	*17p12*	*PMP22*	Deleción	1400
Neurofibromatosis tipo 1	*17q11.2*	*NF1*	Deleción	1500
Enanismo (pituitaria)	*17q23.3*	*GH1*	Deleción	6.7
Carácter farmatogenético CYP2D6	*22q13.1*	*CYP2D6*	Deleción/Duplicación	9.3
Ictiosis	*Xp22.32*	*STS*	Deleción	1900
Ceguera del color rojo o verde	*Xq28*	*RCP y GCP*	Deleción	0

Las variantes estructurales también se han asociado a síndromes y trastornos complejos en una parte de los individuos afectados, como por ejemplo el autismo, esquizofrenia, epilepsia, enfermedad de Parkinson, y de Alzheimer. En el Parkinson, el Alzheimer y la epilepsia, los genes causales fueron identificados en primera instancia a partir del mapeo y análisis de mutaciones puntuales; mientras que para el autismo, la esquizofrenia y el *CMT1A*, se usaron los *CNVs* para identificar el gen de interés [Stankiewicz and Lupski. 2010]. Podemos ver que hay diferentes grados de causalidad por parte de las variantes estructurales, por ejemplo en trastornos monogénicos el grado de causalidad es alto mientras que en síndromes y enfermedades más complejas las variantes explican sólo una parte de los fenotipos o un porcentaje de los individuos afectados. Por último, algunos *CNVs* han sido relacionados con la susceptibilidad a la infección del *VIH*, la glomerulonefritis, la psoriasis, el lupus sistémico eritematoso, enfisemas, etc, y su grado de causalidad puede ser muy bajo [Stankiewicz and Lupski. 2010].

1.4.1.2 Variación entre poblaciones y valor adaptativo

La selección natural limita la mutación en elementos funcionales, y más fuertemente en los implicados directamente en la supervivencia o en la capacidad de tener descendencia. Los análisis a nivel genómico de *CNVs* soportan esta idea. Es decir, como norma general, las poblaciones humanas no presentan diferenciación en cuanto a la frecuencia de *CNVs* [Hurles et al. 2008]. Además las deleciones son las variantes más raras debido a que perder trozos de secuencia importantes está seleccionado negativamente [Hurles et al. 2008]. Por eso, las deleciones tienden a tener un tamaño menor que las duplicaciones. Evidentemente hay excepciones que nos permiten analizar la adaptación, a través de *CNVs* que tienen efectos sobre genes que podrían afectar a la eficacia biológica, y por lo

tanto pueden estar seleccionadas positiva o negativamente. Entendemos por adaptación la capacidad diferencial de generar descendencia viable en un entorno donde hay perturbaciones, por ejemplo la variación de las condiciones ambientales. En ese sentido se ha interpretado la cantidad de *CNVs* que afectan a genes relacionados con la respuesta inmune y la percepción sensorial como una posible señal de selección positiva, es decir, los portadores de más copias de estos genes acabarían teniendo más descendencia que los individuos con menos copias [Nguyen et al. 2008]. Lamentablemente no es fácil analizar la adaptación. La manera más sencilla de analizar el impacto evolutivo de una variante estructural es contando el número de descendientes de los padres portadores de diferentes formas de la variante, pero la posibilidad de realizar estos estudios se da en muy pocas ocasiones.

Alternativamente se usa la frecuencia de la variante en diferentes poblaciones y el análisis de los patrones de variación de las regiones colindantes, ya que cada forma de selección afecta de manera diferente a la frecuencia de la variante estructural y deja una señal o marca en los patrones de variación de los alrededores [Hurles et al. 2008]. Las proteínas APOBEC son un ejemplo de adaptación en *CNVs*. Su principal función es la de defender el organismo de retrovirus mediante la desaminación de las citosinas en uridinas. Estas proteínas están codificadas por una familia pequeña de genes. Algunos individuos tienen los genes *APOBEC3A* y *APOBEC3B*, mientras que otros son portadores de una deleción de unas 30 Kb que produce un gen fusión con la misma secuencia de aminoácidos que *APOBEC3A*, de manera que el efecto es el de eliminar *APOBEC3B* y alterar la regulación de *APOBEC3A* [Kidd et al. 2007]. La frecuencia de la deleción varía significativamente entre las diferentes poblaciones humanas y la diferenciación entre ellas es muy alta, con un índice de estructura poblacional *Fst* de 0.28. La deleción es rara en Africanos y Europeos, con una frecuencia del 0.9% y 6%, respectivamente, mientras que en Asiáticos del Este y en Indios Americanos es del 36.9% y 57.7% respectivamente, y en la poblaciones Oceánicas del 92.9% [Kidd et al. 2007]. La diferenciación entre poblaciones lleva a pensar que la deleción está seleccionada, aunque el fenotipo seleccionado no está claro y se requieren más estudios para llegar a una conclusión.

Otro ejemplo de efecto adaptativo de *CNVs* que afectan a genes es la relación que se observa entre el número de copias de los genes de la amilasa salival *AMY1*, los niveles de proteína y el contenido de almidón de la dieta de diferentes poblaciones humanas. En este caso sí que se puede afirmar que los *CNVs* de la amilasa salival están seleccionados positivamente [Perry et al. 2007]. Es un buen ejemplo porque permite analizar la distribución de *CNVs* entre poblaciones y sus implicaciones adaptativas a través de la historia demográfica que se relaciona, en este caso, con el incremento de la cantidad de almidón que consume la especie humana. En primer lugar se analizó si existía una relación funcional entre el número de copias y la expresión de la proteína amilasa en la saliva. Se estimó el número de copias del gen *AMY1* para 50 Americanos de ascendencia Europea mediante *PCR* cuantitativa y se observó una gran variación en esta población.

Después se analizaron los niveles de proteína amilasa en saliva a partir de experimentos de *western blot* con muestras de saliva de los mismos individuos. A partir de los resultados, se obtuvo una correlación positiva significativa entre el número de copias del gen y la cantidad de proteína en saliva [Perry et al. 2007]. Después se buscaron las diferencias entre las poblaciones con un consumo elevado de almidón y las poblaciones con un consumo bajo. Se estimó el número de copias del gen *AMY1* en 3 poblaciones con dieta rica en almidón y en 4 con dieta pobre en almidón, y se obtuvo una media mayor de copias del gen en las poblaciones con dieta rica en almidón. Entre éstas se encuentran poblaciones Africanas y Asiáticas, por lo que la variación de *CNVs* no sigue un modelo de deriva genética, es decir, no se puede explicar por muestreo aleatorio de alelos. La dieta explica mejor el *CNV* que la proximidad geográfica y los autores sugirieron que la selección natural está implicada. Propusieron un modelo en el que el *CNV* de *AMY1* ha estado sujeto a selección positiva al menos en las poblaciones con dieta rica en almidón, mientras que ha evolucionado de manera neutra en las poblaciones con dieta baja en almidón [Perry et al. 2007].

Otra posibilidad planteada fue que, aparte de la ventaja de los niveles altos de amilasa salival para los individuos que tienen una dieta rica en almidón, el menor número de copias de *AMY1* fuese seleccionado en las poblaciones con dieta pobre en almidón, pero fue descartada porque tener mayor producción de amilasa no tiene un efecto negativo sobre la eficacia biológica. En primer lugar la digestión del almidón comienza en la boca durante la masticación y este proceso gana importancia en episodios de diarrea que sí pueden tener efectos sobre la eficacia biológica. Es en ese momento en que los individuos tienen una enfermedad diarreica cuando esta primera fase de la digestión del almidón es importante para la absorción de energía, ya que una mayor digestión en la boca permitirá absorber más nutrientes dado el funcionamiento anómalo del sistema digestivo [Perry et al. 2007]. De manera normal la amilasa salival llega al estómago donde sigue funcional y aumenta la actividad enzimática de la amilasa pancreática en el intestino fino. Por lo tanto, los niveles más altos de amilasa salival mejoran la eficiencia de la digestión en dietas ricas en almidón tanto en la boca como en el estómago y los intestinos, y ayudan a mantener la absorción de nutrientes en caso de enfermedades intestinales.

En cuanto a las diferencias entre humanos y grandes simios, de media, la especie humana tiene 3 veces más copias que los chimpancés y los bonobos no parece que tengan amilasa salival [Perry et al. 2007]. Esto sugiere que el mayor número de copias fueron duplicaciones en el linaje humano más que deleciones en chimpancé. Siguiendo con la correlación entre mayor número de copias y mayor cantidad de proteína, los niveles de proteína en humanos son de 6 a 8 veces mayores que en chimpancé. Además estos patrones son consistentes con la menor cantidad de almidón presente en la dieta frugívora de chimpancés y bonobos en comparación con la dieta humana. En resumen, este ejemplo muestra que la variación en el número de copias del gen *AMY1* es consistente con las presiones selectivas relacionadas con el contenido de almidón en la dieta durante la

evolución humana.

1.4.2 Inversiones polimórficas

Recientemente se ha visto que las inversiones humanas son más frecuentes de lo que se había especulado a partir de su descubrimiento mediante las técnicas citogenéticas, ya que pueden estar presentes en los cromosomas sin tener un efecto fenotípico visible. Son características muy diferentes de las que se les otorgaron en un principio, como causantes de menor fertilidad y aislamiento reproductivo. Esto ha promovido que se investigue sobre su implicación en la evolución del genoma humano y su relación con caracteres fenotípicos y enfermedades [Alves et al. 2012]. Las inversiones pueden contener genes completos sin que estos vean alterada su expresión, si sus puntos de rotura no los interrumpen, y esto permite que inversiones de un tamaño relativamente grande en comparación con los *CNVs* puedan ser frecuentes en una población; mientras que un *CNV* no puede estar localizado en un gen sin que tenga un efecto de desequilibrio de su dosis génica [Feuk et al. 2010].

En la especie humana existen numerosas inversiones de tamaños diferentes, que segregan en las poblaciones. La mayoría de las inversiones son del orden de kilo bases pero también existen con tamaños mayores de la mega base [Alves et al. 2012]. Tal como se ha comentado pueden ser neutras si no afectan a genes o elementos reguladores y por lo tanto su expansión o desaparición de las poblaciones se da por procesos estocásticos como la deriva genética [Alves et al. 2012]. Pero hay otras inversiones que sí afectan genes o elementos reguladores. Esto ha llevado a los investigadores a plantearse qué mecanismos median en la expansión de las inversiones cromosómicas en las poblaciones [Kirkpatrick et al. 2010]. Evidentemente la combinación de factores que afectan a esta expansión están influenciados por la demografía, la ecología y la historia evolutiva de las poblaciones, por lo que la deriva genética, la selección natural y el flujo genético pueden tener un rol importante en la determinación de la distribución y frecuencia de las inversiones en las poblaciones [Alves et al. 2012].

Además, las inversiones tienen un efecto indirecto muy importante, la supresión de la recombinación en los individuos heterocigotos. Este efecto es clave para su implicación evolutiva. La supresión de la recombinación es resultado de la pérdida de gametos desequilibrados resultantes de la recombinación meiótica entre la zona invertida y no invertida de un chirigotero o de la incapacidad de formar sinapsis de las regiones invertidas en los heteroscios, dependiendo del tamaño de las inversiones [Kirkpatrick et al. 2010]. Aunque sea un efecto indirecto, la supresión de la recombinación puede tener consecuencias drásticas, ya que la inversión actúa como barrera genética y puede mantener un haplotipo en la población [Alves et al. 2012]. La recombinación es uno de los procesos más importantes de la evolución ya que es responsable de la mezcla de genes

y de la introducción de nuevas combinaciones alélicas sobre las que puede actuar la selección natural.

Cuando una inversión se genera en la población, puede perderse, es decir, ningún individuo es portador, fijarse, todos los individuos son portadores, o bien mantenerse en polimorfismo, con individuos portadores y no portadores [Hoffmann et al. 2008]. Según Hoffmann y colaboradores, hay seis explicaciones principales para la propagación y distribución de las inversiones en las poblaciones. En primer lugar, la supresión o reducción de la recombinación facilitaría la propagación de alelos coadaptados, por lo que las inversiones que llevasen alelos favorables se propagarían hasta la fijación a no ser que hubiese migración o selección en contra para evitarla [Hoffmann et al. 2008]. Una hipótesis alternativa es que las inversiones se propagan porque reúnen alelos adaptados localmente incluso sin epistasis, es decir, que actúan aditivamente. La frecuencia de la inversión aumentaría hasta la fijación y solo se mantendría el polimorfismo en el caso de que hubiese migración o que la inversión capturara genes deletéreos [Hoffmann et al. 2008]. La tercera hipótesis ya la hemos comentado, propone que las inversiones están favorecidas por selección directa de los efectos mutacionales de sus puntos de rotura, por lo que la inversión en sí es la diana de selección. Este sería el caso en las enfermedades y trastornos genéticos humanos donde los puntos de rotura de una inversión afectan a la expresión génica a través de la disrupción de un gen o de sus elementos reguladores [Hoffmann et al. 2008]. La siguiente hipótesis también la conocemos, propone que las inversiones son neutras y que su probabilidad de fijación o pérdida depende tan solo del tamaño de población y de la migración. La quinta y sexta hipótesis implican la sobredominancia o subdominancia de las inversiones. La sobredominancia ocurre cuando los heterocigotos para la inversión tienen mayor eficacia biológica que los homozigotos de ambas orientaciones; y la subdominancia ocurre cuando los individuos homozigotos para la inversión tienen una menor eficacia biológica que los heterocigotos u homozigotos estándar [Hoffmann et al. 2008].

1.4.2.1 Efectos de las inversiones

Las inversiones son variantes balanceadas difíciles de detectar y no se han diseñado técnicas de detección a gran escala como son los *arrays* para los *CNVs*, por eso se han detectado y caracterizado menos inversiones polimórficas [Antonacci et al. 2009]. Consecuentemente, las inversiones mejor caracterizadas y que se han genotipado en más individuos siguen siendo las inversiones microscópicas y las inversiones cromosómicas relacionadas con enfermedades y trastornos genéticos, mientras que pocas son las inversiones que se han relacionado con efectos funcionales no relacionados con enfermedades. Dos de ellas están localizadas en los cromosomas *4* y *8*. Ambas tienen puntos de rotura que se encuentran en grupos de genes relacionados con receptores olfativos que tienen gran identidad entre ellos. La inversión del cromosoma *8* tiene un

tamaño de 3.5 Mb y se ha detectado en el 26% de los controles sanos y la inversión del cromosoma *4* tiene un tamaño de 6 Mb y se ha encontrado en el 12.5% de individuos sanos control [Feuk et al. 2010]. En este sentido, los investigadores han hecho un esfuerzo por diseñar técnicas para genotipar estas inversiones, por ejemplo la técnica basada en *FISH* con la que se caracterizaron 6 inversiones de gran tamaño con duplicaciones segmentales en sus puntos de rotura, incluida la del cromosoma *8* [Antonacci et al. 2009].

Tampoco es sencillo establecer una relación de causalidad entre inversiones cromosómicas y enfermedades o trastornos. Hay muchas descripciones de pacientes con fenotipos específicos que también son portadores de una inversión. En estos casos, es difícil encontrar pacientes portadores de la inversión suficientes como para establecer una relación causa-efecto [Feuk et al. 2010] y es uno de los problemas que llevaron a subestimar los efectos de las inversiones. Aún así, los esfuerzos están dirigidos a encontrar la causa de enfermedades. Un ejemplo es el estudio de una inversión cromosómica de 970 *Kb* localizada en el cromosoma *17q21.31*, que predispone a tener una deleción que causa un síndrome con retraso mental [Stefansson et al. 2005]. En la caracterización de la inversión y el análisis del desequilibrio de ligamiento, se vio que, a través de la inhibición de la recombinación, la inversión tiene efectos adaptativos. Este ejemplo nos muestra que el esfuerzo de los investigadores se ha centrado y se sigue centrando mayoritariamente en buscar las causas genéticas de las enfermedades mientras que los efectos fenotípicos de las inversiones no asociados a enfermedad siguen poco explorados.

Aunque se está trabajando en ello. Hay que destacar la reciente publicación de la aplicación de la *PCR* inversa para genotipar inversiones con puntos de rotura en duplicaciones segmentales, que permite genotipar 96 individuos por cada experimento [Aguado et al. 2014]. Estos avances permiten caracterizar a gran escala las inversiones que ya han sido descubiertas y permitirán relacionar mejor las inversiones con sus efectos funcionales, más allá de las enfermedades.

1.4.2.1.1 Efectos de posición y asociación a enfermedades

Los efectos de posición agrupan los efectos mutacionales de los puntos de rotura de las inversiones cromosómicas, tanto si se trata de la disrupción de un gen, la disrupción de un exón o intrón, la disrupción o separación de elementos reguladores del gen o la formación de genes de fusión.

Un ejemplo de efecto de posición es la hemofilia de tipo A, que es una enfermedad ligada al cromosoma *X* y afecta a 1 de cada 5000 hombres [Lakich et al. 1993]. Los hombres que padecen esta enfermedad tienen problemas para coagular la sangre por un déficit del

factor de coagulación VIII. Los pacientes tienen diferentes mutaciones en el gen del factor VIII y mutaciones que se heredan de la madre. Una de las últimas variantes detectadas en esta enfermedad es una inversión presente en el 43% de los pacientes. La caracterización de los puntos de rotura indicó que la inversión de unas 400 Kb se forma por *NAHR* entre duplicaciones segmentales localizadas en el intrón 22 del gen factor VIII y otras copias localizadas 400 Kb aguas abajo del gen. La recombinación homóloga no alélica se daría casi exclusivamente en células germinales masculinas [Lakich et al. 1993], hecho que explica la enfermedad en pacientes sin historia familiar. Al ser uno de los primeros ejemplos de inversiones causantes de enfermedad y dado el tiempo transcurrido, hacen que esta sea una de las inversiones mejor caracterizadas en humanos [Feuk et al. 2010].

Hay otros ejemplos de inversiones relativamente frecuentes en la población humana que causan enfermedades, como la disrupción del gen de la sulfatasa idunorato 2 que provoca la mucopolisacaridosis de tipo 2 en el síndrome de Hunter [Bondeson et al. 1995]. Las inversiones poco frecuentes en la población también causan trastornos o enfermedades por sus efectos mutacionales. La sordera ligada al *X*-2 (*DFNX2*) es un fenotipo asociado a una inversión paracéntrica que se encuentra aguas arriba del gen *POU3F4* [Anger et al. 2013]. La inversión afectaría a los elementos reguladores del gen y la desregulación de éste explicaría parte de los pacientes, ya que también hay una parte que tienen mutaciones en el gen directamente. Por último están los genes de fusión. Como su nombre indica son genes híbridos que se forman a partir de dos genes separados, en este caso juntados por una inversión. Sus nuevas funciones suelen ser oncogénicas si uno de los genes es un proto-oncogén que pasa a ser regulado por un promotor fuerte del otro gen. Por ejemplo, se ha visto que el gen de fusión entre el gen *echinoderm microtubule-associated protein-like 4*, *EML4*, y el gen *anaplastic lymphoma kinase*, *ALK*, se forma por una pequeña inversión en el cromosoma *2p* y explica una parte de los pacientes de cáncer de pulmón de células no pequeñas, *NSCLC* [Soda et al. 2007]. Se detectó el transcrito *EML4-ALK* en el 6.7% de los pacientes de *NSCLC* y aunque estos pacientes se diferencian del resto de *NSCLC* en que tienen mutaciones en otros genes, se engloban dentro de la enfermedad. Otros casos de genes de fusión están relacionados con el cáncer de tiroides, que afecta a una de cada 20000 personas, siendo la enfermedad endocrina más frecuente. Existen distintos tipos de cáncer de tiroides. El de tipo papilar, *PTC*, es el más frecuente con un 80% de los casos. En una parte de los pacientes se han detectado genes de fusión con el gen quinasa del receptor de la tirosina neurotrópica de tipo 1, *NTRK1*, aunque están implicados más genes de fusión en este tipo de cáncer. De todos ellos, el 90% son los genes de fusión *RET/PTC*, formados por una inversión paracéntrica en el cromosoma *10*, aunque otros subtipos están formados por translocaciones [Santoro et al. 2006]. Además, los genes de fusión *NTRK1* también están formados por una inversión paracéntrica en el cromosoma *1q*. En conjunto, varias inversiones forman genes de fusión causantes de cáncer de tiroides de tipo papilar.

1.4.2.1.2 Gametos aberrantes

Las inversiones cromosómicas tanto pericéntricas como paracéntricas pueden producir gametos desequilibrados en los individuos heterocigotos y esto depende mayormente de que se dé recombinación dentro de la región invertida, en concreto un número impar de entrecruzamientos. Las inversiones pericéntricas han sido más estudiadas, con una frecuencia en la población del 1-2% [Morel et al. 2007]. Esta frecuencia es unas 13 veces mayor en hombres infértiles. En un individuo heterocigoto para una inversión, en el que se dé un número impar de entrecruzamientos en el bucle que se forma entre la secuencia invertida y estándar, resultará en una espermatogénesis donde un espermatozoide llevará el cromosoma estándar, otro el cromosoma invertido y dos espermatozoides llevarán cromosomas recombinantes con duplicaciones o deleciones [Morel et al. 2007]. La fecundación de estos gametos recombinantes puede dar lugar a trisomías o monosomías parciales (duplicaciones o deleciones) que pueden causar aborto espontáneo, malformaciones o retraso mental dependiendo de los genes implicados. El mismo proceso en las inversiones paracéntricas dan como resultado cromosomas dicéntricos o acéntricos. Los cromosomas acéntricos se pierden en la meiosis, de manera que en caso de llegar a fecundar un óvulo, el embrión no tendría ese cromosoma o esa parte de cromosoma, por lo que dependiendo de la región no sería viable. En el caso de los cromosomas dicéntricos, sólo sobreviven a la meiosis aquellos que inactivan un centrómero, de manera que pueden llegar a segregar de manera correcta. Aún así, la viabilidad del embrión se ve afectada.

Los estudios de espermatozoides humanos revelan que la incidencia de la recombinación está relacionada con el tamaño de la región invertida, con un tamaño mínimo de 100 Mb, y la inversión al menos del 50% del cromosoma [Anton et al. 2005]. Otro resultado relevante es que hay mucha variabilidad en la producción de gametos recombinantes, con casos en los que no se observan productos recombinantes y casos donde alcanzan el 38% [Anton et al. 2006]. En otro estudio de este tipo se analizaron los gametos recombinantes de 4 individuos portadores de inversiones pericéntricas, uno de ellos portador también de una inversión paracéntrica. Los portadores de inversiones pequeñas de 11 y 29 Mb, no produjeron gametos recombinantes y los portadores de inversiones de mayor tamaño, 65 - 85 Mb, muy pocos. Estos resultados confirmaron que se necesitan inversiones del orden de 100 Mb para que tengan efectos sobre la eficacia reproductiva del portador [Anton et al. 2006]. Recientemente se ha descubierto otro caso de síndrome por gametos desequilibrados. Se trata de la trisomía parcial en el cromosoma *14* a partir de gametos recombinantes [Sgardioli et al. 2013]. Se le ha detectado a una niña que presenta un trastorno de anomalías faciales, hipertelorismo, nariz dismórfica, frente prominente y cara plana, microcefalia, hipotonía, retraso en el desarrollo y malformaciones cardíacas. La afectada tiene una duplicación de aproximadamente 20 Mb entre la región *14q31.3* y la región terminal del brazo y a su madre se le ha detectado una inversión pericéntrica con puntos de rotura en el cromosoma *14p12* y *14q31*.

1.4.2.1.3 Predisposición a otras reorganizaciones

Un efecto indirecto de las inversiones es la predisposición a otras reorganizaciones. Se trata de inversiones que se relacionan con trastornos genéticos, pero que no son la causa sino que incrementan el riesgo de que se den otras variantes estructurales que sí los causan. Para algunos síndromes de microdeleción se ha determinado que uno o ambos padres de los individuos afectados son portadores de una inversión que corresponde con el fragmento delecionado en el hijo. Esta asociación se describió en el síndrome de Williams Beuren, que está causado por una microdeleción de 1.5 Mb en el cromosoma *7q11* [Osborne et al. 2001]. Se realizó un estudio con 12 familias donde el hijo o hija padecía el síndrome a causa de la microdeleción y en el 33% de los padres se detectó la inversión [Osborne et al. 2001]. Hasta aquel momento se conocía la inversión por ser relativamente frecuente en la población, aproximadamente un 5%, pero no se asociaba con ningún efecto fenotípico [Osborne et al. 2001]. Las inversiones cromosómicas podrían tener este efecto al incrementar las probabilidades de alineamiento erróneo entre duplicaciones segmentales no alélicas, donde se encuentran los puntos de rotura de la inversión. Los portadores de la inversión tendrían un riesgo más elevado de sufrir deleciones *de novo* u otras reorganizaciones cromosómicas durante la meiosis [Feuk et al. 2006].

Otros ejemplos se han encontrado, en el síndrome de Angelman, en el que casi la mitad de los padres de individuos afectados son portadores de una inversión de 4 Mb en el cromosoma *15q12*, que tiene una frecuencia del 9% en la población general. El síndrome de Sotos también está causado por una deleción y se ha encontrado una inversión de 1.9 Mb en el cromosoma *5q35* que predispone a tenerla. Algunas de estas inversiones y las reorganizaciones a que predisponen se encuentran en la **Tabla 1.2**. No sólo existen ejemplos de predisposición a trastornos provocados por micro-deleciones, sino que una de las translocaciones constitucionales en el genoma humano también podría estar mediada por una inversión. Se ha descubierto que los genes que codifican para receptores olfativos en los cromosomas *4p16* y *8p23*, están involucrados en la formación de una translocación común en el genoma humano y se ha visto que los padres de los portadores de la translocación son heterocigotos para inversiones en los cromosomas *4p16* y *8p23*. Los descendientes portadores de la translocación presentan fenotipos que pueden ir desde el síndrome de Wolf-Hirschhorn hasta otros conjuntos de características dismórficas menores [Feuk et al. 2006].

Tabla 1.2: Inversiones que predisponen a otras reorganizaciones. Tabla modificada a partir de Feuk et al. 2010.

Localización Cromosómica	Tamaño de la inversión (Mb)	Trastorno o reorganización
3q29	1.9	Síndrome de deleción 3q29
5q35.2-q35.3	1.9	Síndrome de microdeleción de Sotos
7q11.23	1.5	Síndrome de microdeleción de Williams-Beuren
8p23	4.7	Inv dup(8p) y del (8)(p23.1;p23.3)
15q11-q13	4	Síndrome de deleción de Angelman
15q13.3	2	Microdeleción de 15q13.3
15q24	1.2	Microdeleción de 15q24
17q12	1.5	Síndrome de microdeleción Quistes renales y diabetes (RCAD)
17q21.31	0.9	Síndrome de microdeleción de 17q21.31

Por último, otro ejemplo lo encontramos en el síndrome de microdeleción del cromosoma *17q21.31*. Los caracteres fenotípicos asociados al síndrome son el retraso mental, la hipotonía y unas facciones características. Los genes implicados son *MAPT* y *CRHR1* [Koolen et al. 2006]. En un principio se buscaron *CNVs* mediante *array* de *BACs* en 360 individuos con retraso mental y se identificó una deleción de 600 Kb aproximadamente en el cromosoma *17q21.31*. A partir de ahí se analizaron 840 individuos con retraso mental para la deleción mediante *MLPA* con sondas específicas para los genes *MAPT* y *CRHR1*, y se identificaron dos individuos más, portadores de la deleción con el primer punto de rotura idéntico y el segundo separado por 100 Kb. Se confirmó la deleción en los tres individuos por *FISH*. La región delecionada está localizada en la misma región que una inversión polimórfica de 900 Kb común en la población humana, que tiene una frecuencia del 20% en Europeos. Para los tres individuos afectados uno de los padres es portador del haplotipo alternativo. Además en éste se encuentra una duplicación segmental en orientación directa mientras que en el haplotipo principal tiene orientación invertida. Esto indica que la deleción se produce por *NAHR* entre las duplicaciones segmentales en el haplotipo alternativo. Consistentemente, esta duplicación segmental tiene la misma orientación en los tres individuos afectados, y por lo tanto, los portadores de la inversión están predispuestos a que se dé *NAHR* [Koolen et al. 2006] y por extensión al síndrome de microdeleción *17q21.31*. Finalmente se analizaron 1200 individuos Europeos con retraso mental y se detectó la deleción en el 0.3% [Koolen et al. 2006], por lo que se estima que este síndrome tiene una prevalencia de 1 cada 13000-20000 personas, ya que el retraso mental tiene una frecuencia de entre el 2% y el 3% en la población general.

1.4.2.1.4 Inhibición de la recombinación

Los estudios sobre la inhibición de la recombinación en los individuos heterocigotos para una determinada inversión, se realizaron principalmente en especies del género *Drosophila* donde fue descubierta por Dobzhansky y colaboradores [Dobzhansky et al. 1938]. En cambio en humanos se han realizado pocos estudios. Las inversiones mejor caracterizadas a este nivel son la inversión en el cromosoma *17q21.31* de 970 Kb y la del cromosoma *8p23* de 4 Mb. La primera está asociada con la menor expresión del gen *MAPT* que codifica la proteína tau asociada a microtúbulos y además predispone a padecer el síndrome de microdeleción correspondiente [Koolen et al. 2006]. En el caso de la inversión en *8p23*, se asocia con la expresión del gen *BLK* además de afectar la expresión génica de otros 4 genes y confiere riesgo a padecer lupus sistémico eritematoso y artritis reumatoide [Salm et al. 2012].

En ambos casos hay una inhibición de la recombinación a lo largo de la región invertida. La inversión en *8p23* es la inversión polimórfica más grande que se conoce en el genoma humano y muestra una distribución clinal en las poblaciones humanas, con una frecuencia del 80% en África, del 50% en Europa y del 20% en Asia [Salm et al. 2012]. En este caso, se analizaron las diferencias a nivel de patrones de recombinación entre la orientación estándar y la inversión [Salm et al. 2012] y se demostró que la diferenciación genética entre ambas orientaciones está correlacionada con la inhibición de la recombinación. Además estas diferencias se mantienen en todas las poblaciones, por lo que ambas orientaciones han estado evolucionando de manera distinta desde antes de la migración humana fuera de África [Salm et al. 2012]. Por último, como resultado del estudio se encontró una región de 350 Kb que se encuentra en el centro de la inversión y que no es igual de divergente entre ambas orientaciones. Se especula que es debido a que se da recombinación mediante entrecruzamientos dobles [Salm et al. 2012]. Por lo tanto, la inversión no es una barrera total contra la recombinación. Es un buen ejemplo de los efectos sobre la recombinación de las inversiones cromosómicas en el genoma humano, que pueden ser totales para inversiones más pequeñas pero que en el caso de inversiones de tamaño grande pueden no ser totales debido a la posibilidad de formarse entrecruzamientos dobles en los bucles que se forman entre ambas orientaciones en la meiosis, siempre alejados de los puntos de rotura.

El otro caso caracterizado en humanos es la inversión del cromosoma *17q21.31*. Se conocía un bloque de desequilibrio de ligamiento de 1.6 Mb que mantiene dos haplotipos, *H1* y *H2* diferentes para el gen *MAPT* y se demostró la presencia de una inversión polimórfica de 900 Kb responsable del bloque de desequilibrio de ligamiento [Stefansson et al. 2005]. Además se demostró que no ha habido recombinación entre ambos haplotipos. Para ello usaron 6 microsatélites de distribución bimodal y 95 *SNPs* y generaron haplotipos para los 24 cromosomas independientes correspondientes a individuos Europeos de Utah correspondientes al proyecto HapMap. Los haplotipos

generados abarcaron el fragmento invertido de secuencia única de 424 Kb, es decir la parte no duplicada de la inversión. Los haplotipos alternativos *H2* que contienen la inversión se diferenciaron en los 6 microsatélites del haplotipo estándar *H1* y 36 *SNPs* resultaron estar fijados, es decir, uno de los alelos siempre segrega junto a la inversión y el otro junto a la orientación estándar, con lo que demostraron que no se ha dado recombinación dentro de la zona invertida [Stefansson et al. 2005]. En este caso, la inversión en *17q21.31* tiene un efecto de supresión total de la recombinación, posiblemente porque su menor tamaño en comparación con la inversión en *8p23* no permite que se den dobles entrecruzamientos en el bucle meiótico.

1.4.2.2 Polimorfismo y valor adaptativo

En los primeros estudios que se realizaron sobre inversiones cromosómicas en *D. pseudoobscura*, Dobzhansky y colaboradores ya vieron que los alelos invertidos y estándar segregaban en las poblaciones y este polimorfismo se relacionó tempranamente con marcadores genéticos bajo selección natural [Hoffman and Rieseberg. 2008]. La implicación de las inversiones polimórficas en la adaptación local es muy clara cuando la variación de su frecuencia está correlacionada con la geografía, es decir cuando se forman clinas [Kirkpatrick et al. 2010] y esto ocurre claramente en las especies del género *Drosophila*.

Sin embargo en humanos existen pocos ejemplos de inversiones polimórficas que se relacionen con efectos adaptativos y coinciden con las inversiones mejor caracterizadas, de nuevo las inversiones en los cromosomas *17q21.31* y *8p23*. La inversión en *17q21.31* es el mejor ejemplo en humanos de adaptación a través de una inversión hasta el momento. La inversión de 900 Kb inhibe la recombinación y mantiene así dos haplotipos en la población. Los haplotipos alternativos *H2* que contienen la inversión son mucho más homogéneos que los haplotipos estándar *H1* y esto crea un patrón de diversidad inusual, si se suma a la divergencia ancestral entre ambos haplotipos y a la frecuencia del haplotipo *H2* del 20% en la población Europea.

Este patrón de diversidad podría explicarse por selección equilibradora desde la formación de la inversión, sustituida recientemente por selección positiva. Para descartar que se hubiese llegado a la situación actual por evolución neutra se comparó la diversidad de los microsatélites en ambos haplotipos con la de haplotipos generados por simulaciones de coalescencia bajo 4 diferentes situaciones demográficas, incluida una expansión después de un cuello de botella. Se concluyó que por evolución neutra no se puede obtener un patrón de diversidad tan homogéneo en *H2* [Stefansson et al. 2005]. La selección positiva quedaría limitada a las poblaciones Europeas ya que la frecuencia de *H2* es del 20% frente al 6% y al 1% en individuos Africanos y Asiáticos, respectivamente. La diferencia entre los individuos de estas tres poblaciones también se debe al

desequilibrio de ligamiento: en Europeos es más fuerte que en Africanos y Asiáticos, formando un bloque claro, mientras que en individuos no Europeos se divide en bloques más pequeños. Este desequilibrio de ligamiento se puede atribuir al efecto de la inversión y al probable efecto de la selección positiva. El origen Africano del haplotipo *H2* está soportado por la mayor diversidad haplotípica en individuos Africanos, además de contener este grupo los haplotipos fundadores [Stefansson et al. 2005] [Zody et al. 2008]. Por lo tanto las diferencias mutacionales entre los haplotipos *H2* Africanos y Europeos tienen la clave sobre la expansión selectiva de estos últimos [Stefansson et al. 2005].

Para determinar si se ha dado selección positiva en los haplotipos *H2* en Islandia, se genotiparon 29137 Islandeses, 16959 mujeres y 12178 hombres, mediante el marcador *DG17S142* que está fijado respecto a la inversión. Se hizo una regresión del número de descendientes con el número de copias del haplotipo *H2*, con ajuste para el año de nacimiento y sexo, y además se usó una regresión ponderada para tener en cuenta que la gente que tiene más hijos están sobrerrepresentados. Los datos se ajustaron a un modelo dominante de *H2*, es decir que los portadores de *H2* tienen más descendencia que los homozigotos *H1*. Además el efecto es mayor para las mujeres, con un 3.5% más de descendencia, que para los hombres, con un incremento del 2.9% [Stefansson et al. 2005]. Una vez demostrado que el *H2* tiene efectos sobre la eficacia biológica, estudiaron los posibles mecanismos. En ese momento ya se conocía que las mujeres con tasas de recombinación más altas tienden a tener más hijos, así que se estudió el posible impacto de *H2* sobre la recombinación. Se genotiparon 1000 marcadores en todo el genoma de 5012 mujeres además de sus hijos y maridos, en total 20955 individuos. Se realizó una regresión de la tasa estimada de recombinación con el número de copias de *H2* en las madres portadoras, se ajustó por año de nacimiento y edad media de la madre en el momento del nacimiento de su hijo genotipado y el resultado fue que la tasa de recombinación se incrementa en 0.472 Morgans por copia de *H2*.

Este resultado sin precedentes demostró que las tasas de recombinación afectan a la fertilidad de las mujeres y también que la inversión causa un aumento en la fertilidad, aunque sólo explica un 0.3% de la varianza en las tasas de recombinación de las madres. Por lo tanto, existen otros factores que afectan a las bases genéticas de la recombinación que explican el 29.7% de heredabilidad restante de las tasas de recombinación. Esto implica que los haplotipos *H2* afectan a la eficacia biológica mediante otras vías aparte del aumento de la recombinación en las madres. Es por lo tanto este ejemplo el primero en relacionar una variante estructural con una mayor eficacia biológica en humanos y demostrar que está seleccionada positivamente en una población humana. Finalmente los resultados de este estudio soportan la hipótesis que la selección positiva es un factor determinante en la historia evolutiva de las poblaciones Europeas [Stefansson et al. 2005].

En cuanto a la inversión en el cromosoma *8p23*, presenta una distribución poblacional clinal aunque desde un principio se cuestionó que fuera debido a efectos adaptativos. La distribución clinal se correlaciona fuertemente con la distancia de *Addis Ababa*, el sitio de origen de los humanos modernos en Etiopía [Salm et al. 2012]. El alelo invertido es frecuente en el África subsahariana, con un 69.7% de media, mientras que en Europa y Asia la frecuencia disminuye progresivamente, desapareciendo el alelo en América (media 1.3%). Para ver si la distribución de la inversión se podía explicar por factores no demográficos como la selección positiva, se comparó la distribución con la de 19969 *SNPs* autosómicos, seleccionados para representar loci con un frecuencia alélica similar y bajo una putativa evolución neutra. El resultado fue que la distribución de la inversión difiere muy poco de la de los *SNPs* neutros y es consistente con los modelos de expansión fuera de África de la especie humana, donde el papel de la selección positiva o negativa es potencialmente débil [Salm et al. 2012]. Por lo tanto, la inversión parece evolucionar de manera neutra o bajo una muy débil presión selectiva, a pesar de que se han detectado señales de selección natural en la región que abarca [Salm et al. 2012].

1.5 ¿Cuántas inversiones polimórficas hay en el genoma humano?

Una primera estimación del número de inversiones en el genoma humano surge de sumar las inversiones que se han detectado en estudios a nivel de todo el genoma. Se han identificado 1123 inversiones polimórficas en el genoma humano en 9 estudios. En la **Tabla 1.3** se muestra más detalladamente el número de inversiones detectadas en cada estudio, su método de detección y la resolución de sus puntos de rotura, es decir, la precisión con la que han sido definidos.

Tabla 1.3: Inversiones descubiertas en estudios de detección de variación estructural a nivel genómico.

Estudio	Número de inversiones	Método de detección	Resolución
Feuk et al. 2005	3	Comparación genómica	1 pb
Tuzun et al. 2005	56	PEM	> 8 Kb
Korbel et al. 2007	122	PEM	> 3 Kb
Levy et al. 2007	90	Comparación genómica	1 pb
Kidd et al. 2008	224	PEM	> 8 Kb
Wang et al. 2008	17	PEM	> 100 pb
Ahn et al. 2009	415	PEM	> 100 pb
McKernan et al. 2009	91	PEM	> 3.5 Kb
Pang et al. 2010	105	PEM	> 2 Kb

Aunque estos 9 estudios son los que más inversiones polimórficas han detectado y contienen las inversiones detectadas en estudios globales y algunas inversiones detectadas en los estudios que tienen como objetivo detectar la variante causal de una enfermedad,

no todas las inversiones detectadas en estudios dirigidos sobre una enfermedad están en esta lista. De todas maneras usamos la **Tabla 1.3** como orientativa, ya que solamente algunas de las inversiones han sido validadas y por lo tanto seguramente una parte son falsos positivos y por otra parte de manera intuitiva podemos pensar que no todas las inversiones polimórficas en el genoma humano han sido detectadas.

1.5.1 Bases de datos y redundancia

El número de entradas en bases de datos que almacenan información sobre variantes estructurales ha crecido exponencialmente desde la publicación del genoma humano especialmente para los *CNVs*, que se han caracterizado en mayor medida, pero también para las inversiones, sobre todo a partir del desarrollo de la técnica de *PEM*. Un tema importante para las bases de datos es el solapamiento de variantes, es decir, la redundancia de entradas. A día de hoy, la base de datos de variantes genómicas, *DGV* [MacDonald et al. 2014] contiene 2.307.729 entradas, de 55 estudios diferentes, que se convierten en 110.101 variantes estructurales no redundantes, de las cuales 238 son inversiones. La redundancia no es un tema fácil de solucionar. No se evita simplemente superponiendo todas las entradas. Los diferentes estudios definen los puntos de rotura de las variantes con distintos grados de precisión dependiendo del método de detección utilizado y esto hace difícil unificar las variantes detectadas, por ejemplo con *microarrays* o *PEM*.

En el caso de las inversiones polimórficas en humanos, la gran mayoría han sido detectadas por *PEM*, por lo que la definición de los puntos de rotura no sólo es importante para poder analizar sus efectos sino también para diferenciarlas. Otro problema es la fiabilidad, ya que debido a la repetitividad del genoma humano, los mapeos de extremos apareados generan una parte importante de falsos positivos [Lledó and Cáceres. 2013]. *InvFEST* es la base de datos específica para inversiones polimórficas en el genoma humano y su política tiene en cuenta la precisión de cada estudio a la hora de incorporar las inversiones detectadas al conjunto [Martínez-Fundichely et al. 2014]. Los autores han desarrollado además un programa informático denominado GRIAL [Martínez-Fundichely *et al.*, in preparation], que usa la información del mapeo de los extremos apareados para generar nuevas predicciones de inversiones, refinar los puntos de rotura de las predicciones existentes y generar una puntuación sobre su fiabilidad. Los mapeos de los extremos apareados se reagrupan siguiendo unas reglas geométricas, que se basan en las características de las inversiones y esto permite que se seleccionen exhaustivamente los mejores grupos de mapeos de extremos para definir de la manera más precisa posible las inversiones. Este refinamiento de los puntos de rotura permite que después se puedan diferenciar mejor las predicciones y agruparlas en un conjunto no redundante. La puntuación sobre la fiabilidad de la predicción permite además descartar una parte importante de falsos positivos. En resumen, es la base de datos que mejor representa el

número de inversiones que hay en el genoma humano, con menos inversiones redundantes y menos de falsos positivos. En total contiene 1092 predicciones de inversiones, pero si se tiene en cuenta el índice de fiabilidad y las validaciones experimentales, la base de datos cuenta con 617 inversiones polimórficas validadas experimentalmente o fiables [Martínez-Fundichely et al. 2014].

1.5.2 Espectro de variación detectada

Evidentemente 617 es una cifra aproximada de las inversiones reales que existen en el genoma humano y una parte de las que se han detectado. En los últimos años se ha descubierto una parte de la variación que no se consideraba en un principio, las inversiones submicroscópicas muy pequeñas, de menos de 1 Kb. Por ejemplo, en la comparación entre los ensamblajes del genoma de Referencia en humanos y en chimpancés se detectó una gran cantidad de inversiones de menos de 250 pb [Feuk et al. 2005] y aunque según la metodología de comparación pueden estar enriquecidas en falsos positivos, las inversiones pequeñas representan una parte del espectro de variación no explotada. Entre los estudios de detección comentados anteriormente solamente los estudios de comparación de ensamblajes de genomas [Feuk et al. 2005] [Levy et al. 2007] y los estudios de *PEM* que usan insertos pequeños [Wang et al. 2008] [Ahn et al. 2009] detectan esta parte de la variación; mientras que el resto de estudios detectan preferentemente inversiones con tamaño mínimo de 2 Kb y con menos frecuencia este tipo de inversiones pequeñas (**Tabla 1.3**). Si echamos un vistazo a *InvFEST*, 157 predicciones tienen un tamaño menor de 1 Kb, aunque sólo 135 son fiables. Seguramente esta cifra crecerá a medida que los estudios de detección de inversiones basados en *PEM* incluyan tamaños de inserto más pequeños y mayor recubrimiento que permitan detectar la variación estructural de tamaño menor de 1 Kb; aunque esto implique mejorar la técnica en cuanto a la detección de falsos positivos.

Por lo tanto podemos decir que las inversiones de menor tamaño dentro del espectro de variación, no han sido analizadas en el mismo grado que el resto de inversiones. Pero también tenemos evidencias de falsos positivos entre las inversiones detectadas. Hasta el momento todos los estudios de detección de inversiones se basan en algún grado en el genoma de Referencia y éste contiene zonas mal ensambladas o cuya orientación no es la correcta. Dos de estos errores, correspondientes con las regiones *Xp11.3* y *6q27*, fueron detectados por Pang y colaboradores en el año 2013 [Pang et al. 2013]. Más allá de los problemas que impiden la detección de toda la variación estructural de un genoma concreto, estamos comparando constantemente los genomas con el genoma de Referencia.

Se han comparado con el genoma de Referencia, un individuo de origen incierto [Tuzun et al. 2005], ese mismo individuo y un individuo de origen Africano [Korbel et al. 2007], un individuo de origen Europeo [Levy et al. 2007], uno Asiático [Wang et al. 2008], cuatro Africanos, dos Asiáticos, dos Europeos y el individuo de origen incierto [Kidd et al. 2008], y otro individuo Asiático [Ahn et al. 2009]. Algunos de estos individuos han sido comparados en varios estudios. En resumen, a *grosso modo* se han comparado unos 10 individuos provenientes de diferentes poblaciones humanas, con un genoma que aunque fue secuenciado a partir de un grupo de individuos, proviene en más del 70% de un único individuo [Tuzun et al. 2005]. Estos datos nos indican que hay una gran cantidad de variación por descubrir y la detección de más variación estructural pasa por la comparación de más genomas entre sí.

El ensamblaje de novo y la comparación de genomas son la solución para conocer la variación estructural completa de un genoma y por extensión la del genoma humano. La comparación de genomas no tiene los sesgos y problemas que hemos comentado del mapeo de los extremos apareados y permite detectar inversiones de cualquier tamaño. Es por lo tanto el método que detecta toda la variación estructural de un genoma con la máxima precisión. Por el momento las limitaciones están relacionadas con el coste de la secuenciación y sobre todo del ensamblaje.

Es cierto que las técnicas de secuenciación de última generación permiten la secuenciación de genomas con menor coste y de una manera más rápida pero el tamaño de las secuencias dista mucho del necesario para generar un ensamblaje nuevo sin utilizar el genoma de Referencia. Es por eso que hasta el momento se ha usado la secuenciación clásica *Sanger*, en la que se obtienen fragmentos más largos y con una calidad mayor que las técnicas de última generación. Como contrapartida es un proceso mucho más lento y costoso. Idealmente se necesitarían secuencias de más de 100 Kb con una tasa de error menor al 0.1% para obtener ensamblajes de una calidad comparable al actual genoma de Referencia [Alkan et al. 2011], cifras lejanas a las que se manejan en *NGS*.

1.5.3 Validación experimental y genotipación

Para determinar los efectos de las inversiones sobre la expresión de genes, caracteres fenotípicos y su implicación evolutiva, es necesario en primer lugar validar experimentalmente las inversiones, es decir, asegurarse de que no se trata de falsos positivos, de errores propios de la técnica de detección o bien de los genomas a comparar; y en segundo lugar genotipar el mayor número posible de individuos de poblaciones diferentes para conocer su distribución poblacional. Hasta el momento disponemos de esa información para pocas inversiones, aunque el número de inversiones validadas experimentalmente no es tan bajo. Esto se da porque aparte de los estudios de detección de inversiones a nivel global, hay estudios dirigidos a encontrar la variante causal de una

enfermedad que aportan inversiones a esta lista. En total, según *InvFEST*, se han validado experimentalmente mediante *FISH* y *PCR* unas 54 inversiones entre los estudios de detección global y los estudios dirigidos publicados. En la **Tabla 1.4** se muestran las inversiones validadas experimentalmente en cada estudio y el número de individuos genotipados y su origen si se conoce. Hay que destacar que algunas inversiones se validan en más de un estudio.

Tabla 1.4: Inversiones validadas experimentalmente y genotipadas en individuos de distintas poblaciones humanas.

Estudio	Inversiones validadas experimentalmente	Número de individuos y origen
Osborne et al. 2001	*7q11.23*	-
Giglio et al. 2002	*4p16, 8p23*	-
Gimelli et al. 2003	*15q11*	6 mujeres ?
Feuk et al. 2005	*7p22, 7q11, 16q24*	10 Europeos
Stefansson et al. 2005	*17q21.31*	29137 Europeos
Gilling et al. 2006	*10p11*	60 Europeos
Korbel et al. 2007	45 inversiones	1 Africano, 1 ?
Deng et al. 2008	*8p23*	2 Africanos, 1 Asiatico, 7 Europeos
Antonacci et al. 2009	*3q29, 8p23, 15q13.3, 15q24, 17q12, 17q21.31*	11 Africanos, 8 Asiáticos, 8 Europeos
Bosch et al. 2009	*8p23.1*	24 Europeos
Entesarian et al. 2009	*10q11.22*	8000 Europeos aprox.
Pang et al. 2010	*16p12.2*	8 Individuos ?
Salm et al. 2012	*8p23*	15 Africanos, 17 Asiáticos, 68 Europeos
Pang et al. 2013	*3q26.1*	-
	Xp11.3, 7q11.22, 4q22.1, 1q31.3, 6q27, 16q24.1	10 Africanos, 20 Asiáticos, 11 Europeos, 1 ?
	16q23.1	871 individuos de 57 poblaciones
Aguado et al. 2014	17 inversiones	10 Africanos, 2 Asiáticos, 70 Europeos, 1 ?

Hay que destacar dos métodos de validación experimental entre los más usados, la hibridación in situ fluorescente, *FISH*, y la *PCR*. Estos métodos se han usado tanto para la validación como para la genotipación de individuos de diferentes poblaciones, en estudios de detección global y en los estudios dirigidos. Por ejemplo Antonacci y colaboradores en el año 2009 usaron *FISH* sobre metafases para genotipar 6 inversiones asociadas a enfermedades en 27 individuos de tres poblaciones HapMap [Antonacci et al. 2009]. Korbel y colaboradores validaron experimentalmente 45 inversiones en el año 2007, 4 por *FISH* y 41 por *PCR* en dos individuos, uno de origen Africano y otro de origen no conocido. Los estudios de Deng y colaboradores en el año 2008 y Salm y colaboradores en el año 2012, genotiparon 10 y 100 individuos respectivamente, de origen Africano, Asiático y Europeo, para la inversión *8p23* mediante *FISH* y luego usaron *SNPs* que segregan siempre con la inversión para genotipar 209 y 1894 individuos de varias poblaciones respectivamente. En cuanto a los estudios que más inversiones han validado y genotipado experimentalmente son los estudios de Pang y colaboradores en el año 2013 y de Aguado y colaboradores en el año 2014. En el primero validaron mediante *PCR* 8 inversiones y 6 de ellas se genotiparon por *PCR* en 42 individuos de cuatro poblaciones HapMap y una inversión en 871 individuos de 57 poblaciones del panel de diversidad del genoma humano *HGDP* [Pang et al. 2013]. En el segundo estudio que es el más reciente,

se han validado y genotipado 17 inversiones, con puntos de rotura en duplicaciones segmentales, en 83 individuos de 3 poblaciones HapMap mediante *PCR* inversa [Aguado et al. 2014]. En este estudio se presenta una nueva aplicación de la *PCR* inversa que permite genotipar a gran escala 90 individuos por reacción [Aguado et al. 2014]. Es por lo tanto el primer método de genotipación experimental a gran escala de inversiones con puntos de rotura en duplicaciones segmentales. En resumen, las inversiones para las que se han genotipado individuos de varias poblaciones humanas y de las que se tiene conocimiento de su distribución poblacional son las mejor caracterizadas para comenzar a analizar sus efectos sobre la evolución del genoma humano, como ya se ha hecho en las inversiones del cromosoma *17q21.31* y *8p23*.

1.5.4 Inversiones detectadas entre los genomas de J. Craig Venter y Referencia

La publicación del genoma humano revolucionó en su momento la biología y permitió que ahora podamos detectar y genotipar inversiones en individuos de diversas poblaciones humanas, pero también fue muy importante la publicación de un segundo genoma, correspondiente a J. Craig Venter (HuRef), ensamblado de manera independiente al genoma de Referencia [Levy et al. 2007]. La comparación de ambos ensamblajes permitió descubrir la variación estructural que lo diferencia del genoma de Referencia con precisión máxima y sin límites de tamaño o de localización. En cuanto a las inversiones cromosómicas, se detectaron 90, de tamaños que van desde menos de 100 pb hasta unas 15 Mb. Evidentemente, pueden existir falsos positivos, por ejemplo errores en el ensamblaje del nuevo genoma, del genoma de Referencia o bien errores en la comparación de ambos. Por esos motivos y para poder analizar sus efectos, es importante la validación experimental de este conjunto de inversiones y la genotipación de individuos de diferentes poblaciones humanas. Tres años más tarde de la publicación del genoma de J. Craig Venter y del resultado de la comparación con el genoma de Referencia, se publicó un estudio en el que se detectó más variación estructural en *HuRef* a partir del análisis del mapeo de los extremos apareados generados durante la secuenciación [Pang et al. 2010]. En este estudio se publicaron 79 nuevas inversiones en *HuRef* respecto al genoma de Referencia, aunque el método por el cual fueron detectadas, *PEM*, hace que estas inversiones no representen de manera insesgada la variación entre ambos genomas.

1.5.4.1 Secuenciación y ensamblaje de HuRef

El genoma de J. Craig Venter fue secuenciado mediante la técnica clásica de *Sanger* siguiendo una estrategia *shotgun*, en la que se fragmentó el ADN y se secuenciaron los fragmentos. En total *HuRef* contiene 2810 Mb de secuencia contigua con una cobertura de 7.5 veces para cualquier región [Levy et al. 2007]. Se generaron *contigs* que se pudieron

conectar mediante el uso de extremos apareados formando 4528 *scaffolds* o conjuntos de *contigs*. En este estudio se modificó el ensamblador original de Celera Genomics [Istrail et al. 2004] para permitir agrupar los *reads* de alelos diferentes y ensamblar el genoma diploide [Levy et al. 2007]. Los *reads* que engloban una variante se separaron por alelos y un alelo se definió por dos o más *reads* con secuencia idéntica. A continuación, se les asignó una fiabilidad a cada alelo en función de la suma de la calidad media de la secuencia del *read* localizada en la variante de todos los *reads* con secuencia idéntica, se usó el alelo con mayor fiabilidad como secuencia consenso y se reportaron por separado los alelos alternativos. Por último se usaron los extremos de *BACs* generados a partir del genoma humano de Referencia para alinear los 553 *scaffolds* de *HuRef* de al menos 100 Kb. Estos *BACs* ya habían sido usados en el ensamblaje del genoma de Celera Genomics en el proyecto de secuenciación del genoma humano. Se mantuvieron los extremos de *BACs* que mapearon de manera única y cerca del extremo de un *scaffold*, de manera que el otro extremo estuviese necesariamente fuera. Se usaron las parejas de extremos de *BACs* en que ambos cumplieran este criterio y de esta manera se dieron 144 uniones de *scaffolds* con dos parejas de extremos de *BACs* como soporte mínimo y 98 uniones con una pareja de extremos como soporte. Así se redujo el número de *scaffolds*. Además para ordenarlos en cromosomas se usó el mapeo uno contra uno con el genoma de Referencia (versión NCBI36).

1.5.4.2 Comparación genómica

La comparación uno contra uno de los genomas *HuRef* y Referencia se realizó usando el software *A2Amapper* [Istrail et al. 2004]. Este software consigue un buen mapeo de regiones con secuencia repetitiva, por ejemplo duplicaciones segmentales casi idénticas, siempre que no sean muy grandes. Además en los genomas secuenciados a partir de la estrategia *shotgun* hay una menor representación de duplicaciones segmentales, que suelen ser eliminadas para evitar problemas en el ensamblaje. Por otra parte, las secuencias duplicadas solamente en uno de los dos genomas no se incluyeron en el mapeo uno contra uno [Levy et al. 2007]. Cada mapeo uno contra uno tiene tres niveles: emparejamientos o *matches*, recorridos o *runs* y bloques o *clumps*. Un emparejamiento es un alineamiento local de alta identidad, que suele acabar en un indel o un *gap* en uno de los ensamblajes. El número total de pares de bases en emparejamientos es una medida de cuánta secuencia se comparte entre ambos genomas. En cambio, los recorridos son conjuntos continuos de emparejamientos con la misma orientación que pueden tener un número de bases diferente para cada ensamblaje, porque pueden contener *indels* o *gaps* entre emparejamientos. Contienen regiones sin reorganizaciones entre ensamblajes y su número es una medida de las diferencias de orden y orientación. Los bloques son grupos de recorridos que pueden contener pequeñas discontinuidades como inversiones o *CNVs* y son agrupaciones más cercanas a los *scaffolds*. Por ejemplo, en el ensamblaje de *HuRef* se detectaron *scaffolds* que contienen más de un bloque de al menos 5 Kb en comparación

con el genoma de Referencia y se etiquetaron como potencialmente quiméricos.

Las variantes se caracterizaron mediante el alineamiento de los *reads* de secuenciación en el ensamblaje de *HuRef* y por la comparación de las regiones diferentes en el mapeo uno contra uno. Las variantes heterocigotas fueron detectadas por el ensamblador en la agrupación de los *reads* por alelos y en el mapeo uno contra uno mientras que las variantes homocigotas se detectaron solo por diferencias en el mapeo uno contra uno. Por ejemplo, en inserciones y deleciones, *HuRef* tiene más o menos secuencia en comparación con el genoma de Referencia, y en el caso de las inversiones, tiene diferente orientación. Dependiendo del tamaño de las variantes, se incluyeron en los bloques y por tanto en los *scaffolds* mapeados. Es el caso de las variantes pequeñas. Por otro lado, las variantes grandes se encuentran en los *scaffolds* no mapeados. En total inicialmente se identificaron 5.061.599 variantes en el mapeo uno contra uno que representan variantes heterocigotas, incluyendo *SNPs*, indels y variaciones de más de un nucleótido. Se pasaron por varios filtros de calidad y resultaron en 3.325.530 variantes heterocigotas fiables. Además este mapeo produjo aproximadamente 150 Mb de secuencia *HuRef* no mapeada, formada por *scaffolds* parcialmente mapeados y no mapeados. Se recuperaron 233.796 variantes heterocigotas de esta secuencia no alineada, que se sumaron a las detectadas en el mapeo uno contra uno. También se detectaron variantes homocigotas por comparación de la secuencia de *HuRef* no mapeada con el genoma de Referencia. Después de pasar por los mismos filtros de calidad resultaron en 275.512 inserciones, 283.961 deleciones y 90 inversiones.

1.5.4.3 Análisis y validación de inversiones en HuRef

Es importante conocer qué inversiones de este conjunto han sido validadas experimentalmente y para cuantas de ellas se tienen datos de su frecuencia y distribución poblacional. Las inversiones detectadas no son únicas de *HuRef* ni del genoma de Referencia de manera que pueden haber sido validadas en otros estudios y en otros individuos. Esto es lo que ocurre con un estudio que se realizó en el mismo año que se publicó el genoma *HuRef* y cuyo objetivo fue la detección de variación estructural mediante *PEM* [Korbel et al. 2007]. Se detectaron algunas inversiones que están presentes en *HuRef* y se validaron 12 de ellas por *PCR* (*HsInv0024*, *HsInv0030*, *HsInv0031*, *HsInv0042*, *HsInv0045*, *HsInv0050*, *HsInv0061*, *HsInv0062*, *HsInv0064*, *HsInv0066*, *HsInv0073* y *HsInv0074*) y una por *FISH* (*HsInv0002*). Hemos querido referenciar estas inversiones con su código identificativo correspondiente a la base de datos *InvFEST* en lugar de sus coordenadas, y además coincide con su orden en la lista de inversiones reportadas en la comparación de ambos genomas [Levy et al. 2007]. En este estudio se genotiparon dos individuos que corresponden al proyecto HapMap, un individuo de origen Africano de código NA18505 y un individuo de origen no conocido NA15510, por lo que no se puede establecer ninguna distribución poblacional.

Por otra parte, en el año 2013 se publicó un estudio directamente relacionado con las variantes descubiertas en *HuRef*, con el objetivo de establecer los mecanismos de formación en esas variantes y dar una idea de su impacto y características en el genoma humano [Pang et al. 2013]. Se analizaron experimentalmente 8 inversiones, de las cuales 7 habían sido detectadas por comparación genómica [Levy et al. 2007] y la inversión restante por *PEM* usando las secuencias generadas en el ensamblaje de *HuRef* [Pang et al. 2010]. De éstas, 6 inversiones fueron validadas experimentalmente mediante *PCR (HsInv0004, HsInv0030, HsInv0031, HsInv0052, HsInv0063 y HsInv0095)* y dos (*HsInv0073* y *HsInv0062*) resultaron ser posibles errores en el genoma de Referencia, aunque no fue demostrado. En cuanto a la distribución poblacional, las inversiones se genotiparon por *PCR* en 10 individuos de origen Africano, 20 de origen Asiático, 11 de origen Europeo y 1 de origen no conocido, pertenecientes a poblaciones del proyecto HapMap y 1 chimpancé con el que se estableció la orientación ancestral. En 3 de ellas la inversión se genotipó también por *SNPs* marcador en 60 individuos Europeos, 88 individuos Asiáticos y 59 individuos Africanos. Además, a partir de las frecuencias imputadas de los *SNPs* marcador, se calculó la diferenciación genética entre poblaciones mediante el índice de estructura poblacional *Fst* (**Tabla 1.5**).

Tabla 1.5: Resultado de la genotipación experimental y mediante *SNPs* marcador de inversiones en *HuRef*. Tabla modificada a partir de Pang et al. 2013

Identificador *InvFEST*	Localización Cromosómica	Alelo	Europeos	Chinos	Japoneses	Yoruba	Estado Ancestral	*SNP* marcador	Alelo del *SNP*	*Fst*
HsInv0052	*3q26.1*	Multi-alélica	-	-	-	-	-	-	-	-
HsInv0073	*Xp11.3*	Inversión	19	14	14	16	Inversión	-	-	-
		Referencia	0	0	0	0				
HsInv0063	*7q11.22*	Inversión/Deleción	14	13	15	3	Referencia	-	-	-
		Referencia	10	7	5	17				
HsInv0063	*7q11.22*	Inversión/Deleción	83	65	59	28	Referencia	*rs1525303*	A	0.38
		Referencia	37	25	27	90			T	
HsInv0030	*16q23.1*	Inversión	15	19	20	20	Inversión	-	-	-
		Referencia	5	0	0	0				
		Deleción	4	1	0	0				
HsInv0095	*4q22.1*	Inversión	14	16	17	18	Inversión	-	-	-
		Referencia	10	4	3	2				
HsInv0095	*4q22.1*	Inversión	81	66	68	98	Inversión	*rs1477602*	A	0.08
		Referencia	39	24	22	16			G	
HsInv0004	*1q31.3*	Inversión	8	2	2	0	Inversión	-	-	-
		Referencia	16	18	18	20				
HsInv0004	*1q31.3*	Inversión	24	13	10	1	Inversión	*rs1627999*	G	0.17
		Referencia	96	77	80	115			A	
HsInv0062	*6q27*	Inversión	24	20	20	20	Inversión	-	-	-
		Referencia	0	0	0	0				
HsInv0031	*16q24.1*	Inversión	12	10	11	15	Referencia	-	-	-
		Referencia	12	10	9	5				
HsInv0031	*16q24.1*	Inversión	81	51	51	77	Referencia	*rs9933231*	T	0.03
		Referencia	37	39	39	37			A	

Dos inversiones (*HsInv0052* y *HsInv0063*) resultaron pertenecer a variantes complejas. La primera está relacionada con un *CNV* que fue validado por *PCR* cuantitativa, llegando a la conclusión de que la inversión y el *CNV* aparecieron a la vez (**Figura 1.18**), por lo que se consideró una variante multialélica y no se analizó su distribución poblacional.

Además no sólo hay una deleción asociada a la inversión sino que toda la región está afectada por una deleción mucho más grande que es independiente a la variante compleja.

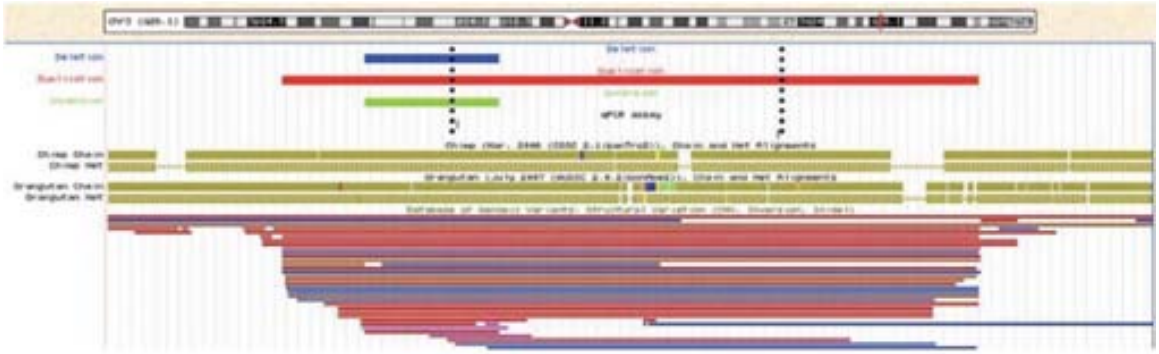

Figura 1.18: Imagen del navegador genómico en la región de *HsInv0052*. En la parte superior, la inversión está representada por una barra de color verde, la deleción por una barra de color azul y la duplicación por una roja. En la parte inferior se pueden ver las entradas de variantes estructurales en *DGV*. Entremedio y de color verde pistacho podemos ver información relacionada con los ensamblajes de chimpancé y orangután. Las líneas negras discontinuas verticales representan ensayos de *PCR* cuantitativa realizados en el estudio. Imagen tomada de Pang et al. 2013.

En el caso de la inversión *HsInv0063*, se detectó una variante compleja que incluye inversión y deleción. Esta inversión tiene una frecuencia mayor en Europeos y Asiáticos en comparación con los individuos Africanos, y un valor *Fst* de 0.38 que indica una varianza mayor de lo normal entre poblaciones. En ausencia de elementos funcionales, los autores la atribuyen a un efecto fundador en la población Euroasiática sumado a la deriva genética.

Por otro lado la inversión *HsInv0030* intercambia el exón 1 entre dos genes *CTRB1* y *CTRB2*, que codifican para proteínas precursoras de quimotripsinógenos B (**Figura 1.19A**). Aunque los genes tienen una similitud del 97%, los primeros exones son un poco más diferentes, con un 82% de similitud. El estado ancestral corresponde a la orientación invertida que tiene una frecuencia alta en la población. Además, se detectó una deleción adyacente a la inversión en algunos individuos Europeos y un individuo Asiático. Esta deleción afecta al exón 6 del gen *CTRB2*, y mediante el análisis de los genotipos de primates se postuló como un alelo derivado del haplotipo invertido. Para la inversión y la deleción se genotiparon por *PCR* 871 individuos del panel de diversidad del genoma humano *HGDP-CEPH* provenientes de 57 poblaciones. La orientación invertida resultó ser la mayoritaria y la deleción está ligada a ella. Las poblaciones con mayor frecuencia para el haplotipo conjunto de la inversión y la deleción son Surui (47.2%), Vasco Franceses (20.5%), Italianos del Norte (17.9%) y Drusos (15.4%), mientras que no se encontró en Yorubas, Yakut, Sindhi y Pigmeos Mbuti (**Figura 1.19B**). El índice de estructuración de la población tiene un valor de 0.53, por lo que resultaron evidentes las diferencias de frecuencia entre poblaciones. Mediante el análisis de los transcritos se

llegó a la conclusión de que puede estar relacionada con adaptación a diferencias en la dieta a partir de la enzima digestiva quimotripsina.

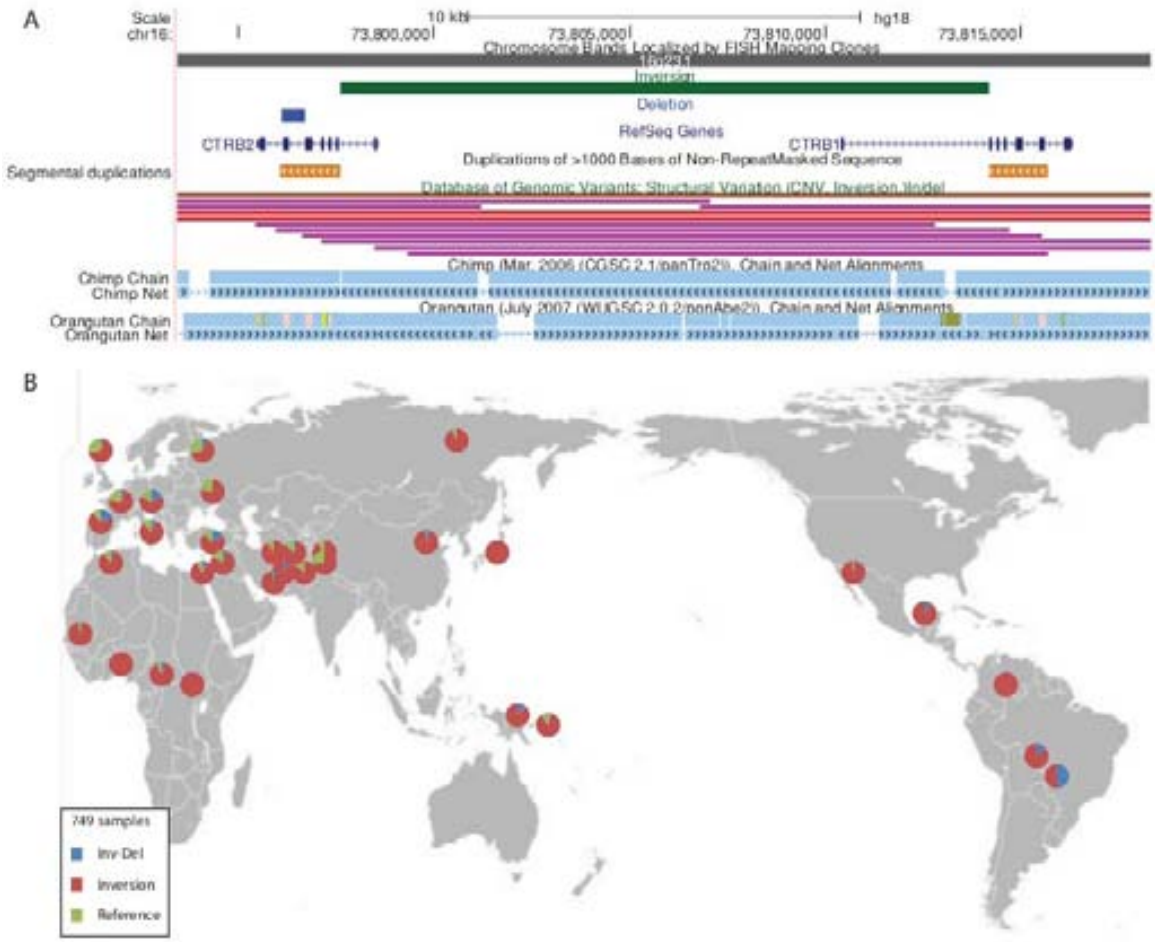

Figura 1.19: Esquema y distribución poblacional de la inversión *16q23.1* y su deleción asociada. (A) Imagen del navegador genómico donde se muestra la inversión en verde, la deleción en azul, las duplicaciones segmentales en naranja y los genes afectados. (B) Distribución poblacional de los 3 haplotipos en 749 individuos de poblaciones *HGDP-CEPH*. Sólo se muestran las poblaciones con al menos 10 individuos genotipados. Imagen tomada de Pang et al. 2013.

Por último, dos inversiones (*HsInv0001* y *HsInv0042*) fueron descartadas como errores en el ensamblaje del genoma de Referencia por el consorcio encargado de su revisión, *GRC* (del inglés Genome Reference Consortium). En total, 16 inversiones descubiertas por Levy y colaboradores en la comparación de *HuRef* y el genoma de Referencia han sido analizadas experimentalmente en dos estudios [Korbel et al. 2007] [Pang et al. 2013].

1.6 Objetivos

El objetivo principal de esta tesis es analizar bioinformática y experimentalmente las inversiones descubiertas por la comparación de los genomas de J. Craig Venter y de Referencia que han sido ensamblados de forma independiente. En concreto, la validación de las inversiones y el descarte de posibles falsos positivos contribuye a la construcción de un catálogo fiable y no redundante de inversiones en el genoma humano. Además, la genotipación de individuos de diferentes poblaciones humanas y otras especies también es un punto importante, ya que permite tanto analizar la distribución poblacional de las inveriones como entender mejor su origen. Con este estudio se espera aportar conocimiento sobre las características de sus puntos de rotura, mecanismos de formación, estado ancestral, origen, distribución, posibles efectos sobre genes y posibles efectos adaptativos en el genoma humano, para de este modo determinar el impacto funcional y evolutivo de las inversiones en el genoma humano.

Los objetivos concretos son los siguientes:

1. Análisis exhaustivo de las diferentes inversiones predichas para descartar falsos positivos y validación experimental de las verdaderas inversiones polimórficas.

2. Definición de los puntos de rotura con la máxima precisión posible y determinación del mecanismo de origen y el estado ancestral de las diferentes inversiones polimórficas identificadas.

3. Estudio de la frecuencia y distribución de las inversiones en distintas poblaciones humanas.

4. Análisis de la variación nucleotídica y haplotípica de las dos ordenaciones.

5. Identificación de las inversiones candidatas a tener efectos funcionales sobre genes así como de las que muestren patrones indicativos de selección positiva.

2. MATERIALES Y MÉTODOS

2. MATERIALES Y MÉTODOS

2.1. Obtención de la secuencia, alineamiento, definición de los puntos de rotura y anotación de las inversiones.

En primer lugar, se obtuvieron las coordenadas de las 90 putativas inversiones potenciales predichas por Levy y colaboradores [Levy et al. 2007], en el genoma de J. Craig Venter, *HuRef*, de la sección *Additional Data* en la página web del proyecto (http://huref.jcvi.org). Las coordenadas corresponden al ensamblaje *NCBI36/hg18* del genoma humano de Referencia publicado en marzo del año 2006. Se extrajo la secuencia correspondiente a las regiones publicadas usando el browser genómico *UCSC* (http://genome.ucsc.edu/) [Kent et al. 2002] y añadiendo 5 Kb de secuencia flanqueante a cada lado. A continuación, se realizó una búsqueda mediante *Blastn* [Altschul et al. 1990] en otros genomas humanos disponibles en la web del *National Center for Biotechnology Information*, *NCBI* (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome) para identificar las secuencias correspondientes en *HuRef* y obtener las coordenadas. Éstas permitieron extraer la secuencia a partir de los cromosomas ensamblados de *HuRef*, que están disponibles en el archivo *ftp* del *NCBI* (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.3/Assembled_chromosomes/). Para cada región, las dos secuencias genómicas correspondientes a *HG18* y *HuRef* se alinearon mediante *Blast2seq* [Altschul et al. 1990.] En el alineamiento se distinguieron las partes de la secuencia que alineaban en la misma orientación que corresponden a las secuencias flanqueantes de 5 Kb, de las partes que alineaban en orientación invertida que corresponden a la región prueba (**Figura 2.1**). En ocasiones se usaron más de 5 Kb de secuencia flanqueante para obtener alineamiento en orientación directa, debido a la presencia de inserciones, deleciones, duplicaciones segmentales o *gaps*.

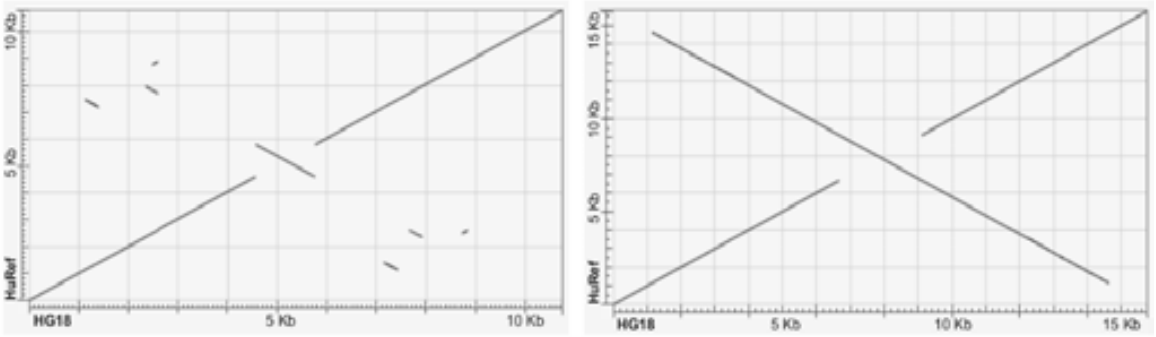

Figura 2.1: Ejemplo de representación *dot plot* de un alineamiento. El programa *Blast2seq* [Altschul et al. 1990] genera este tipo de gráficos para representar el alineamiento de dos secuencias. En el eje de las abscisas está representado el genoma de Referencia y en el de las ordenadas, *HuRef*. Cada punto en el gráfico representa un nucleótido alineado y la sucesión de puntos forman las líneas que representan los fragmentos del alineamiento. Los fragmentos invertidos están representados por líneas con pendiente negativa y los fragmentos que alinean en la misma orientación por líneas con pendiente positiva, que en este caso corresponden a la secuencia flanqueante. En el ejemplo de la izquierda se muestra una inversión con puntos de rotura sencillos, es decir, no localizados en repeticiones invertidas, por lo que no hay solapamiento de los fragmentos directos con el fragmentos invertido. En el ejemplo de la derecha podemos ver que sí hay solapamiento, esto es debido a que los puntos de rotura de la inversión se encuentran en duplicaciones segmentales invertidas, que evidentemente pueden alinear en dos lugares y dos orientaciones, esto hace que los gráficos tomen forma de cruz, donde la línea de pendiente negativa que no solapa con ningún otro fragmento representa la región invertida.

Se comprobó manualmente que las secuencias mantienen una estructura *A-B-C-D* para la orientación estándar de referencia y *A-C'-B'-D* para la orientación invertida, correspondiente a *HuRef*. *A* y *D* corresponden a las secuencias flanqueantes y por lo tanto tienen la misma orientación en ambos genomas. Las secuencias *B* y *C* están situadas dentro de la región invertida, se tomó como referencia la orientación del genoma de Referencia, por lo tanto estándar. En *HuRef*, estas secuencias tienen un orden invertido, *C'-B'*, donde *C'* es la secuencia reversa complementaria de *B* en la orientación estándar, y lo mismo ocurre con *B'* y *C*. Se definieron los puntos de rotura en las intersecciones de las secuencias en el primer punto de rotura, *A-B* para el genoma de Referencia y *A-C'* para *HuRef* y *C-D* y *B'-D* en el segundo. Se definieron las coordenadas como un rango para incluir los puntos de rotura que están localizados en repeticiones invertidas (RIs). Precisamente en estas inversiones se alinearon las RIs para definir en qué parte se encuentran los puntos de rotura. Se usó el programa de alineamiento múltiple, *MUSCLE* [Edgar et al. 2004], para alinear las RIs de las secuencias de ambos genomas. El análisis manual del alineamiento múltiple se realizó a través del programa *Bioedit 7.2.5* [Hall et al. 1999]. Se identificaron los cambios nucleotídicos que diferencian ambas RIs en la secuencia estándar y dónde se intercambian en *HuRef* por la aparición de la inversión (**Figura 2.2**). Los puntos de rotura se definieron así en el intervalo formado entre tres diferencias entre las dos copias parálogas de las RIs sin intercambio y tres diferencias

intercambiadas entre las RIs, correspondientes a la orientación invertida (**Figura 2.2**).

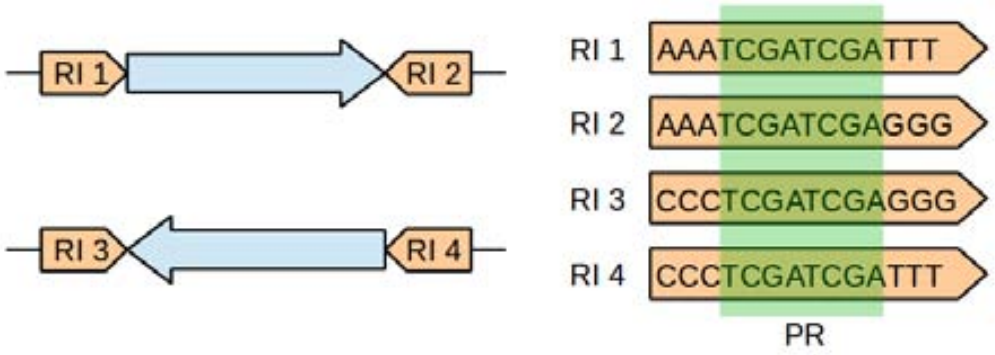

Figura 2.2: Definición de los puntos de rotura en RIs por alineamiento múltiple. En la parte izquierda se muestra el esquema de las RIs de ambas orientaciones. En la parte derecha se muestra el alineamiento múltiple de las cuatro RIs, en verde se muestra la región del punto de rotura, de secuencia idéntica en todas las RIs, delimitada a la izquierda por tres diferencias sin intercambio entre las RIs correspondientes a una y otra orientación; y a la derecha por tres diferencias con intercambio entre las RIs de la orientación invertida.

Una vez definidos los puntos de rotura, se procedió a anotar su secuencia junto a la de la inversión, en la región formada por la posible inversión y las regiones flanqueantes de 5 Kb. Se usó el programa *CLC Main Workbench 7.0.3* [http://www.clcbio.com]. También se anotaron las repeticiones invertidas en el caso de las inversiones con puntos de rotura localizados en RIs y cualquier elemento que afecta a los puntos de rotura como por ejemplo, elementos móviles, inserciones, deleciones, secuencias de micro-homología y *SNPs* (**Figura 2.3**).

Además se anotaron de forma manual los genes localizados en la secuencia completa, es decir, en la región invertida con sus puntos de rotura y en las secuencias flanqueantes. Para ello se usó el navegador genómico *UCSC* (http://genome.ucsc.edu/) [Kent et al. 2002]. En primer lugar, se usó la utilidad *liftover* del propio navegador para convertir las coordenadas del ensamblaje *HG18* del genoma de Referencia *HG18* a las correspondientes del ensamblaje *HG19*. Después se visualizó la información correspondiente a genes *UCSC*, ARNs mensajeros humanos, *ESTs* procesadas y elementos funcionales del proyecto *GENCODE* [Harrow et al. 2012], y se anotó de manera manual sobre la secuencia completa. Este análisis manual de los genes afectados por los efectos posicionales de las inversiones se extendió a los genes localizados en las 50 Kb flanqueantes a los puntos de rotura, ya que pueden tener afectada su regulación; pero no fueron anotados al no estar localizados en la región invertida o en los puntos de rotura de las potenciales inversiones.

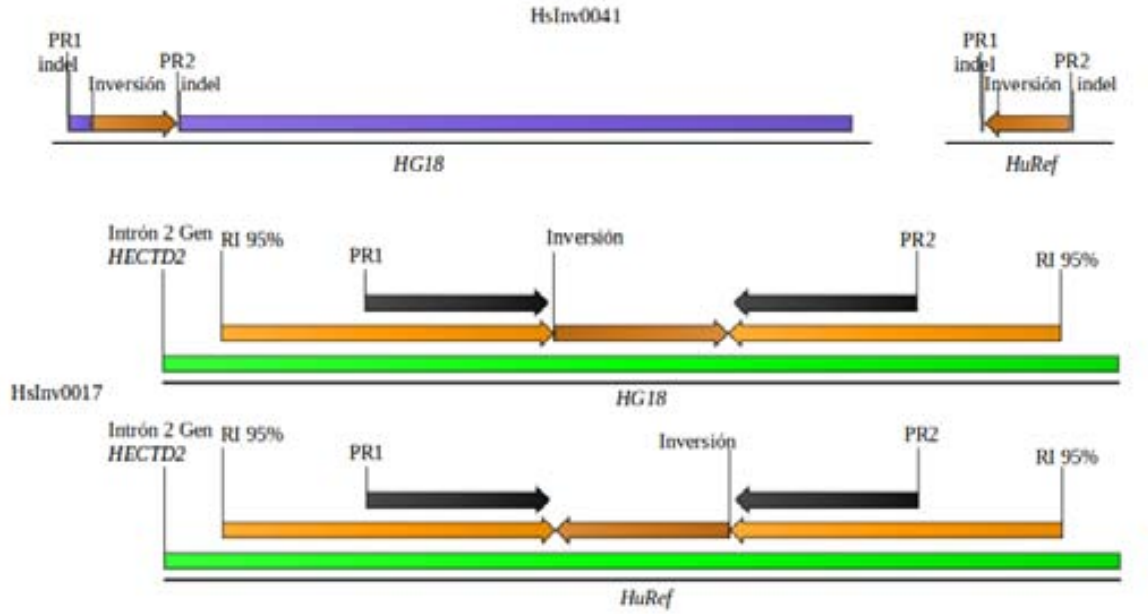

Figura 2.3: Ejemplos de anotación de las potenciales inversiones. Se muestran los elementos anotados sobre la secuencia representada por una línea negra. Arriba el esquema representa la anotación de la inversión *HsInv0041* para el genoma de Referencia *HG18* y para *HuRef*. Destacan en violeta las inserciones/deleciones. Abajo se muestra el esquema de la inversión *HsInv0017*, donde se anotaron las repeticiones invertidas RIs. Además se puede ver la anotación del gen *HECTD2*, ya que toda la región está localizada en el interior del intrón 2 de este gen.

Finalmente, se compararon manualmente los puntos de rotura publicados por Levy y colaboradores [Levy et al. 2007] con los resultados de análisis manual en nuestro estudio para evaluar su definición. Se consideraron bien definidos los puntos de rotura con una diferencia igual o inferior a 3 nucleótidos respecto a las coordenadas resultantes de nuestro análisis. Se tuvo en cuenta la definición correcta de ambos puntos de rotura para concluir que una inversión estuvo bien definida en el artículo original.

2.2 Soporte por mapeo de extremos apareados (PEM) de fósmidos

Una vez comprobado el alineamiento invertido entre ambas secuencias y definidos sus puntos de rotura, se procedió a analizar el soporte de fragmentos con extremos apareados, proveniente de los individuos usados en el *PEM* realizado por Kidd y colaboradores en el año 2008 [Kidd et al. 2008]. La información proviene de las librerías de fósmidos usadas para detectar variación estructural en 9 individuos procedentes de diferentes poblaciones (NA12156, NA12878, NA15510, NA18507, NA18517, NA18555, NA18956, NA19129 y NA19240). Los extremos de los insertos usados en *PEM*, en este caso fósmidos, se secuencian en dirección opuesta entre sí, por lo que en un fósmido concordante que tiene la misma orientación que el genoma de Referencia, tienen orientaciones invertidas, +/-. Por el contrario, los fósmidos discordantes tienen sus extremos mapeados en la misma

orientación, +/+ o -/-. El análisis consistió básicamente en comprobar qué tipo de fósmidos se localizan en la misma zona que los puntos de rotura (**Figura 2.4**). En el caso de tratarse de inversiones polimórficas en el genoma humano, encontraríamos fósmidos concordantes y discordantes en los puntos de rotura, siempre y cuando la cobertura de la región por la librería fuese suficiente. Por otro lado, la presencia solo de fósmidos concordantes para todos los individuos, indicaría un posible error en *HuRef*; de la misma manera que de tratarse de fósmidos discordantes se trataría de un posible error en el genoma de Referencia *HG18*. En el análisis se contaron el número de fósmidos concordantes y discordantes de los 9 individuos en cuestión, que tienen un extremo alineado dentro de la región invertida y el otro extremo alineado fuera, sin tener en cuenta los mapeos de extremos en RIs. Las regiones con menos de 3 fósmidos mapeados no se clasificaron.

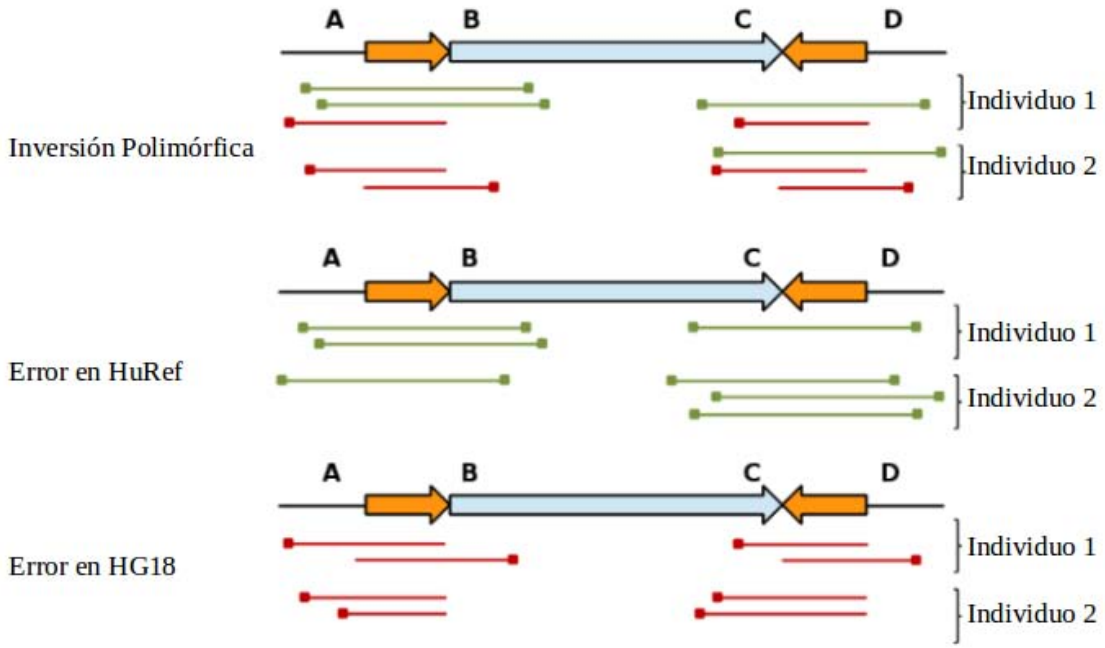

Figura 2.4: Ejemplos de análisis del soporte por *PEM*. Se muestra la orientación estándar de la zona potencialmente invertida, representada por una flecha azul, y las RIs donde se encuentran los puntos de rotura representadas por flechas de color naranja. Los fósmidos concordantes están representados por líneas verdes y sus extremos por cuadrados del mismo color. Las líneas y cuadrados rojos representan los fósmidos discordantes y sus extremos. Arriba se muestra una inversión potencialmente polimórfica con soporte de fósmidos concordantes y discordantes, en medio una inversión que es un error potencial en *HuRef*, indicado por los fósmidos concordantes en los puntos de rotura y abajo un error potencial en el genoma humano de Referencia, indicado por los fósmidos discordantes.

Se realizó el análisis de manera manual utilizando el navegador genómico del proyecto de Variación Estructural en el Genoma Humano *HGSV* [The Human Genome Structural Variation Working Group. 2007]. Éste consiste en una instalación propia del navegador genómico *USCS* [Kent et al. 2002], que contiene el ensamblaje *HG18* del genoma de

Referencia y la información del mapeo de los extremos de las librerías de fósmidos de varios estudios, entre otros del estudio mencionado anteriormente. Se encuentra disponible en la web (http://hgsv.washington.edu/).

2.3. Muestras de ADN

Se usaron 96 muestras correspondientes a individuos de la fase I del proyecto HapMap [The International HapMap Consortium. 2005], incluyendo 90 individuos de origen Europeo (población CEU) divididos en 30 trios padres-hijo, 4 individuos independientes de origen Africano (población YRI) y 2 individuos independientes de origen Asiático (poblaciones CHB y JPT). Además, se usó la muestra de ADN del individuo *NA15510* de origen desconocido, que había sido usado previamente por Kidd y colaboradores en el año 2008 para generar librerías de fósmidos [Kidd et al. 2008]. El ADN genómico de cada individuo se obtuvo a partir de líneas celulares de limfoblastocitos B transformados por virus *Epstein-Barr* [Coriell Cell Repositories, Camden, New Jersey, USA]. Para la mayoría de individuos la extracción se realizó a partir de 10 ml de cultivo celular crecido conforme el procedimiento recomendado, mediante un método basado en fenol cloroformo, modificado para obtener ADN de alto peso molecular [Sambrook and Russell 2001] [Aguado et al. 2014]. Posteriormente se confirmó la identidad de todas las muestras de ADN mediante el *kit* de microsatélites *MSK* [Coriell Cell Repositories, Camden, New Jersey, USA]. El ADN de J. Craig Venter (*HuRef*) y de otras muestras de las que no se disponía la línea celular, se adquirió directamente de *Coriell Cell Repositories* [Coriell Cell Repositories, Camden, New Jersey, USA]. Por otro lado, los clones *BAC* usados en el ensamblaje del genoma humano de Referencia se adquirieron de *CHORI BACPAC Resources Center* (Oakland, California, USA), de *RIKEN Bioresource Center DNA Bank* (Ibaraki, Japón) y de *Source BioScience* (Nottingham, UK). Las bacterias crecieron en placas de agar LB con 30 μg/ml de cloranfenicol o kanamicina dependiendo del clon. Se extrajo el ADN del *BAC* mediante el *Plasmid Mini Kit* (Qiagen). También se usaron muestras de ADN de 4 chimpancés y 2 gorilas incluyendo una pareja padre-hijo de cada especie. El ADN genómico del chimpancé *N457/03* y de los dos gorilas *Z01/03* y *Z02/03* se extrajo de tejido procedente de córtex frontal obtenido de *Banc de Teixits Animals de Catalunya* (BTAC, Bellaterra, Barcelona, España). El ADN genómico de los 3 chimpancés restantes *PTR1211*, *PTR1213* y *PTR1215* se extrajo de líneas celulares de limfoblastocitos B transformados por virus *Epstein-Barr*, generadas a partir de muestras de sangre provenientes del Zoo de Barcelona. Todos los procedimientos que incluyeron el uso de muestras humanas y de primates fueron previamente aprobados por la Comisión de Ética en la Experimentación Animal y Humana (CEEAH) de la Universitat Autònoma de Barcelona.

2.4. Diseño de cebadores, validación por *PCR* y *PCR* inversa (iPCR) y genotipación experimental

Los cebadores que se usaron en la validación por *PCR* y *iPCR*, se diseñaron usando la aplicación *web Primer3Plus* [Untergasser et al. 2007]. Se estableció una temperatura de fusión *Tm* de entre 58 ºC y 62 ºC y se determinó el tamaño máximo de los productos a amplificar en 4 Kb. El tamaño óptimo de los cebadores se estableció en 21 pb y el mínimo y máximo en 19 y 26 pb respectivamente, aunque algunos cebadores sobrepasaron estos límites por la dificultad de su diseño en determinadas regiones del genoma. Se estableció la complementariedad máxima de su extremo 3' en 2 bases, tanto consigo mismos como con su pareja, salvo en casos de difícil diseño en que se admitieron 3 bases. Se comprobó la ausencia de *SNPs* mediante el navegador genómico *USCS* [Kent et al. 2002], usando todas las capas de información disponibles sobre ellos provenientes de la base de datos *dbSNP*, tanto en el ensamblaje *HG18* como en *HG19*. En el protocolo normal de la *PCR* se usaron parejas de cebadores flanqueando los puntos de rotura de las inversiones potenciales para determinar la orientación de la secuencia. En cada punto de rotura, cada una de las dos parejas de cebadores fue asociada a una de las orientaciones de la secuencia, para amplificar un fragmento de un tamaño diferente al de la otra pareja. De esta manera se pudieron diferenciar los productos por electroforesis en un gel de agarosa. Además se realizó una comprobación adicional para descartar la generación de productos inespecíficos en la *PCR*, una búsqueda de cuántos fragmentos se pueden amplificar en el genoma humano para cada pareja de cebadores. Debido a la repetitividad del genoma humano, se pueden amplificar fragmentos inespecíficos más pequeños o del mismo tamaño que el producto deseado, que pueden interferir en la *PCR*. Se comprobó mediante *primer-BLAST* que no se amplificasen fragmentos de este tipo [Ye et al. 2012]. Finalmente, siempre que fue posible, se diseñaron los cebadores externos, los que hibridan fuera de la región invertida, para que fueran compatibles con ambos cebadores internos, que son específicos de cada orientación en un punto de rotura. De esta manera y teniendo en cuenta que amplificasen fragmentos de distinto tamaño, se diseñaron para realizar una *PCR* multiplex, es decir, una *PCR* en la que se amplifican a la vez fragmentos de la orientación estándar y de la orientación invertida, siempre que el individuo sea heterocigoto para la inversión. Si no lo es, se amplifica solo el fragmento correspondiente al alelo presente (**Figura 2.5**). En este caso, se comprobó que los dos cebadores específicos de cada alelo no formaran una pareja que amplificase fragmentos que pudiesen interferir con el producto de la *PCR* debido a su tamaño y que permitiesen diferenciar los productos en la electroforesis. Además, siempre que no hubiera problemas de incompatibilidad, se usaron los cebadores internos para ambas orientaciones conjuntamente con el cebador externo del mismo punto de rotura.

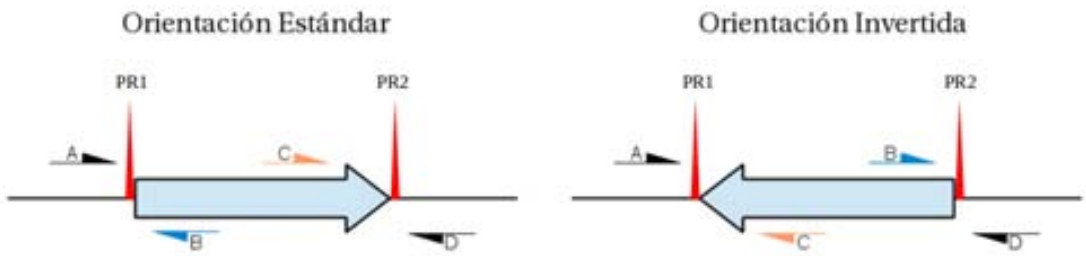

Figura 2.5: Esquema del diseño experimental de *PCR* para validar inversiones cromosómicas. En la parte izquierda se muestra la orientación correspondiente al alelo estándar y en la parte derecha la del alelo invertido. El nombre de los cebadores coincide con el esquema *A-B-C-D* de las secuencias alrededor de los puntos de rotura (PR) en la orientación estándar. Se diseñaron 3 cebadores para cada punto de rotura, uno común, de color negro, localizado fuera de la inversión, que es específico del punto de rotura; y dos específicos de cada orientación localizados dentro de la región invertida. Generalmente estos dos cebadores se pueden usar para ambos puntos de rotura, por ejemplo el cebador *B* de color azul es específico del alelo estándar para el primer punto de rotura y del alelo invertido para el segundo punto de rotura. Las dos parejas de cebadores implicados en cada punto de rotura se diseñaron para amplificar fragmentos de tamaño diferente de manera que los productos se pueden diferenciar en un gel de agarosa, por ejemplo *A-B* y *A-C* para el primer punto de rotura.

Las reacciones de *PCR* se prepararon en un volumen total de 25 µl, que contenía 1x *buffer*, 1.5 mM $MgCl_2$, 0.2 µM de cada *dNTP*, 0.4 µM de cada cebador, 1.5 *U* de *Taq polimerasa* (BioTherm) y 50-100 ng de ADN genómico. En el caso de las *PCR* multiplex, se usó una concentración final de 0.8 µM para el cebador común compartido por los fragmentos específicos de uno y otro alelo, y de 0.4 µM para el resto. Cada reacción siguió el siguiente protocolo, en primer lugar un paso de desnaturalización de 5 minutos a 95 ºC, seguido de 30-35 ciclos a 95 ºC por 30 segundos, un paso de apareamiento de 59-62 ºC por 30 segundos, un paso de extensión por 20-120 segundos y un paso de extensión final de 7 minutos a 72 ºC. En las *PCR* de rango largo que amplifican productos de 5 a 10 Kb, se usaron 100-200 ng de ADN genómico y 2.5 unidades de ADN polimerasa *Pfu Turbo* (Stratagene). Las condiciones de reacción fueron 92 ºC por 2 minutos, 35 ciclos a 92 ºC por 10 segundos, 60-66 ºC por 30 segundos, 68 ºC por 10-15 minutos y 68 ºC por 10 minutos. Los productos de *PCR* se analizaron por electroforesis en gel de agarosa al 1.5-2% teñido con bromuro de etidio; para los productos de *PCR* más largos se usaron porcentajes de entre 0.8-1%. En el caso de las amplificaciones a partir de clones *BAC*, no se extrajo el ADN, se resuspendió una colonia del clon en 100 µl de *TE* (con una proporción 10:0.1 de Tris:EDTA) y se usaron 2 µl como molde. En la *PCR* inversa, el ADN genómico se cortó con enzimas de restricción que generan extremos cohesivos. Las enzimas que se usaron tienen dianas de restricción dentro y fuera de la región invertida, pero no en los puntos de rotura. Se muestran en la **Tabla 2.1**.

Tabla 2.1: Enzimas de restricción usadas en el protocolo de *PCR* inversa.

Identificador	Enzima
HsInv0023	*NsiI*
HsInv0024	*NsiI*
HsInv0029	*EcoRI*
HsInv0031	*EcoRI*
HsInv0038	*NsiI*
HsInv0040	*HindIII*
HsInv0045	*SacI*
HsInv0046	*NsiI*
HsInv0052	*HindIII & NsiI*
HsInv0053	*NsiI*
HsInv0055	*BamHI*
HsInv0057	*KpnI*
HsInv0061	*HindIII*
HsInv0067	*ApaI*
HsInv0069	*NsiI & EcoRV*
HsInv0072	*HindIII*
HsInv0090	*NsiI*

De manera general se digirieron 150 ng de ADN genómico durante toda la noche a 37 ºC, usando 3 unidades de enzima de restricción. Después se inactivó la enzima de restricción mediante la incubación a alta temperatura (temperatura de inactivación por calor específica de cada enzima). Posteriormente se circularizó el ADN digerido, mediante una reacción de ligación a 25 ºC durante 3 horas. Las proporciones usadas en las reacciones de ligación fueron 1x *buffer* y 400 unidades de ligasa de ADN *T4* (New England Biolabs) en un volumen total de 175 µl. Se inactivó por calor la enzima ligasa durante 10 minutos a 65 ºC y finalmente se usaron 10 µl de producto de ligación (aproximadamente 8.6 ng de ADN ligado) como molde en la reacción de *PCR*. Los cebadores amplificaron alrededor de los sitios de restricción y ligación. En el caso concreto de la inversión *HsInv0090* se llevó a cabo una digestión parcial con la enzima *NsiI*. Se digirió el ADN durante 40 minutos después de probar diferentes tiempos de digestión entre 10 minutos y 3 horas. La digestión parcial permite usar enzimas de restricción que tienen dianas en los puntos de rotura, ya que no se digiere totalmente el ADN, de manera que se generan los fragmentos deseados, en los que la enzima ha cortado dentro y fuera de la inversión pero no en el punto de rotura.

En la validación por *PCR* y *iPCR*, se seleccionaron las inversiones potencialmente polimórficas según su soporte por *PEM*, que tuviesen tamaños mayores o iguales a 1 Kb o que pudiesen tener efectos posicionales de sus puntos de rotura sobre genes. La validación consistió en la genotipación de 10 individuos, 9 individuos de distintas poblaciones humanas (anteriormente nombrados) [Kidd et al. 2008], y J. Craig Venter. Una inversión fue validada como polimórfica en el genoma humano cuando se encontraron ambos alelos entre los cromosomas de los 10 individuos. En ese caso, fueron

genotipadas experimentalmente del mismo modo en 90 individuos Europeos. En caso de amplificar solo alelos invertidos, se genotipó el *BAC* correspondiente a la región en el genoma de Referencia para analizar la posible existencia de un error en el ensamblaje. En el caso de amplificar solo alelos estándar en *HuRef*, se comprobó su secuencia mediante la secuenciación de los fragmentos amplificados y el alineamiento con el genoma de Referencia.

2.5. Secuenciación

En aquellas situaciones en que fue necesario analizar si el producto amplificado en la *PCR* correspondía con el producto esperado, se secuenció la banda separada en la electroforesis. Es el caso de los errores en *HuRef*, regiones donde la muestra de ADN de J. Craig Venter amplificó productos correspondientes a la orientación estándar (**Figura 2.6**).

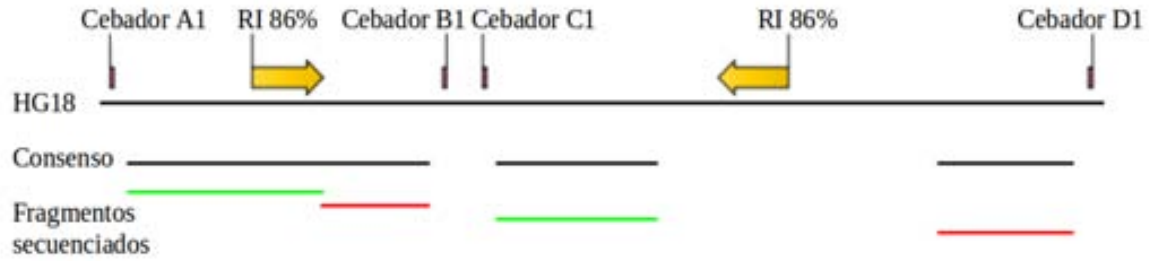

Figura 2.6: Alineamiento de los fragmentos secuenciados con el genoma de Referencia. Se muestra la secuencia del genoma de Referencia *HG18* representada por una línea negra, donde se han anotado los cebadores y las RIs, con un 86% de identidad, de color amarillo. En este ejemplo la secuencia amplificada a partir de *HuRef* tiene la misma orientación que el genoma de Referencia. Se trata del error en *HuRef HsInv0078*. Los fragmentos se amplificaron a partir de los cebadores que se muestran en la parte superior de la muestra de ADN de J. Craig Venter. Se muestran en verde los fragmentos amplificados a partir de cebadores directos, como *A1* y *C1*, y en rojo los fragmentos de cebadores reversos, *B1* y *D1*. En negro debajo de la secuencia del genoma de Referencia se muestra la secuencia consenso formada por las secuencias de los fragmentos amplificados, que alinea perfectamente con el de Referencia.

En primer lugar se separaron los productos de *PCR* mediante electroforesis en gel de agarosa y se extrajo el *ADN* mediante el *kit Qiaquick Gel Extraction Kit* (Qiagen). Los productos de *PCR* se secuenciaron mediante el método *Sanger* por la empresa *Macrogen* en *Seoul, Korea*. Los fragmentos secuenciados se alinearon con el genoma de referencia para comprobar si se trataba de la misma secuencia. Para ello se usó el programa *CLC Main Workbench 7.0.3* [http://www.clcbio.com].

2.6. Genotipación bioinformática *in sílico*

Se usaron los datos de secuenciación de los individuos correspondientes a la primera fase del proyecto de los 1000 Genomas [Abecasis et al. 2012] para genotipar las inversiones, mediante el mapeo de *reads* en los puntos de rotura. Se genotiparon 1092 individuos no relacionados pertenecientes a 14 poblaciones humanas de todo el mundo para 7 inversiones polimórficas con puntos de rotura no localizados en RIs. En primer lugar se preparó una librería de puntos de rotura, cada secuencia contenía el punto de rotura y 50 pb de secuencia flanqueante a cada lado. Para cada inversión se usaron 4 secuencias correspondientes a las dos orientaciones y los dos puntos de rotura. Además se añadieron a la librería secuencias adicionales en representación de inserciones o deleciones, en los casos en que se detectaron en los puntos de rotura, con el objetivo de incluir todas las combinaciones posibles. En la **Tabla suplementaria S8** del apartado de resultados se muestra la librería. Se descargaron los *reads* de los individuos secuenciados del servidor *ftp* del proyecto de los 1000 Genomas [Abecasis et al. 2012]. En el caso de los individuos para los que está disponible el alineamiento de los *reads* en formato *SAM*, se usó *SAMtools* [Li et al. 2009] para descargar solo los *reads* correspondientes a las regiones de los puntos de rotura, además de los *reads* no mapeados. Los *reads* en formato *SAM* fueron convertidos a formato *fastq*. En los casos en que no estaba disponible el alineamiento de los *reads*, se descargaron los archivos *fastq* que contienen la información de todos los *reads* en bruto, evitando otras fuentes como exomas, etc. El genotipado se realizó mediante una versión ligeramente modificada del *software BreakSeq* [Lam et al. 2010] [Lucas-Lledó et al. 2014].

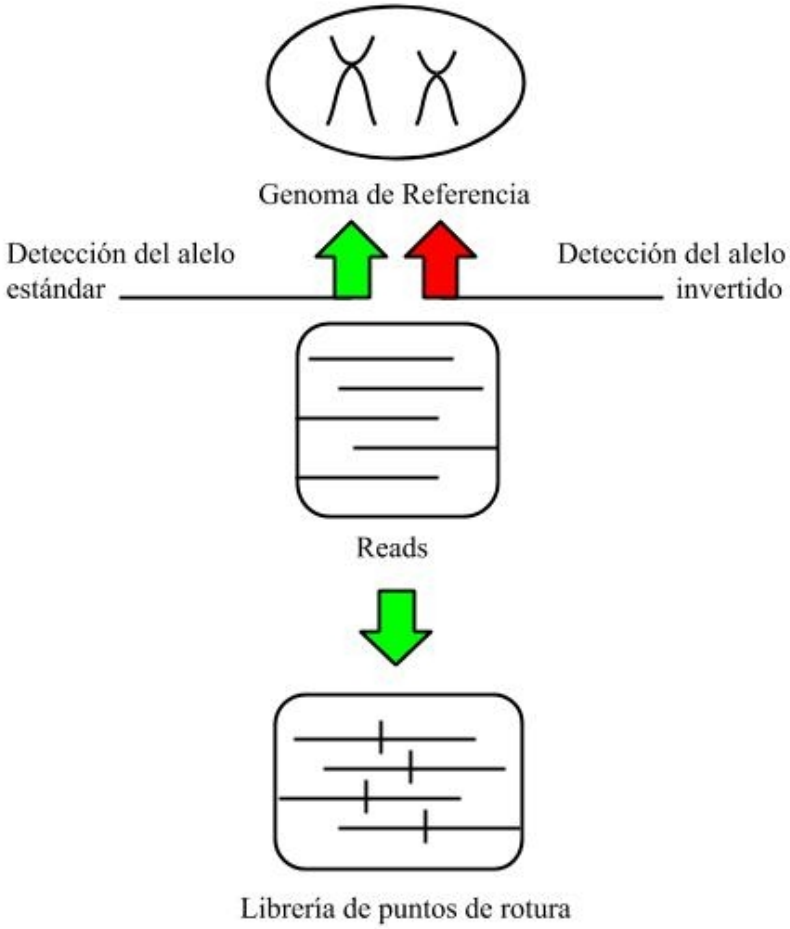

Figura 2.7: Esquema del funcionamiento del software *BreakSeq*. Se muestra como el programa determina si las observaciones corresponden a un alelo estándar o invertido. Las flechas verdes indican que hay alineamiento y las rojas que no hay alineamiento. Los *reads* que alinean de forma única en un punto de rotura de la región correspondiente a la inversión en el genoma de Referencia y en una secuencia correspondiente al alelo estándar de la librería de puntos de rotura, se consideran observaciones del alelo estándar. Por el contrario, los *reads* que no alinean en los puntos de rotura de la inversión el genoma de Referencia y sí en una secuencia correspondiente al alelo invertido de la librería, se consideran observaciones del alelo invertido.

En primer lugar se mapearon los *reads* a la librería de puntos de rotura mediante el programa de alineamiento *Bowtie2*, una modificación respecto al *BreakSeq* original que usa *Bowtie*. Los *reads* que alinearon en la región de cualquier punto de rotura presente en la librería y 10 bases a un lado y a otro, se guardaron y fueron alineados con el genoma de Referencia. Se conservaron los *reads* que mapearon en la secuencia correspondiente a la orientación estándar para un punto de rotura en la librería, y de éstos, sólo se conservaron los que mapearon de manera única en la región del punto de rotura en el genoma de Referencia. El resto con alineamientos múltiples fueron descartados para evitar errores de genotipado debidos a la repetitividad del genoma humano. También se descartaron por el mismo motivo, los *reads* que mapearon en la secuencia correspondiente al alelo invertido en la librería de puntos de rotura y en el genoma de Referencia. De esta manera, los *reads*

restantes que mapearon en las secuencias correspondientes a ambos alelos en la librería de puntos de rotura contaron como observaciones de uno y otro alelo en la genotipación (**Figura 2.7**).

Lamentablemente, debido a la poca cobertura de los genomas analizados (2-6x), las observaciones de los alelos fueron insuficientes para determinar el genotipo de cada individuo de manera fiable, ya que existía el riesgo de genotipar como homocigotos individuos heterocigotos. Esta limitación nos llevó a utilizar el número de *reads* mapeados para estimar los genotipos y la frecuencia alélica gracias al programa *svgem* [Lucas-Lledó et al. 2014]. Este programa implementa un algoritmo de esperanza-maximización que determina la incertidumbre de cada genotipo y entrega estimas de la frecuencia alélica basadas en la máxima verosimilitud. Además tiene en cuenta el sesgo en la observación de los alelos, es decir, la mayor probabilidad de observar uno de los dos alelos en un individuo heterocigoto (**Tabla 2.2**).

Tabla 2.2: Datos sobre el sesgo de detección estimado para ambos alelos de las 7 inversiones genotipadas bioinformáticamente.

Identificador	Valor de landa
HsInv0003	1.02454
HsInv0006	3.39188
HsInv0041	0.907808
HsInv0058	1.91213
HsInv0059	1.24146
HsInv0063	0.906018
HsInv0068	1.04286
HsInv0095	0.974667

Se usaron resultados de la genotipación por *PCR* para comprobar que los individuos homocigotos no mostraban recuentos espurios para el alelo ausente. Se detectó un número muy bajo de observaciones erróneas por lo que se las consideró negligibles y se fijó el parámetro correspondiente en 1.0E-05. Finalmente se ejecutó *svgem* sobre los recuentos de observaciones de alelos para las 7 inversiones, sin asumir equilibrio de Hardy-Weinberg, y se estimaron las frecuencias de los alelos invertidos para cada población, por grupos continentales y también a nivel global.

2.7. Análisis de la variación nucleotídica

Para analizar la variación nucleotídica asociada a estas inversiones se obtuvo la información relativa a *SNPs* para la región de la inversión y 10 Kb de secuencia flanqueante a cada punto de rotura. Los *SNPs* provinieron de la fase I del proyecto de los

1000 Genomas [Abecasis et al. 2012] y del proyecto HapMap en su publicación de datos 27 [Altshuler et al. 2010]. En el caso de los *SNPs* del proyecto *1000GP* se obtuvieron datos para 35 individuos no relacionados, genotipados experimentalmente para las 17 inversiones. En el caso del proyecto HapMap se obtuvieron datos para 60 individuos independientes. No se incluyeron los *SNPs* localizados en RIs.

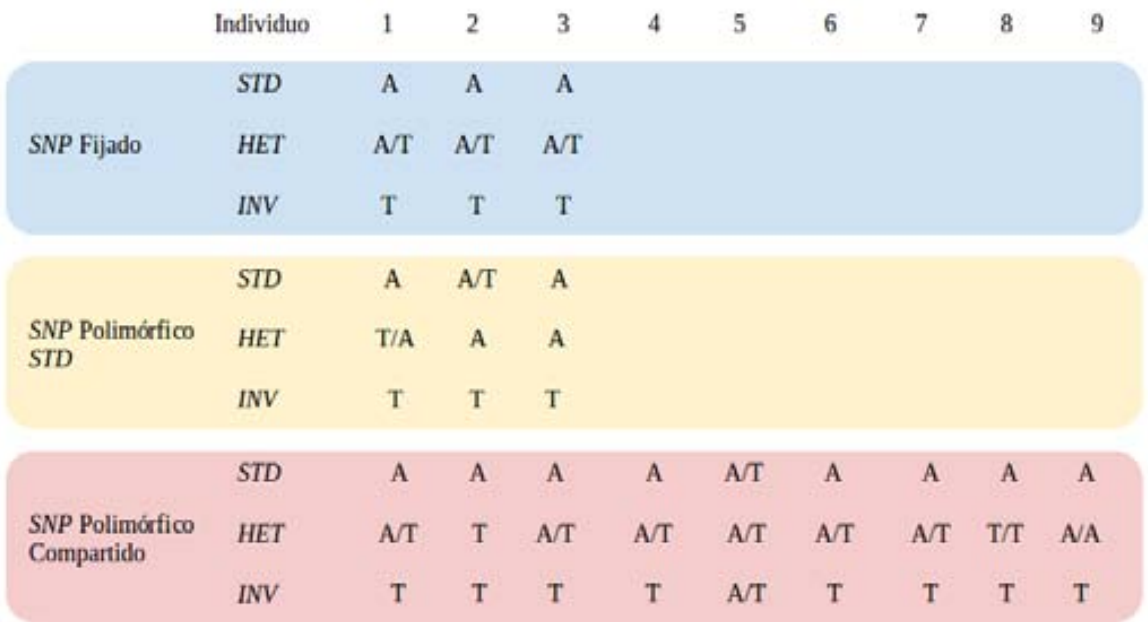

Figura 2.8: Esquema de *SNPs* fijados y compartidos entre ambas orientaciones de las secuencia. Se muestra en la parte superior el identificador de individuo y en la izquierda el tipo de *SNP* detectado en relación con los alelos de la inversión. Cada fila contiene la información del genotipo de cada individuo para la inversión (*STD* corresponde a homocigoto estándar, *HET* a heterocigoto e *INV* a homocigoto para la inversión) y para el *SNP*. Se muestran ejemplos de cómo se detecta un *SNP* fijado, un *SNP* polimórfico para la orientación estándar y 3 casos de *SNPs* polimórficos compartidos entre ambos alelos de la inversión.

El análisis de *SNPs* consistió en la búsqueda de los fijados y compartidos entre ambos alelos de la inversión. Se usaron *scripts* en *Perl* y *Bash* para relacionar los alelos de los *SNPs* y de las inversiones [Aguado et al. 2014]. Los *SNPs* fijados segregan siempre con la inversión, es decir, uno de los alelos se encuentra siempre junto a uno de los alelos de la inversión. Estos *SNPs* indican que la inversión se ha generado una vez y que ha inhibido la recombinación dentro de la región invertida en los individuos heterocigotos, de manera que al divergir las dos orientaciones capturan un alelo de cada *SNP*. Los *SNPs* compartidos o polimorfismos compartidos entre los cromosomas invertidos y estándar se detectaron por la presencia de *SNPs* polimórficos en los individuos homocigotos para el alelo estándar y el alelo invertido, es decir, al detectar los dos alelos de un *SNP* en un individuo homocigoto para la inversión o para el alelo estándar (**Figura 2.8**). Otra situación en que se detectaron *SNPs* compartidos fue en los individuos heterocigotos para la inversión en los que unos individuos son homocigotos para un alelo del *SNP* y otros

para el otro [Aguado et al. 2014]. Por último, también se consideraron *SNPs* compartidos en individuos heterocigotos para la inversión, los *SNPs* polimórficos en un alelo de la inversión que no lo son en el otro. Estos *SNPs* indican que ha habido recombinación en la región invertida de manera que uno de los alelos del *SNP* no está siempre junto a la misma orientación.

El desequilibrio de ligamiento entre los *SNPs* y las inversiones polimórficas se determinó mediante el programa *Haploview 4.2* [Barret et al. 2005] y el estadístico r^2. Se usaron los genotipos de los individuos *Europeos* para las inversiones obtenidos experimentalmente, y se correlacionaron con los genotipos de *SNPs* localizados en la región invertida y las regiones flanqueantes, para detectar los *SNPs* fijados, que tienen valores de r^2 de 1. Estos *SNPs* están en completo desequilibrio de ligamiento con la inversión y se definieron como *SNPs* marcador en la población Europea. En el caso de los *SNPs* marcador globales para todas las poblaciones, se analizó el desequilibrio de ligamiento en todas las poblaciones entre los *SNPs* marcador identificados en individuos Europeos. El objetivo era ver en qué casos el desequilibrio de ligamiento entre estos *SNPs* era específico de la población Europea o se mantenía en todas las poblaciones. Para cada inversión se obtuvieron los genotipos de los *SNPs* marcador en individuos Europeos para 1092 individuos no relacionados, correspondientes a las 14 poblaciones analizadas en el proyecto de los 1000 Genomas [Abecasis et al. 2012]. Se analizó el desequilibrio de ligamiento entre parejas de *SNPs*, usando el programa *Haploview*. Se seleccionaron como *SNPs* marcador para todas las poblaciones las parejas de *SNPs* con valores de r^2 iguales o superiores a 0.99. De cada pareja seleccionada se usó el *SNP* más próximo a uno de los puntos de rotura para estimar la frecuencia de la inversión. En ocasiones se encontró más de una pareja con el mismo valor y en esos casos sólo se consideraron si forman parte de un bloque de desequilibrio de ligamiento con valores de r^2 iguales o superiores a 0.8 para todas las parejas. En el caso de las inversiones para las que se disponía de individuos genotipados bioinformáticamente, se comprobó el desequilibrio de ligamiento usando los genotipos de los individuos y se definieron los *SNPs* marcador globales de manera manual y mediante *Breakseq* + *svgem*, usando el mismo valor de r^2 como umbral. Estos *SNPs* marcador globales son más fiables debido a que realmente se analizó el desequilibrio de ligamiento entre el alelo invertido y los alelos de los *SNPs*, mientras que el análisis por parejas de *SNPs* puede tomar como *SNPs* marcador los que están en desequilibrio de ligamiento con su pareja pero no con la inversión. En el caso de las inversiones no genotipadas bioinformáticamente, en que no se detectaron *SNPs* marcador globales válidos o bien se encuentran lejos de los puntos de rotura, se usaron los *SNPs* marcador en población Europea, concretamente los localizados en el interior de la región invertida y que están localizados más cerca de uno de los puntos de rotura, ya que es la localización donde es menos probable que ocurra la recombinación y por lo tanto, donde es más probable que los alelos de los *SNPs* y la inversión segreguen juntos.

Los haplotipos para los cromosomas invertidos y no invertidos, se estimaron mediante el programa *Phase 2.1.1* [Stephens et al. 2001]. Se usaron los genotipos de los *SNPs* de la región invertida y las regiones flanqueantes pero no los de las RIs. También se usaron los genotipos de las inversiones. Se generaron haplotipos usando la información de *SNPs* del proyecto de los 1000 Genomas para 35 individuos no relacionados y por otra parte la información del proyecto *HapMap* para 90 individuos agrupados en 30 familias padre-madre-hijo, que permite crear haplotipos con mayor fiabilidad. Se pudo analizar la variación haplotípica asociada a cada alelo de la inversión a partir de los haplotipos estimados por el programa. Finalmente se generaron redes de haplotipos usando un algoritmo *median-Joining* incluido en el programa *Network 4.612* [Bandelt et al. 1999], para analizar de manera visual la diversidad haplotípica del alelo estándar y el invertido. Se analizaron de manera manual los grupos de haplotipos de cada alelo de la inversión, por ejemplo un haplotipo perteneciente a un alelo de la inversión localizado en la red junto a los haplotipos del otro alelo puede indicar recurrencia en el origen de la inversión. En general, se comprobó que las redes representaban los grupos de haplotipos de cada orientación de manera separada y que el alelo más frecuente de la inversión tenía mayor diversidad haplotípica. Este patrón corresponde a un origen único de la inversión y a una acumulación de cambios entre ambas orientaciones de la secuencia. Cualquier alteración del patrón se analizó como posible indicación de recurrencia en el origen de la inversión, pero el origen único o recurrente de las inversiones se determinó mediante el análisis del estado ancestral, el análisis del desequilibrio de ligamiento entre *SNPs* e inversión y el análisis de los haplotipos.

2.8 Análisis de la frecuencia y distribución poblacional

En primer lugar se analizó si la población Europea cumple el equilibrio de Hardy-Weinberg para las inversiones genotipadas por *PCR*, mediante un *test* chi-cuadrado. En segundo lugar se determinó el estado ancestral tanto bioinformáticamente como experimentalmente. Se alineó la secuencia del genoma humano de Referencia (ensamblaje HG18) con la de chimpancé (ensamblajes panTro1,2,3 y 4), la de gorila (ensamblaje gorGor3) y la de mono *Rhesus* (ensamblaje macRhe3) mediante *Blast2seq* [Altschul et al. 1990]. Previamente se obtuvieron las secuencias de las especies de primates a partir de la secuencia en humanos, mediante la opción de conversión entre ensamblajes que ofrece el navegador genómico *USCS* [Kent et al. 2002]. Posteriormente se realizaron experimentos de *PCR* o *PCR* inversa en 4 chimpancés y dos gorilas (nombrados anteriormente en el apartado de muestras de ADN), para confirmar la orientación en estas especies.

La variación de las frecuencias alélicas en las diferentes poblaciones humanas se analizó mediante el estadístico *Fst*. El índice de fijación o *Fst* es una medida de la diferenciación poblacional debido a la estructura genética de la población. Se usó el programa *Arlequin*

3.0 [Excoffier et al. 2005] para calcular el índice de fijación por población y también por continente a partir de las frecuencias para las inversiones genotipadas en las 14 poblaciones humanas correspondientes al proyecto de los 1000 Genomas [Abecasis et al. 2012]. Además se calculó el *p-valor* para ambos valores.

3. RESULTADOS

3. RESULTADOS

3.1 Análisis, validación y estudio poblacional de las inversiones entre dos genomas humanos

El apartado de resultados de esta tesis corresponde por completo con un artículo que aún no se ha publicado pero que se muestra a continuación.

David Vicente-Salvador, Marta Puig, Magdalena Gayà-Vidal, David Izquierdo, Alexander Martínez-Fundichely, Aurora Ruiz-Herrera, Xavier Estivill, Cristina Aguado, José Ignacio Lucas-Lledó and Mario Cáceres Validation and population analysis of inversions between independently assembled human genomes.

Validation and population analysis of inversions between independently assembled human genomes

David Vicente-Salvador[1], Marta Puig[1], Magdalena Gayà-Vidal[1], David Izquierdo[1], Alexander Martínez-Fundichely[1], Aurora Ruiz-Herrera[1,2], Xavier Estivill[3,4], Cristina Aguado[1], José Ignacio Lucas-Lledó[1] and Mario Cáceres[1,5]*

[1] *Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.*
[2] *Departament de Biologia Celular, Fisiologia i Immunologia, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.*
[3] *Centre for Genomic Regulation (CRG), Barcelona, Spain.*
[4] *Universitat Pompeu Fabra (UPF), Barcelona, Spain.*
[5] *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.*

***Corresponding author:** Mario Cáceres
ICREA Research Professor
Institut de Biotecnologia i de Biomedicina
Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona)
Spain
Phone: +34 93 586 8726
Fax: +34 93 581 2011
E-mail: mcaceres@icrea.cat

ABSTRACT

The increasing number of structural variants catalogued in the human genome has often overlooked inversions as one of the most difficult types of variation to detect and study, even though they have been proven to affect different phenotypic traits in diverse organisms. In this work we have analyzed in detail 90 inversion predictions derived from the comparison of two independent human genome assemblies: the reference genome (NCBI36/HG18) and HuRef. We have found that only 29 of these predictions (32.2%) correspond to possibly true polymorphic inversions and we have validated experimentally 18 of them. The remaining predictions represent either errors in assembly comparison (31, 34.4%) or errors in one of the assemblies (30, 33.3%). This includes 25 errors in the reference genome between pairs of inverted repeats (IRs) that can now be corrected. All validated inversions have a small size (average 2958 bp), which suggests the limited ability of this strategy to detect longer inversions with more complex breakpoints. In addition, we have experimentally determined the ancestral alleles and frequencies in Europeans for 17 inversions (DAFs between x and y) and for 12 inversions also the frequencies in 14 worldwide-distributed populations based on either tag SNPs or the detection of breakpoint junctions in available next-generation reads from the 1000 Genomes Project. Among the validated polymorphic inversions, 9 (50%) have IRs at their breakpoints, and three show nucleotide variation patterns consistent with a recurrent origin that contrasts clearly with that of the inversions without IRs. Interestingly, seven of those (39%) show deletions at the breakpoint junctions in the derived allele which in many cases are mediated by microhomology sequences, highlighting the importance of mechanisms like FoSTeS/MMBIR in the generation of complex rearrangements in the human genome. Finally, we have found several inversions located within genes that may be affecting gene expression, and at least one inversion candidate to be positively selected in African populations.

INTRODUCTION

Although inversions have been known for a long time to exist as polymorphic variants in many species (Kirkpatrick 2010; Hoffmann and Rieseberg 2008), it is not yet clear the prevalence and importance of this type of variants in the human genome (Feuk et al. 2005; Feuk 2010; Alves et al. 2012; Martínez-Fundichely et al. 2014). An inversion implies the change of orientation of a segment of DNA within a given chromosome and by its own nature their detection is particularly challenging compared to that of other structural variants like CNVs (Redon et al. 2006; Conrad et al. 2010) or indels (Abecasis et al. 2012; Montgomery et al. 2013). First, inversions are balanced changes that often do not involve a change in the amount of DNA, and where the order of the different elements is the only alteration. Second, inversion generation is usually mediated by repeated sequences, which means that their breakpoints tend to be located within repeated (and often complex) elements of the genome that complicate their identification and characterization.

Even though the changes they cause in the DNA sequences may seem subtle, inversions are known to be able to affect phenotype and can be adaptive in many species (Lowry and Willis 2010; Thomas et al. 2008; Hoffmann and Rieseberg 2008; Joron et al. 2011; Jones et al. 2012). Their phenotypic effects may derive from the reduction in recombination between arrangements that takes place within the inverted sequence (Kirkpatrick 2010; Hoffmann and Rieseberg 2008) or from the mutational effects of their breakpoints (Puig et al. 2004; Imsland et al. 2012). In that sense, prevalence of inversions could be directly related to their mechanisms of origin as well as their effect on phenotype. One possibility is that they occur by recombination between oppositely oriented copies of a repeated sequence (Non-Allelic Homologous Recombination or NAHR), which causes the inversion of the intervening sequence (Lam et al. 2010; Kidd et al. 2010; Pang et al. 2013). Another option is that inversions originate through mechanisms that repair double or single-stranded DNA breaks. If no homology is detected at both inversion breakpoints, it is considered that the inversion was generated by Non-Homologous End Joining (NHEJ) or microhomology-mediated end joining (MMEJ) if microhomology sequences are involved (Onishi-Seebacher and Korbel 2011; Hastings et al. 2009). Finally, replication-based mechanisms such as FoSTeS/MMBIR (Slack et al. 2006; Lee et al. 2007) could also have a role in the generation of inversions and other rearrangements, which often show microhomology sequences at their breakpoints. However the relative importance of each of these mechanisms in inversion generation is not clear yet.

So far, only a handful of polymorphic inversions have been characterized in detail in humans (Feuk 2010; Martínez-Fundichely et al. 2014; Aguado et al. 2014), and in many less their distribution and frequency in human populations has been determined (Stefansson et al. 2005; Salm et al. 2012; Pang et al. 2013). Some of these inversions also seem to affect either gene expression (Salm et al. 2012) or certain phenotypes (González

et al. 2014) and even might be positively selected in some populations (Stefansson et al. 2005) suggesting that they may have roles in human phenotypic variation, disease or evolution. Nonetheless, the catalogue of polymorphic human inversions is far from complete yet.

One of the first challenges is the identification of these polymorphic inversions. High-throughput techniques like paired-end mapping (PEM) (sequencing of the two ends of libraries of DNA fragments of known size) have provided a wealth of information with respect to structural variants in the human genome, including inversions (Kidd et al. 2008; Ahn et al. 2009; McKernan et al. 2009; Korbel et al. 2007; Wang et al. 2008). However, the high-repeat content of the human genome as well as the inter-individual variability found in the regions where these variants tend to occur, affect the reliability of the mapping of sequence reads to the reference genome, making validation of the predicted structural variants an essential part of the detection process (Martínez-Fundichely et al., in preparation). Comparison of independently assembled genomes could be a good alternative to overcome these limitations and avoid biases of inversion detection from PEM (Lucas Lledó and Cáceres 2013). Nevertheless, *de novo* assembly is a computationally complex and costly procedure, especially with short reads from next-generation sequencing (NGS) technologies (Alkan et al. 2011).

Once a polymorphic inversion is identified and the exact localization of its breakpoints known, the second challenge is large-scale genotyping to investigate inversion frequency, geographic distribution, relationship to nucleotide variation or association to expression changes or phenotypic traits. While several techniques that attempt to detect breakpoint junctions like PCR (Pang et al. 2013), inverse PCR (iPCR) (Aguado et al. 2014), FISH (Antonacci et al. 2009) or alternative large-clone assemblies (Martin et al. 2004; Kidd et al. 2010) can be used to experimentally validate the existence of inversions with breakpoints of increasing length and complexity, not all of these methods can be applied easily to a large number of samples. Inversions with simple breakpoints could in theory be genotyped using NGS data with enough depth, but genotyping of inversions with inverted repeats (IRs) at the breakpoints is a more complex problem. Therefore, the use of SNP alleles in high linkage disequilibrium with the inversion alleles has been adopted as an indirect but easily scalable method to infer inversion genotypes in large numbers of individuals (Steinberg et al. 2012) using data from available SNP and haplotype collections like the 1000 Genomes Project (1000GP) (Abecasis et al. 2012) or HapMap (The International HapMap Consortium 2005). One danger of this is that, as recent evidences suggest, recurrence of inversions mediated by IRs might be more common than previously thought (Aguado et al. 2014; Antonacci et al. 2009), and for many inversions it is not possible to determine their genotypes and possible effects through the association with SNPs.

The HuRef genome (Levy et al. 2007), sequenced from a single individual (J. Craig Venter), represents a notable exception and a good opportunity for the identification of a whole set of structural variants in a single genome. This genome was sequenced by the Sanger method that provides relatively long reads with an average 7.5x coverage, and independently assembled into 4,528 scaffolds containing 2,810 Mb of contiguous sequence. Many variants were found in the comparison with the reference genome assembly (NCBI36/HG18), including 3,213,401 SNPs, ~53,800 block substitutions, ~851,000 indels, and 90 inversions. These inversions were not validated in the original work, but Pang et al. (2013) (in a study that was published during the course of the present work) analyzed seven of these predictions with more detail and found that two of them were potential HG18 assembly mistakes while five could be experimentally validated as real polymorphic inversions, highlighting the need of further analysis to clarify if the remaining predictions correspond to real polymorphic variants.

In this work we have carried out a detailed bioinformatic and experimental analysis of the 90 inversion predictions obtained from the HuRef-HG18 comparison. Using PCR and iPCR we have validated and genotyped the candidate inversions in individuals from different human populations. In addition, we have combined the genotype information with available nucleotide variation data to obtain insights about the evolutionary history and possible functional consequences of these inversions.

RESULTS

Sequence alignment and breakpoint definition

The alignment of the two genome assemblies (HuRef and HG18) for the 90 regions reported as inversions by Levy et al. (2007) plus 5 Kb of flanking sequence to each side, resulted in 59 regions showing an inverted segment (Table S1). The remaining 31 predictions did not represent true inversions since no unique inverted alignment was detected in the reported region and were classified as assembly comparison errors. Even though we found several types of errors (Figure 1), they all seem to be caused by the local mapping process used to call the variants between the two assemblies. Most of these errors are related to repeated sequences, polymorphisms with different alleles in both genomes or sequence gaps in the HuRef assembly.

In 14 cases (Table S2), the predicted inverted region maps completely within one copy of a pair of inverted segmental duplications (SDs) in the HG18 sequence. The detection of an inverted alignment is caused by the incorrect mapping of a small portion of one of HuRef SDs into the paralogous copy of the pair in HG18 in opposite orientation. This situation is usually accompanied by the absence (due to either an indel polymorphism or a sequencing gap) of all or part of one of the SD copies in the HuRef sequence (Figure 1A).

In some cases this mapping within the oppositely-oriented SD copy might be caused by a true inversion polymorphism of the sequence between both SDs, which would also include the internal part of the SDs. The presence in HuRef of the inverted allele and chimeric SDs may explain the observed crossed mapping pattern. Twelve of these cases are related to nine putative polymorphic inversions of larger size (Table S2), all of which have paired-end mapping (PEM) support (Martínez-Fundichely et al., in preparation) or have been experimentally validated (Aguado et al. 2014). The other two regions might represent misassembled SDs in one of the genome assemblies, or inversions for which no other evidence is yet available. We tested in HuRef DNA Hsinv0396, previously assayed by iPCR (Aguado et al. 2014), which is associated to HsInv0080, a 754-bp inversion prediction located within one of two 9.5-kb SD copies with 99.7% identity that mediate the inversion. However, iPCR revealed that HuRef carries the *Std* allele (data not shown), indicating that the prediction of HsInv0080 would be caused by an incorrectly assembled hybrid SD in HuRef genome or gene conversion between the SDs, as has been suggested in two other cases (Table S2, Martínez-Fundichely et al., in preparation).

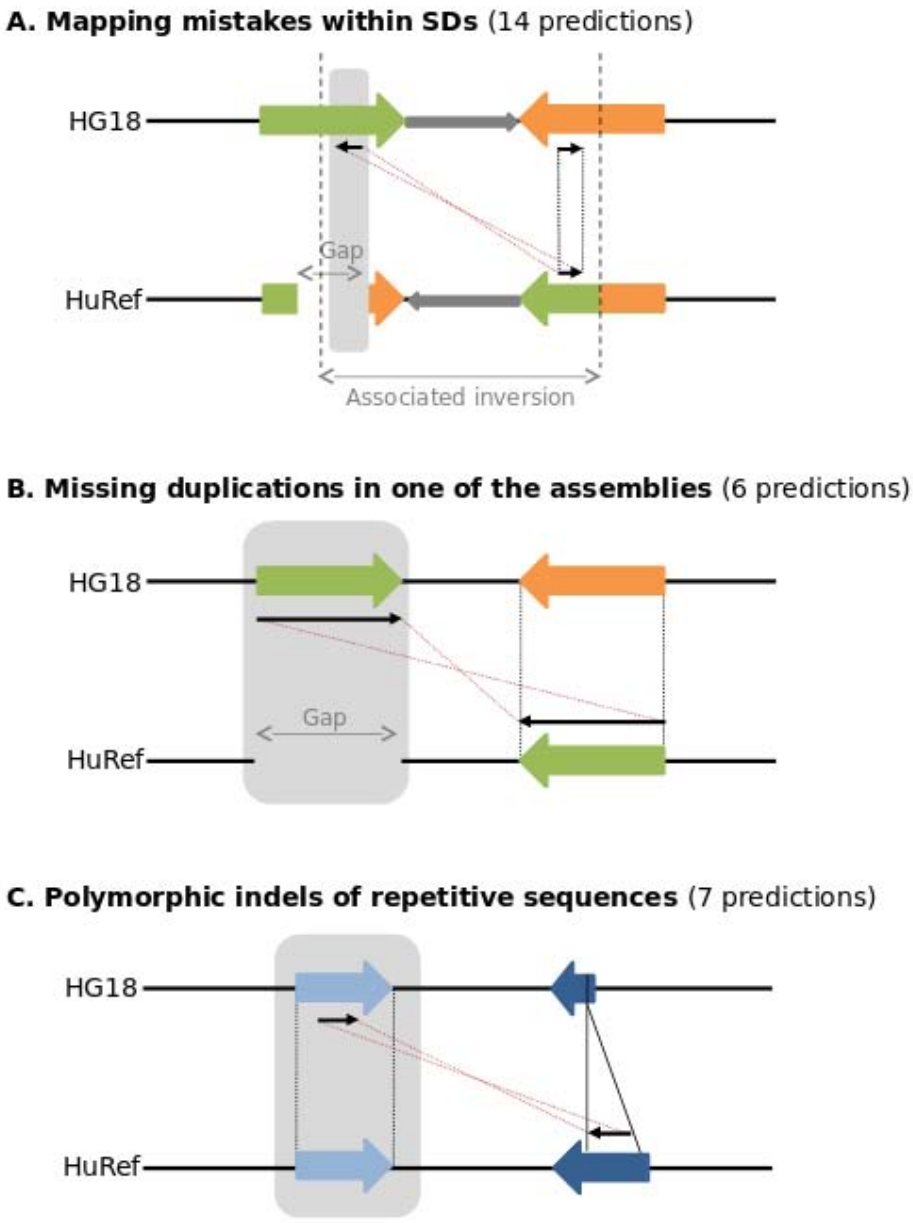

Figure 1. Errors in genome assembly comparison.
A. Inverted sequences detected in the HuRef-HG18 comparison due to mapping mistakes within inverted segmental duplications (SDs) can be caused by exchange between both SD copies (due either to gene conversion or a polymorphic inversion) or misassembly of the SDs in one of the genomes. **B.** Errors caused by the presence of missing inverted duplicated sequences in one assembly. **C.** Errors due to polymorphic repetitive element insertions or deletions. Inverted SD sequences are represented by green (SD1) and orange (SD2) arrows and inversion predictions in HuRef by thinner black arrows. Blue arrows indicate repetitive sequence insertions. Mappings between both genome assemblies that generated the inversion prediction are connected with red dashed lines and other possible direct mappings with black dashed lines. The shaded part corresponds to the alignment between orthologous sequences in both genomes corresponding to HuRef inversion prediction coordinates.

In six additional predictions, the inverted mapping is caused by the presence in HG18 of a duplicated inverted sequence for which only one copy was assembled in HuRef. In these regions a whole copy of a SD pair is missing in HuRef and the extant copy maps better with the sequence in inverted relative orientation in HG18 (Figure 1B), although a direct alignment could also be detected. In seven other regions, the inversion predictions fall entirely within transposable element (TE) copies or simple repeat sequences, and the inverted mapping is probably caused by the absence of the true hit in HG18. This causes that a different copy of the repeat, which happens to be in inverted orientation, is detected as the best possible mapping (Figure 1C). For the last four cases, the causes of the error could not be determined. In addition to not detecting unique inverted alignments between HG18 and HuRef sequences, no further evidence of inversion polymorphisms existed in any of these 31 regions according to available PEM data (Kidd et al. 2008; Ahn et al. 2009; McKernan et al. 2009; Wang et al. 2008). Therefore, all these errors in the comparison between genome assemblies were excluded from any further experimental validation.

Once eliminated the assembly comparison errors, we defined inversion breakpoints (BPs) as precisely as possible from our local sequence alignments for the 59 regions candidate to contain polymorphic inversions using all available additional sequences, such as BACs or fosmids (Kidd et al. 2010). For simplicity, we will always refer to the standard allele (*Std*) as the orientation found in the reference genome (HG18) independently of which allele represents the ancestral arrangement, and the inverted allele (*Inv*) as that in HuRef. Breakpoints were defined in all cases as a range of positions to include easily those located within inverted repeats (IRs) or other regions of uncertainty, like microhomology sequences. When IRs were found at inversion breakpoints, we tried to identify as precisely as possible the point of exchange between the two copies in inverted chromosomes (Figure 2A). The comparison of the breakpoint coordinates reported by Levy et al. (2007) with those derived from our analysis, revealed that 18 putative inversions (30.5%) showed the same breakpoints (<3 bp difference) and 41 inversions (69.5%) had at least one breakpoint incorrectly defined (>3 bp difference) (Table S3).

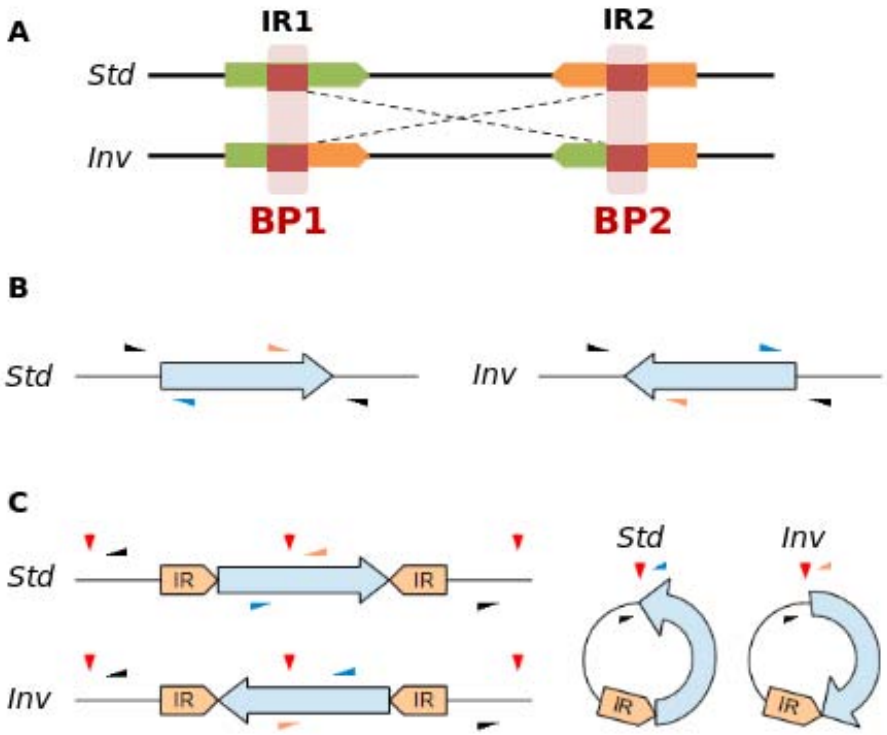

Figure 2. Breakpoint analysis and genotyping strategies.
A. Alignment between the two inverted repeats (IRs) at both breakpoint (BP) junctions in standard (*Std*) and inverted (*Inv*) chromosomes used to define the position where the exchange that caused the inversion took place (in red). IRs in BP1 and BP2 in the Std chromosome are shown in green and orange, respectively. IRs in the *Inv* chromosome are colored depending on their homology to each of the Std copies. The exchange between IRs originating an inversion can generate the observed hybrid pattern. **B.** Detection of *Std* and *Inv* alleles in inversions with simple breakpoints was performed by PCR across the breakpoints using four primers (small arrows): two primers (shown in color) located inside the inverted sequence (large blue arrow) each close to one breakpoint, and another two primers (black arrows) outside. Combinations of these four primers allow the amplification of the two inversion breakpoints in both chromosomal arrangements (*Std* is shown on the left and *Inv* on the right). **C.** Inversions with inverted repeats at the breakpoints were analyzed by inverse PCR (iPCR). Genomic DNA was digested with a staggered-end restriction enzyme able to cut within the inverted region and outside (red arrowheads) but not within the IRs (represented as orange boxes), and was then circularized to be amplified across the restriction-ligation point. An example of genotyping of BP1 with primers A, B and C on digested and circularized DNA is shown at the bottom right corner. Both breakpoints were tested in most cases to confirm that the detected structural change really corresponds to an inversion.

Experimental validation of inversions

In order to select the inversions to be experimentally validated, the 59 candidate regions were classified into three categories (polymorphic inversion, HuRef assembly error candidates, or HG18 assembly error candidates) according to their support in fosmid

PEM data from nine individuals (Kidd et al. 2008) (Table S1). The detection of both concordant and discordant fosmids in 13 regions allowed their classification as polymorphic inversion candidates. Five regions were considered HuRef assembly error candidates since only concordant fosmids were detected in all the individuals analyzed by PEM, leaving HuRef as the only genome with the inversion. Finally, 27 regions where all available fosmids map discordantly in the analyzed individuals were regarded as possible HG18 errors, since apparently the reference genome is the only one carrying the *Std* allele. When there were less than four fosmids mapping in one of the two possible orientations or the available fosmids have paired-end reads that do not map uniquely in the genome, the region was left unclassified (14 regions).

We tested experimentally by PCR or inverse PCR (iPCR), depending on the inversion features, (Figure 2) a total of 46 putative inversions (Table S1) including all those polymorphic candidates and unclassified inversions with sizes >1 kb or that could have effects on genes (16 inversions), as well as all those previously classified as errors in either genome assembly (30 inversions, excluding two already corrected in more recent versions of the human reference genome). Of those, 28 inversions were tested by direct PCR across the breakpoints and 18 using an inverse PCR approach (Table S1). The remaining 11 inversion candidates were not analyzed experimentally due to their small size (inverted region <1 kb), or to the presence of large SDs (>50 kb) at their breakpoints or lack of restriction enzyme targets for iPCR, which hinder inversion validation using available techniques (Table S1).

Validation experiments for the 16 regions candidate to be polymorphic were performed with ten samples corresponding to the nine individuals from different populations (European, African and Asian) previously analyzed by fosmid PEM in Kidd et al. (2008), together with HuRef DNA (Figure 3A). An inversion was considered to be validated and represent a true polymorphic inversion in human populations if a single allele with the *Inv* orientation was detected in this 10-individual panel (Figure 3A). To confirm that they are real inversions (instead for example inverted duplications in tandem), both breakpoints were analyzed for all of them except HsInv0052, which contains a large 19.6-kb indel at BP1 (Figure 4A, see below). The 16 regions (100%) could be validated as polymorphic inversions since both arrangements were detected and concordance of inversion genotypes by both breakpoints was perfect for all individuals (Table S4).

The five regions classified by PEM data as putative errors in HuRef assembly were tested by PCR directly in this DNA (Table S4). Both breakpoints were analyzed by direct PCR for all inversions except HsInv0005, in which only BP1 was tested. In all five regions, only *Std* AB and CD products were obtained and at least one of the PCR products was completely sequenced in each case to demonstrate that the sequence in HuRef has the same orientation that the HG18 reference (Figure 3A). In addition, none of these regions showed evidence of an inversion taking into account available PEM data (Kidd et al.

2008). Therefore, the prediction of inverted regions in these regions is caused by errors in HuRef genome assembly and they do not correspond to real polymorphic inversions.

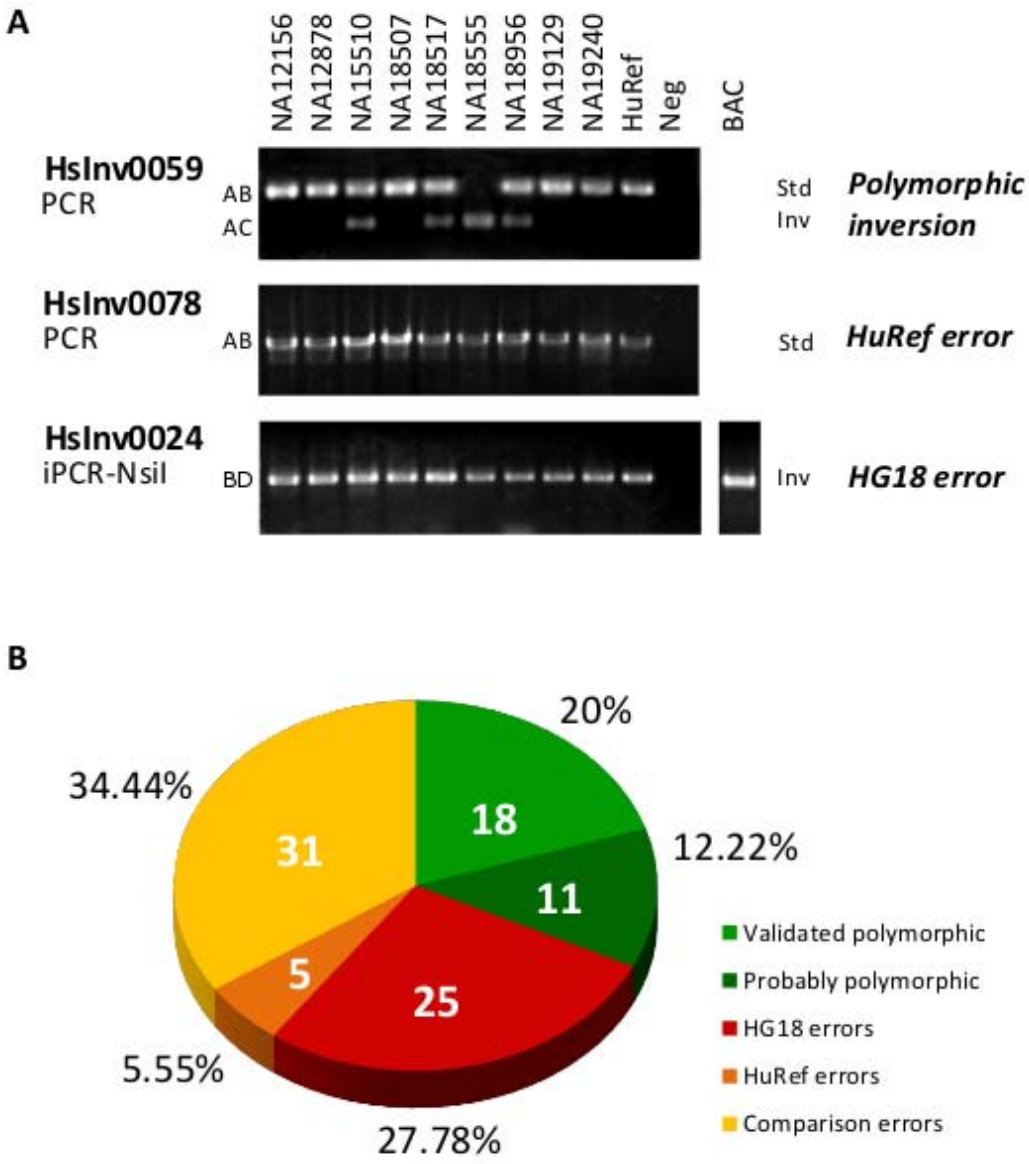

Figure 3. Validation of polymorphic inversions.

A. Inversion genotyping assays were used on a 10-individual panel formed by nine individuals with available fosmid paired-end mapping (PEM) data (Kidd et al. 2008) together with HuRef DNA. Inversions were considered as true polymorphic changes if a single allele carrying the alternative allele (*Inv*) with respect to reference genome HG18 (*Std*) was detected (for example, HsInv0059, top). For those inversions predicted by PEM data to be errors in HuRef, in all the analyzed individuals and HuRef DNA the only amplified allele was the same as in HG18 (*Std*) (HsInv0078, middle). In those inversions predicted to be errors in HG18, all tested individuals as well as the corresponding clones (usually BACs) that are the source of the reference genome sequence show the *Inv* allele by PCR or inverse PCR, confirming that these regions need to be

corrected in future releases of the reference genome (HsInv0024, bottom). **B.** Pie chart showing the final results for the analysis of the 90 inversion predictions in Levy et al. (2007). Only 29 predictions (in green) represent true polymorphic inversions. Light green corresponds to inversions validated experimentally in this study and dark green those that are likely polymorphic but that have not been confirmed yet (see main text). The different types of errors are shown in different colors as explained in the enclosed legend. The number of inversions in each group is indicated within the chart and the percentage over the 90 initial predictions is also shown.

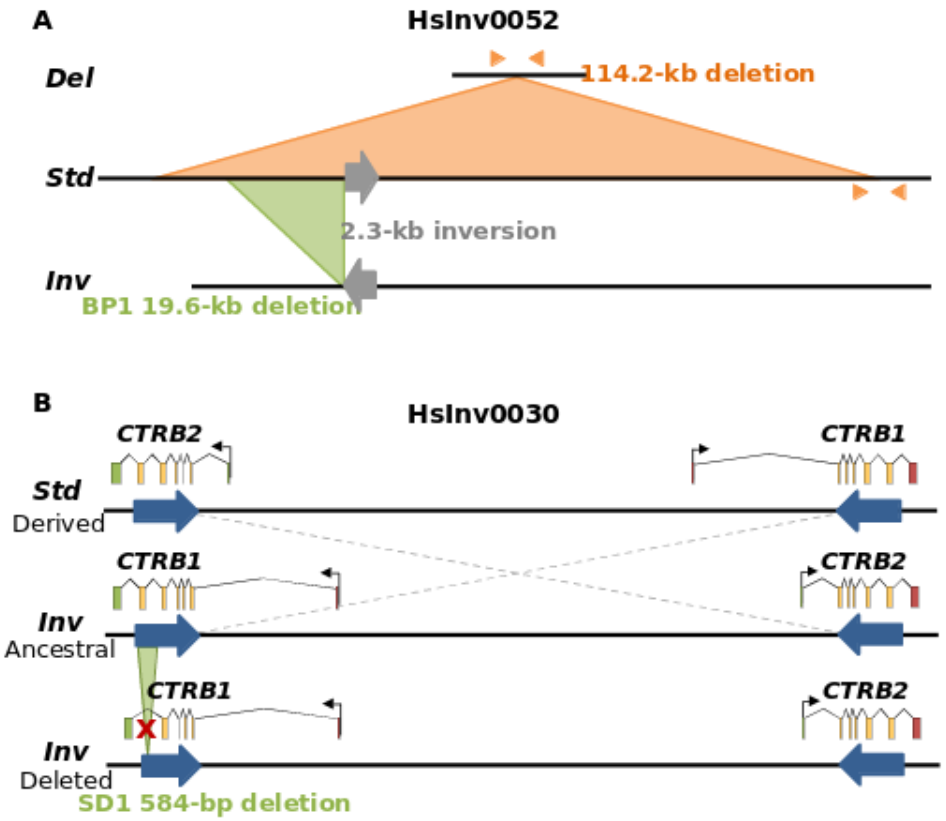

Figure 4. Examples of validated inversions.
A. HsInv0052 is a 2.3-kb inversion (grey arrow) identified within a 114.2-kb sequence that is a polymorphic deletion in human populations (shown in orange). This region, therefore, has three possible alleles (*Std*, *Inv*, and deleted or *Del*) with frequencies 0.29, 0.31 and 0.4, respectively in the CEU population. Orange arrowheads indicate primers used to genotype the large deletion. HsInv0052 BP1 is associated to a 19.6-kb deletion (in green) present in all *Inv* chromosomes. **B.** HsInv0030 is a 16.5-kb inversion mediated by SDs (blue arrows) that exchanges the first exons (green and red boxes) of two genes of the same family: *CTRB1* and *CTRB2*. Genes are reconstructed in the inverted alignment since the remaining exons (yellow boxes) located within the inverted SDs at the breakpoints encode exactly the same protein sequences with only the last exons (also shown in green and red) showing some different amino acids. In this case, the *Inv* allele represents the ancestral arrangement. Independently of the inversion, there is a third allele in this region carrying a deletion of one exon of the *CTRB* gene in BP1 (shown in green) that disrupts the corresponding protein.

Finally, the HG18 error candidate regions were tested by examining directly DNA of the same BACs from which the HG18 reference genome sequence derives as well as some genomic DNAs (Figure 3A and Table S5). Within the 27 HG18 error candidates, three regions (HsInv0001, HsInv0042 and HsInv0050) had been already corrected in the HG19 assembly and only one of these (HsInv0050) was tested experimentally because its correction was based on the substitution of one BAC for a different one, which does not necessarily imply that the inversion does not exist. For another two regions (HsInv0074 and HsInv0086) the sequenced BAC was not available and both the panel of 10 individuals of diverse origins and 90 European individuals (30 parent-child trios) were genotyped instead to try to detect at least one *Std* allele. Out of the 25 inversions tested using one of these approaches, two (HsInv0040 and HsInv0061) turned out to be real polymorphic inversions since both arrangements were detected in the 10-individual panel (Table S4). For the remaining inversions, both genomic DNA and BAC clone analysis showed the presence of the *Inv* allele exclusively, and the orientation of these sequences (22 in total, since HsInv0050 is already corrected in HG19) is in the process of being reversed in subsequent versions of the human genome reference sequence in collaboration with the Genome Reference Consortium (Church et al. 2011). In fact, in the recent release of the new human genome assembly (GRCh38/HG38), five of these regions detected here as HG18 errors already show the proper orientation. In the two cases where the primary clone was not available, all tested individuals were homozygotes for the inverted allele, which is consistent with these inversions being errors in HG18 genome assembly or with polymorphic inversions with a very low frequency in human populations. In fact, the sequences of these two inversion predictions have already been reversed in the HG38 assembly by using additional BAC sequences with the *Inv* orientation.

Origin of polymorphic inversions

In total, only 18 out of the 46 tested inversions (39.1%) were validated as true polymorphic inversions (Tables 1 and 2). They include inversions with sizes between 83 bp (HsInv0006) and 16.5 kb (HsInv0030), being most of them smaller than 10 kb, and with a random distribution across chromosomes. To determine the ancestral orientation of these regions, we genotyped 17 of the inversions in DNA from four chimpanzees and two gorillas (Table S6). Reliable results were obtained for the 17 inversions in at least one non-human species (15 in chimpanzee and 17 in gorilla), with 15 inversions tested by both BPs and two (HsInv0052 and HsInv0055) by only one. The HsInv0045 region is completely deleted in the chimpanzee genome and accordingly, no PCR products were obtained in this species, so the ancestral allele is based on the gorilla genome alone. Thus, ancestral state could be experimentally determined for the 17 tested inversions (Table S6). No inversions polymorphic in other non-human primate species were detected, but HsInv0055 shows a different orientation in chimpanzees and gorillas, as has already been found for other inversions (Aguado et al. 2014). The most recent genome assemblies for

these species together with that of the rhesus macaque were also analyzed (Table S6) and in most cases are consistent with experimental results. The exceptions are HsInv0030, HsInv0069 and HsInv0072, which show a *Std* orientation in at least one of the non-human primate genomes while we have amplified only inverted products from our DNA samples. The three cases correspond to inversions with IRs at the breakpoints, which as we have seen tend to cause assembly errors (all cases in which the orientation could not be determined in the sequence analysis correspond to this type of inversions, but for a single case) (Table S6). However, they could also represent inversions polymorphic in non-human primates that we have not detected due to the low number of chimpanzee and gorilla individuals analyzed.

Next, in order to investigate the mechanisms of generation, the breakpoint sequences for these 18 polymorphic inversions were examined. For nine (50%) of these inversions IRs were found at both breakpoints. The IRs range between 207 bp that are part of *Alu* elements in HsInv0031, to 7-kb SDs in HsInv0069 (Table 3). The two copies of most of these repeats have also a high identity (>96%), with only HsInv0031 and HsInv0045 *Alu* repeats showing an identity around ~85%. Therefore all these inversions were probably generated by non-allelic homologous recombination (NAHR) between the two IRs. The remaining nine inversions (50%) are not mediated by IRs, but eight of them (88.9%) have indels associated to one or both breakpoints (Table 4). These indels have sizes between 2 bp and 19.2 kb and the presence or absence of these sequences is completely linked to the two different arrangements. Moreover, with the exception of just a few duplicated nucleotides, in all these cases it is the ancestral orientation the one that has sequences at the breakpoints that have been deleted in the derived rearranged chromosome. The detection of linked structural changes at the breakpoints (deletions in these cases) that were most likely generated in the same event as the inversion itself, suggests that replication-based mechanisms like FoSTeS/MMBIR (Slack et al. 2006; Lee et al. 2007) could have originated these inversions with complex breakpoint structures. The presence of microhomology sequences at several of the inversion breakpoints reinforces this possibility. These sequences might also be involved in the generation of these complex events in successive steps by microhomology-mediated end joining (MMEJ) (McVey and Lee 2008), although it seems less likely. The other two inversions appear to be generated by non-homologous end joining (NHEJ) mechanisms, which in one case (HsInv0006) would have involved staggered breaks and duplication of the flanking sequences in the process.

Frequencies of polymorphic inversions in human populations

To study the inversion frequencies, we used PCR and iPCR assays for a single inversion breakpoint to genotype 17 of them in 90 HapMap individuals of European origin (CEU) comprising 30 father-mother-child trios (Table S7). HsInv0036 could not be genotyped in this larger panel since the 200 ng of DNA required for the long-range PCR genotyping assay were not available for all HapMap CEU DNAs. All genotyped inversions follow a perfect Mendelian transmission from parents to children and are in Hardy-Weinberg equilibrium in this population. The frequencies for the inverted allele in the 60 independent individuals analyzed range between 0.2 and 0.989 and are given in Table 1.

For seven inversions that have relatively clean breakpoints without repeated sequences, an alternative method was also used to genotype them in a higher number of human populations: the detection of reads spanning inversion breakpoints in the 1000 Genomes Project (1000GP) data (Abecasis et al. 2012) using the BreakSeq pipeline (Lam et al. 2010). A library including the two Std breakpoint junctions (AB and CD) as well as the two Inv breakpoints (AC and BD) for seven inversions was generated (Table S8) and compared to the mapped and unmapped reads from the 1,092 genomes from 14 different human populations of the 1000GP phase I (Abecasis et al. 2012). Reads comprising breakpoint junctions were detected for 344-589 individuals depending on the inversion, with an average number of 3.5 reads per inversion and individual. The algorithm svgem (Lucas-Lledó et al. 2014) was then used to estimate the most likely inversion frequencies for each population from the BreakSeq data. The inversion frequencies in each continental group for these seven inversions with simple breakpoints determined using BreakSeq and svgem are shown in Table 2 (see Table S9 for inversion frequencies in the 14 populations separately). Although the low coverage for most individuals of the 1000GP does not allow an unequivocal genotyping of each particular individual, in the CEU population, in silico genotypes are in complete agreement with PCR results for the seven analyzed inversions, with inverted reads detected only among known inversion carriers.

Table 1. Summary information of validated polymorphic inversions.

Basic information about the genomic location, inversion size, frequency, possible generation mechanisms, number of tag SNPs for each inversion allele and unique or recurrent origin is given for each polymorphic inversion characterized in this work. In bold are shown the frequencies of derived alleles with higher frequency than the ancestral allele. ND = not determined. See main text for the full name of generation mechanisms.

Inversion	Genomic location	Inversion size (bp)	Inv freq (CEU)	Std freq (CEU)	Generation mechanism	Ancestral arrangement	Tag SNPs	Origin
Inversions with IRs								
HsInv0030	chr16:73797599-73814159	16,560	0.84	0.16	NAHR	*Inv*	0	Recurrent
HsInv0031	chr16:83746237-83747302	1,036	0.68	0.32	NAHR	*Inv*	10	Unique
HsInv0036[1]	chr18:12134389-12137214	1,441	ND	ND	NAHR	*Std*	ND	ND
HsInv0040	chr2:138721419-138725673	3,589	0.79	0.21	NAHR	*Inv*	18	Unique
HsInv0045	chr21:26942554-26943508	940	0.45	0.55	NAHR	*Std*	1	Unique
HsInv0055	chr5:63800180-63811001	6,252	0.79	0.21	NAHR	*Inv*	0	Recurrent
HsInv0061	chr6:107275899-107277573	1,674	0.98	0.02	NAHR	*Inv*	0	Unique
HsInv0069	chr9:114913782-114914991	1,200	0.63	0.38	NAHR	-	0	Recurrent
HsInv0072	chrX:45433293-45435559	2,486	0.99	0.01	NAHR	*Inv*	0	Unique
Inversions without IRs								
HsInv0003	chr1:185733101-185733350	251	**0.75**	0.25	MMBIR / FoSTeS	*Std*	24	Unique
HsInv0004	chr1:196023846-196024607	1,192	0.20	**0.80**	NHEJ / MMEJ	*Inv*	52	Unique
HsInv0006	chr1:203445294-203445397	83	0.59	0.41	NHEJ	*Inv*	10	Unique
HsInv0041	chr2:225001224-225001326	136	0.45	0.55	MMBIR / FoSTeS	*Std*	12	Unique
HsInv0052[2]	chr3:164028056-164030335	2,279	0.31	0.29	MMBIR / FoSTeS	*Std*	40	Unique
HsInv0058	chr6:31117201-31118074	873	0.61	0.39	MMBIR / FoSTeS	*Inv*	19	Unique
HsInv0059	chr6:89980353-89980661	308	0.82	0.18	MMBIR / FoSTeS	*Inv*	1	Unique
HsInv0063	chr7:70064121-70076815	12,693	**0.69**	0.31	MMBIR / FoSTeS	*Std*	19	Unique
HsInv0068	chr9:76087960-76088209	249	0.78	0.23	MMBIR / FoSTeS	*Inv*	11	Unique

[1] HsInv0036 has not been genotyped in the CEU individuals nor tested experimentally in non-human primates. Ancestral allele has been determined only based on the gorilla genome sequence analysis.

[2] Std and Inv allele frequencies do not add 1 because there is a third allele that removes completely the inverted sequence segregating with a frequency of 0.4 in the CEU population.

Table 2. Polymorphic inversion frequencies in different continents.
The frequency of the Inv allele (alternative allele compared to the reference genome) is shown for each continental group in the 1000GP data calculated by at least one of two methods (tag SNPs and/or BreakSeq/svgem analysis, see main text). For global tag SNPs the SNP used to establish frequencies is indicated and the total number of individuals analyzed is shown in the last line of the table. For the BreakSeq/svgem method, the number of individuals analyzed varies for each inversion and population and is shown in parentheses. Fst values are calculated comparing continental group frequencies. EUR = European, AFR = African, ASN = Asian, AMR = American. See Table S8 for frequencies in each individual population.

Inversion	Method	SNP	ALL	EUR	AFR	ASN	AMR	F_{st}
HsInv0003	Global tag SNPs	rs2383648	0.83	0.75	0.76	0.99	0.84	0.09
	BreakSeq/svgem	-	0.84 (560)	0.76 (225)	0.79 (95)	0.99 (165)	0.81 (75)	0.09
HsInv0041	Global tag SNPs	rs6733222	0.46	0.38	0.69	0.40	0.42	0.08
	BreakSeq/svgem	-	0.48 (575)	0.42 (218)	0.66 (112)	0.45 (172)	0.49 (73)	0.04
HsInv0059	Global tag SNPs	rs282113	0.71	0.83	0.90	0.38	0.71	0.25
	BreakSeq/svgem	-	0.69 (589)	0.81 (218)	0.85 (112)	0.41 (172)	0.72 (87)	0.20
HsInv0063	Global tag SNPs	rs6460609	0.56	0.63	0.30	0.70	0.57	0.12
	BreakSeq/svgem	-	0.65 (532)	0.71 (210)	0.31 (99)	0.74 (165)	0.72 (58)	0.15
HsInv0006	BreakSeq/svgem	-	0.51 (545)	0.67 (200)	0.10 (120)	0.57 (160)	0.61 (65)	0.25
HsInv0058	BreakSeq/svgem	-	0.60 (585)	0.63 (225)	0.63 (110)	0.52 (178)	0.70 (74)	0.01
HsInv0068	BreakSeq/svgem	-	0.71 (344)	0.62 (189)	0.77 (47)	1.00 (89)	0.08 (19)	0.32
HsInv0004	Global tag SNPs	rs1775463	0.14	0.20	0.03	0.13	0.20	0.05
HsInv0031	Global tag SNPs	rs9933231	0.64	0.69	0.60	0.58	0.68	0.02
HsInv0040	Global tag SNPs	rs4350808	0.79	0.74	0.76	0.89	0.77	0.03
HsInv0045	Global tag SNPs	rs7283610	0.52	0.49	0.53	0.57	0.49	0.00
HsInv0052	Deletion tag SNP	rs206276	0.53	0.29	0.59	0.89	0.37	0.30
	Inversion tag SNP	rs13073727	0.22	0.41	0.07	0.03	0.36	0.22
N (1000GP)			1092	379	246	286	181	

Nucleotide variation analysis

It is known that inversions cause a reduction in recombination within the inverted segment, which generates linkage disequilibrium (LD) with variants located within and around the inversion. For this reason, nucleotide variation was analyzed for the 17 inversions that have been experimentally genotyped in the CEU population using the 35 and 60 unrelated individuals with 1000GP and HapMap SNP data, respectively. These two approaches are complementary because the 1000GP provides a higher number of genotyped SNPs but less individuals with inversion genotypes are available, while HapMap allows the analysis of more individuals even though it will be based on fewer SNPs. The analysis performed including the inverted segment as well as 10 kb of flanking sequence up and downstream identified two clearly different patterns and yielded very similar results with both SNP datasets, although the number of SNPs is always lower in HapMap (Table S10). Fixed SNPs in perfect LD with the inversion ($r^2 = 1$) were identified for 12 inversions using the 1000GP SNP data (in eight of these the tag SNPs were also present in the HapMap data analysis). For these inversions, some shared SNPs (both SNP alleles are found in the two arrangements) were also detected around the inverted

segment, but in eight of them all SNPs inside the inverted region carry a different allele in each arrangement. On the other hand, we found three inversions (HsInv0030, HsInv0055 and HsInv0069) showing only shared SNPs within and around the inverted region and no fixed SNPs. Finally, in two inversions (HsInv0061 and HsInv0072) no shared or fixed SNPs were found within the analyzed region. The existence of shared SNPs between *Std* and *Inv* chromosomes suggests either the presence of gene conversion tracks or a recurrent origin for the inversion, which could have been generated more than once on chromosomes carrying different haplotypes. Fixed SNPs that distinguish both arrangements can be expected if the inversion origin is unique.

To check this, we also inferred the SNP phase and determined the complete haplotypes for the 1000GP and HapMap individuals, in which case, since we used the trio information, haplotypes were more reliable. We constructed haplotype networks to explore the relationships between the different haplotypes for the two SNP data sets using only the sequences within the inverted segment. As expected, for 10 of the 14 inversions with fixed SNPs or no SNP information in the unphased data we found that the haplotypes for the two arrangements were clearly differentiated (Figure 5A). In the other four inversions only one haplotype was found in the derived arrangement, but this haplotype was shared with the ancestral orientation (Figure 5B). However, when we considered the whole analyzed region including sequences flanking the inversion, the derived haplotypes are clearly separated from the ancestral ones, highlighting their unique origin (data not shown). For the other three inversions with a high proportion of shared SNPs, we found several quite differentiated haplotypes shared between *Std* and *Inv* chromosomes (Figure 5C). All individuals carrying unusual combinations of haplotype and inversion allele that suggest recurrency were genotyped experimentally a second time to confirm inversion status. Therefore, our analysis indicates that 14 out of 17 (82.35%) inversions show patterns compatible with a unique origin in the CEU population while 3 (17.65%) might represent recurrent events (Tables 1 and S10). Interestingly, these last three inversions have IRs at their breakpoints (Table 3) and the repeated occurrence of the inversions in these particular regions of the genome can be explained by NAHR (Aguado et al. 2014). These results are consistent with the existence of the two HsInv0055 arrangements in chimpanzees and gorillas, which provides additional evidence for recurrence in this region.

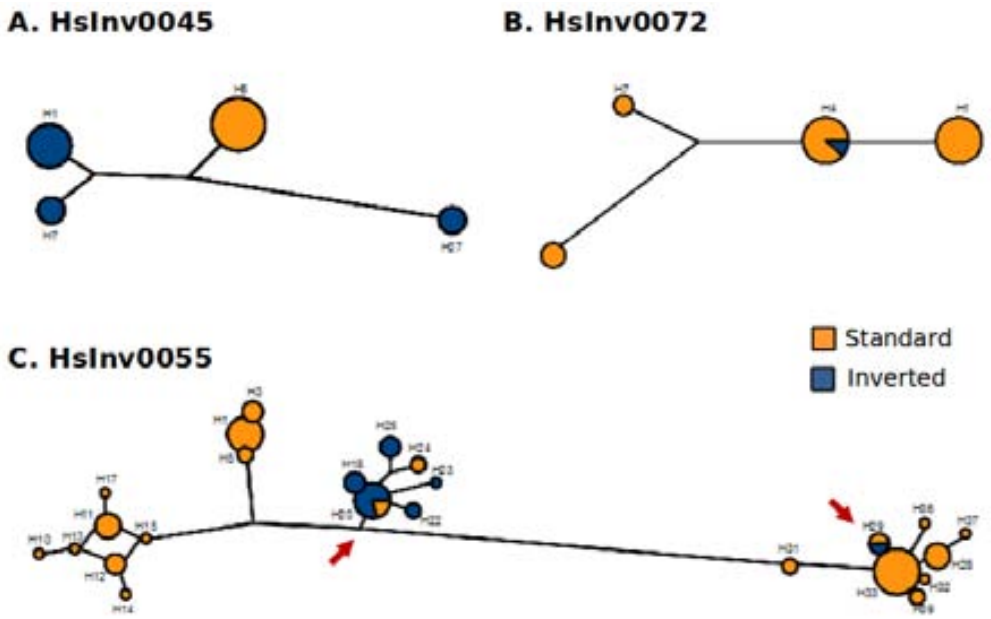

Figure 5. Haplotype networks of inversions with unique and recurrent origin.
Each circle represents a haplotype and circle size is proportional to the number of chromosomes carrying that particular haplotype. Circles connected with lines are related, and the distance between them is proportional to the number of changes that differentiate the two haplotypes. Orange and blue denote *Std* and *Inv* haplotypes, respectively. **A.** Inversion with a unique origin, where all derived chromosomes carry a single haplotype. **B.** Inversion where one haplotype is shared by *Std* and *Inv* chromosomes, a pattern compatible with a unique origin and a derived chromosome (the *Std* allele in case of HsInv0072 represented here) that has not yet accumulated changes that differentiate it from the corresponding ancestral haplotype (*Inv*). **C.** Inversion that shows more than one shared haplotype, which suggests a recurrent origin, that is, that several events of inversion (and sometimes reversion) took place on chromosomes carrying different haplotypes. In case of HsInv0055, the derived *Std* allele was generated at least twice from two different haplotypes carrying the ancestral *Inv* allele (indicated by red arrows). All haplotype networks shown here are based on the 1000GP SNP data within the inverted region exclusively.

Table 3. Features of polymorphic inversions with inverted repeats (IRs) at inversion breakpoints.

Genomic position of defined inversion breakpoints (they do not necessarily correspond to the IR coordinates if the exchange positions could be narrowed down) is indicated together with other data about these repeats. In the case of inverse PCR (iPCR) assays, the restriction enzyme used is also indicated. SD = segmental duplication (defined as duplications ≥1 kb and with ≥90% identity), and TE = transposable element.

Inversion	Chr	BP1	BP2	Inversion size (bp)	IR size (bp) BP1/BP2	IR Identity (%)	IR type	Genotyping assay
HsInv0030[1]	16	73797041-73797599	73814160-73814718	16560	1548/1546	99.60	SD	PCR
HsInv0031	16	83746215-83746259	83747296-83747305	1036	207/207	84.60	TE (AluSq2-AluSx1)	iPCR-EcoRI
HsInv0036[2]	18	12131460-12135081	12136523-12140144	1441	3647/3647	99.90	Several TEs	Long-range PCR
HsInv0040	2	138720715-138721469	138725059-138725814	3589	755/756	99.90	IR	iPCR-HindIII
HsInv0045	21	26942303-26942558	26943499-26943755	940	256/257	85.60	TE (AluSx1)	iPCR-SacI
HsInv0055[1]	5	63797584-63802465	63808718-63813599	6252	5999/6002	96.70	TE (L1PA7-L1PA3)	iPCR-BamHI
HsInv0061	6	107275246-107275899	107277574-107278229	1674	654/656	99.70	IR	iPCR-HindIII
HsInv0069[1]	9	114907364-114913787	114914988-114921411	1200	7035/7027	98.60	SD	iPCR-NsiI
HsInv0072	X	45432027-45433183	45435670-45436826	2486	1448/1444	98.40	TE (L1PA13)	iPCR-HindIII

[1] Recurrent inversions.

[2] HsInv0036 could not be genotyped in the CEU individuals because the long-range PCR requires 200 ng of DNA in the amplification reaction and this amount was not available for all DNAs.

Table 4. Features of polymorphic inversions without inverted repeats at the breakpoints.
Genomic position of defined inversion breakpoints together with the associated indels. In bold, indels where HG18 carries the deleted derived allele and extra sequence is found in HuRef.

					Breakpoint 1		Breakpoint 2		
Inversion	**Chr**	**BP1**	**BP2**	**Inversion size (bp)**	**Indel position**	**Indel size (bp)**	**Indel position**	**Indel size (bp)**	**Deleted chr.**
HsInv0003	1	185733099-185733101	185733353-185733355	251	185731452-185733101	1650	185733351-185733352	2	*Inv*
HsInv0004	1	196023411-196023414	196024607-196024608	1192	-	-	-	-	-
HsInv0006[1]	1	203445230-203445294	203445378-203445457	83	203445254-203445294	41	203445378-203445416	39	-
HsInv0041	2	225001193-225001195	225001332-225001334	136	225001196-225001224	29	225001335-225002195	861	*Inv*
HsInv0052[2]	3	164028056-164028057	164030337-164030338	2279	164008437-164028056	19620	164030338-164030349	12	*Inv*
HsInv0058	6	31117199-31117201	31118075-31118075	873	31117200-31117201	2187	31118075-31118076	630	*Std*
HsInv0059	6	89980347-89980353	89980662-89980668	308	-	-	89980771-89980772	618	*Std*
HsInv0063	7	70064121-70064122	70076816-70076823	12693	70058906 -70064121	5216	-	-	*Inv*
HsInv0068	9	76087959-76087960	76088210-76088213	249	76087958-76087961	1231	76088209-76088210	325	*Std*

[1] Indels at HsInv0006 breakpoint junctions correspond to insertions created by the repair of staggered single-strand breaks.
[2] For HsInv0052 some individuals carry a polymorphic 114.2-kb deletion that removes the whole region including the inverted segment.
This deletion has been genotyped with primers flanking the deletion point that produce specific PCR products confirming the absence or presence of this sequence.
Inversion status can only be genotyped in those individuals that do not have this sequence deleted.

Evolutionary history and worldwide distribution of inversions

For unique inversions, fixed SNPs ($r^2 = 1$) (listed in Table S11) can be used as tag SNPs to identify both arrangements in new individuals not genotyped experimentally. However, we found that not all tag SNPs in the CEU population for a given inversion show the exact same frequency in other populations of the 1000GP, suggesting that not all these SNPs remain completely linked to the inversion alleles in other populations. Since we detected several tag SNPs in the CEU population for many of the inversions (up to 52 tag SNPs in HsInv004), we took into account the *in silico* genotyping results, which include individuals from many different populations, to select as tags those SNPs that seem to be more clearly associated to the inversion alleles and estimate inversion frequencies in the complete set of 1,092 genomes. It is important to notice that svgem calculations are based on genotype likelihoods both for the inversion and the SNPs. Only those SNPs maintaining an $r^2 > 0.95$ with the inversion in the total population were considered global tag SNPs, and the one located closer to the inversion breakpoints was used to estimate inversion allele frequencies in the 1000GP populations (Table S11). For those inversions not analyzed by BreakSeq/svgem, we selected as global tag SNP the tag SNP detected in the CEU individuals genotyped by PCR that is located within the inverted segment and is closest to one of the breakpoints, where recombination is less likely to separate both variants. Nevertheless, in some cases inversions without tag SNPs considering svgem data from all populations, may still have tag SNPs in a particular group of populations that are lost in other groups. For example, HsInv0006 has tag SNPs in European and African individuals but the tag SNPs (all located outside the inverted segment, data not shown) have been separated from the inversion by recombination in Asian individuals and their alleles can not be used to predict inversion genotypes in these populations.

In HsInv0052, validated previously by Pang et al. (2013), a 114.2-kb deletion polymorphic in human populations removes completely the sequence involved in the 2.3-kb polymorphic inversion (Figure 4A), creating a multi-allelic locus. A PCR experiment designed to detect this deletion, revealed a frequency of 40% in the CEU population (Table 3). Therefore, to establish the genotypes of the 1000GP individuals, we first searched for tag SNPs for this deletion based on the 35 and 60 genotyped individuals in common with the 1000GP and HapMap SNP data, respectively. Two tag SNPs with complete linkage to the deletion (rs206286 and rs206276) were found in both SNP datasets and the one closest to the indel breakpoint (rs206276 is only 250 bp away) was used to identify deleted chromosomes. After removing these deleted chromosomes, we used the chosen tag SNP for the inversion (Table S11) to distinguish *Std* and *Inv* alleles. Thus, in this case the alleles of both tag SNPs (deletion+inversion) were combined to obtain the genotypes of each individual in the 14 populations of the 1000GP.

In total, inversion frequencies in the 1000GP populations were obtained for 12 inversions using either BreakSeq/svgem frequency estimates (seven inversions) or global tag SNPs

linked to the inversion (nine inversions). Similar frequency values were obtained for the four inversions that could be analyzed by the two approaches (Tables 4 and S9). F_{st} values comparing the 14 populations independently and grouped by continent were calculated using the inversion allele frequencies determined with both methods (Tables 4 and S9). Two classes of inversions are clearly observed. Seven inversions show low F_{st} values (<0.1) suggesting that there are not significant differences in inversion frequencies between the different worldwide populations analyzed (Abecasis et al. 2010). However, another five inversions show clearly different frequencies in different populations or continents, with F_{st} values of 0.1-0.25 among populations and 0.11-0.32 among continents. This suggests that some evolutionary process might be generating the observed differences among these groups. Since F_{st} values for populations and continents are very similar, we can conclude that differentiation among populations is explained mainly by differences among continents. Africa or Asia, depending on the inversion, are often the continents that present unusual inversion frequencies compared to the other groups (Tables 4 and S9).

Finally, while in 13 of the inversions the ancestral allele shows an overall higher frequency, in 3 inversions (HsInv0003, HsInv0004 and HsInv0063) the derived allele is clearly the most frequent in human populations, suggesting a fast increase in frequency of these inversions within the human lineage. The remaining inversion is HsInv0052 where both *Std* and *Inv* allele show approximately the same frequency and the deletion of the complete region is the most frequent allele (40%). However, the most striking situation is that of HsInv0006, which exhibits an extremely high frequency (94.2%) of the derived allele (*Std*) in African populations, while it shows intermediate frequencies in all the other populations (F_{st} = 0.25 among continents) (Tables 2 and S9, and Figure 6). This pattern is consistent with selection in the African continent of the *Std* derived chromosome (Xue et al. 2009).

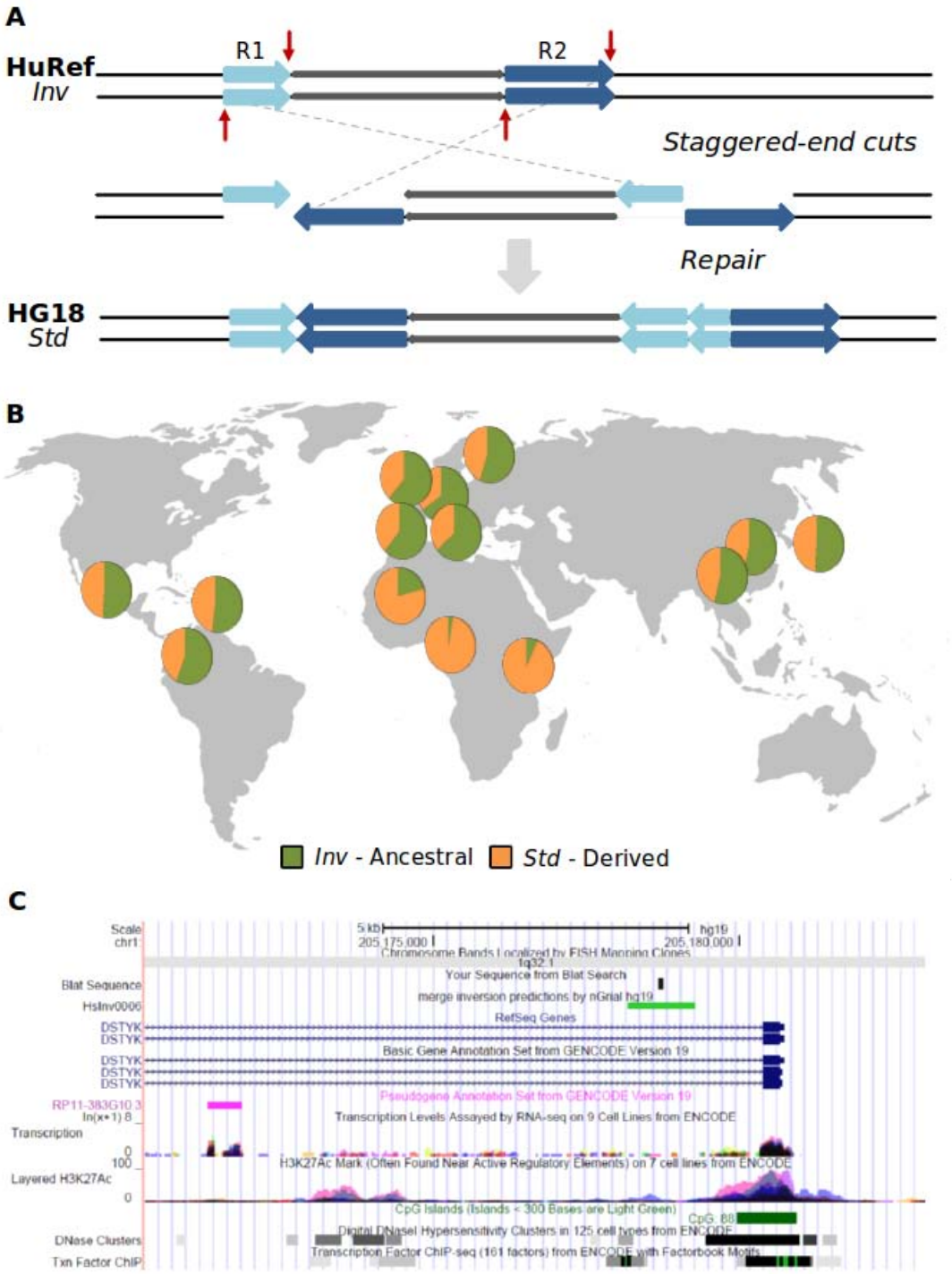

Figure 6. HsInv0006 generation, distribution and proximity to gene *DSTYK*.
A. Generation of HsInv0006 by staggered single-strand breaks (red arrows) and repair. A duplication of 41-bp long Repeat 2 (R2) is generated at the BP1 junction in the derived *Std* chromosomes. At BP2 an additional copy of 24-bp long Repeat 1 (R1) together with a 15-bp partial copy of this same repeat create a 39-bp insertion in the derived arrangement. **B.** Distribution of *Std* and *Inv* alleles in different human populations based on BreakSeq data (detection of next-generation sequence reads containing inversion breakpoints in the 1000GP Phase I data). The high frequency of the derived *Std* allele in Africa is consistent with positive

selection of this allele in this continent. **C.** Proximity of the 83-bp inverted sequence (black line in the first track) to the transcription start site and regulatory regions (indicated in the CpG island, histone modifications or transcription factor binding site tracks) of gene *DSTYK*, as shown in the UCSC genome browser. Genome coordinates correspond to the HG19 assembly to include tracks with ENCODE regulatory information.

Functional consequences of inversions

Since inversions are known to affect phenotypic traits in some species, we checked the relative position of the polymorphic inversions with respect to adjacent genes to try to find genes that might be affected by the inversion. RefSeq and GENCODE v19 annotations were used as reference, but expression evidences (ENCODE RNA-Seq data, ESTs, mRNAs) were also taken into account. Ten inversions are located in intergenic regions that do not seem to affect directly any coding region. In addition, we found that three inversions (HsInv0006, HsInv0059 and HsInv0061) are located completely within an intron of a gene: protein-coding genes *DSTYK* and *GABRR1*, and predicted non-coding RNA *LOC100422737*, respectively (Table S12). Three more inversions are located within introns of putative RNAs: HsInv0055 within a unitary pseudogene, and HsInv0031 and HsInv0052 within ESTs or mRNAs found in GenBank but that have not been integrated into any annotated transcript.

In two more cases (HsInv0030 and HsInv0069) two genes of the same family overlap each of the two SD copies at the inversion breakpoints. In HsInv0069, two non-coding RNA genes (*FAM225A* and *FAM225B*) are completely included within the 6.9-kb SDs. Since the two SDs have a high identity (99.1%) the exchange is not likely to cause any effect on the sequence and functionality of these genes. In HsInv0030, the genes at the breakpoints are *CTRB1* and *CTRB2* that encode chymotrypsinogen precursors expressed in pancreas. In this case, the promoter and the first exon of both genes, which includes the signal peptide, is exchanged between both copies in the alternative arrangement (Figure 4B). As already described by Pang et al. (2013), the few nucleotide differences between the two first and last exons have allowed us to detect hybrid mRNAs in GenBank, which confirm the existence of transcripts from the alternative arrangement: out of six chymotrypsinogen mRNAs available, four exhibit combinations of first exon of one *CTRB* gene and final exon of the other copy (the central exons are located within the SD and are identical between both gene copies), two correspond to *CTRB2* transcripts and none presents complete identity to *CTRB1*. However, it seems unlikely that this exchange causes any functional effects since both genes are reconstructed in the inverted chromosome, which just a slightly different amino acid combination in the beginning and end of the protein. What indeed disrupts one of the genes is a previously reported 584-bp deletion within SD1 that removes *CTRB2* exon 6 (Pang et al. 2013). This deletion occurred in the ancestral arrangement (*Inv* for this inversion since HG18 captured the

derived *Std* orientation) independently of the inversion (Figure 4B).

DISCUSSION

Inversions are one type of structural variant that is especially difficult to detect and validate since they often do not represent gain or loss of DNA and are often mediated by repeated sequences. In this work we have analyzed in detail each of the 90 regions predicted to harbor polymorphic inversions in the comparison of two independently assembled genomes (Levy et al. 2007). The first surprising result is that more than one third of these regions (31) correspond to variant calling errors. The errors in inversion prediction were mainly caused by mapping mistakes due to: (1) hybrid SDs in HuRef assembly (compared to HG18 SDs) due to either gene conversion between paralogous copies of the repeats (Martínez-Fundichely et al. in preparation), a different allele of another larger inversion (often missed in the HuRef-HG18 comparison) or to incorrectly assembled SDs, which allow that divergent segments among the two SD copies are detected as inverted sequences; (2) incomplete assembly of SDs in HuRef (in many cases only one in the inverted pair is present) which allows the detection of the single complete copy as an inverted alignment with one of the HG18 copies; or (3) absence of the true hit in HG18 because of polymorphic indels of TEs, that prompts the alignment with the next best hit in the genome, which happens to be in the opposite orientation. It is worth noting that at least in some cases the source of the error may be a real polymorphic inversion that causes an exchange of the internal part of two inverted copies of a SD pair. In these situations, it is possible that HuRef carries the inverted allele of the true inverted region and therefore a truly hybrid SD. However, if SDs are long the sequence between both SDs can easily be assembled in the incorrect orientation, and the real inversion allele can not be detected by sequence comparison only, making PEM data essential for the detection of inversions with long IRs at the breakpoints. Thus, the local mapping-based strategy for polymorphic inversion discovery used by Levy et al. (2007) turned out to be highly sensitive to gaps in HuRef, indel polymorphisms in either assembly or mapping into repeated sequences, which are precisely the regions where inversion breakpoints tend to occur, being these the main primary causes of detection of erroneous inverted alignments. While this approach may be useful for detecting indel variants, it does not seem to work well for inversions and these errors should be corrected by using a global whole-genome alignment approach.

Another 30 regions where inversions were predicted (33%) turned out to represent assembly errors in one of the genomes. These cases were solved by analyzing the DNA that is the primary source of the sequence (J. Craig Venter DNA or the BAC clones used in the reference genome assembly). It is remarkable that most of these errors (25 out 30, or 83.3%) correspond to misassembled regions in the reference genome assembly, rather than in HuRef, sequenced using a whole-genome shotgun strategy. Three of these regions

had been already corrected in the HG19 release and the orientation of five additional regions has been changed in the recent HG38 release, but 17 regions remain to be modified in future versions of the genome. Two of these errors (HsInv62 and HsInv0073) had already been reported before (Pang et al. 2013), although they had not been validated. As could be expected, all of these erroneous inversion predictions but one (HsInv0044) show IRs at the reported breakpoints (SDs, TEs…) that can explain the assembly in the incorrect orientation of the sequence comprised between them (Figure 7). Some have really large and highly identical SDs (10-11 kb in HsInv0029 with 6.5 kb of >99% identity at the breakpoints), which make them extremely complicated regions to assemble. Curiously, most HG18 errors correspond also to the longest inversion predictions. Five more inversion predictions (16.67%) were errors in the HuRef assembly. While it might be possible that HuRef is a heterozygote for these inversions and that the *Inv* allele is going undetected in our experiments, this seems unlikely since the results are the same when the two inversion breakpoints are tested. In two of these regions the segment incorrectly oriented in HuRef is comprised between two inverted *Alu* copies. In the other three cases, there are gaps in HuRef that may have contributed to the misassembly of these genomic regions. Therefore, these results exemplify the need for a careful analysis and validation of available structural variant predictions to be able to make reliable conclusions.

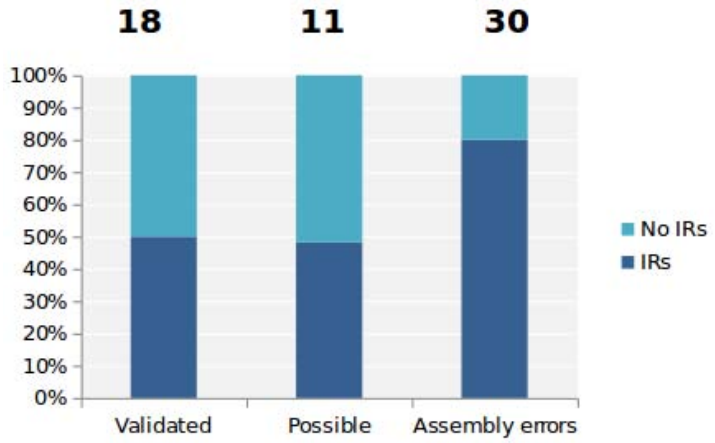

Figure 7. Inversion breakpoints in different groups of predictions.
Proportion of inversion predictions with (dark blue) and without (light blue) IRs at the breakpoints in different groups: validated polymorphic inversions, possible true inversions that have not been yet experimentally validated, assembly errors (both in HG18 and HuRef), and errors in assembly comparison. The number above each column indicates prediction counts included in each class.

Finally, only the final third, 29 regions, probably represent true inversion polymorphisms in human populations. We have validated experimentally 18 (62%) of these inversions and detected additional evidences (PEM data) for another 5 (79.3% all together). For the last 6 (20.7%), the comparison of HuRef and HG18 remains the only evidence for polymorphism but there is no reason to discard these predictions as false, specially taking

into account that all experimentally tested inversions that were classified as polymorphic candidates or were left unclassified according to PEM support, resulted in true polymorphisms. The only inversions that have not been experimentally validated are those with sizes <1 kb that were located in intergenic regions or those with extremely large SDs at the breakpoints because no genotyping assay able to cross the BPs was available. In these cases, the absence of other evidences of the inversion existence can be explained because small inversions are less likely to be represented in PEM data predictions (it is unlikely that one read of the pair maps within the inverted segment, and even less likely that this happens for several pairs) and only large PEM fragments can span across big SDs (Lucas Lledó and Cáceres 2013).

It is worth noting that, with sizes ranging from 83 to 16,560 bp, only relatively small inversions have been validated in this work as true polymorphic inversions. Long inversions are expected to be harder to detect by comparison of genome sequences due to the complicated repeats that flank many of them (Stefansson et al. 2005; Salm et al. 2012; Antonacci et al. 2009; Aguado et al. 2014) and that make these regions particularly challenging to assemble properly. So, the initial set might be already skewed towards small inversions due to the detection method, and therefore, it does not represent the complete set of inversions differentiating these two genomes. We have also seen here that the inverted region in inversions with IRs at their breakpoints tends to be longer (~2x) than in inversions with simple breakpoints (average length of validated inversions is 3,908 and 2,007 bp, respectively) suggesting that the presence of repeats facilitates the rearrangement of larger segments of DNA sequence (Pang et al. 2013).

In our analysis we have genotyped 17 validated polymorphic inversions in 90 CEU individuals organized in 30 family trios. We have not observed any *de novo* generation events and all the analyzed inversions seem to be segregating in human populations following Mendelian transmission. However, the analysis of nucleotide variation has revealed that 3 inversions out of the 17 genotyped in the CEU population (17.6%), show patterns consistent with recurrence. The main evidence derives from shared SNPs or haplotypes between *Std* and *Inv* chromosomes, which contrasts clearly with the high proportion of fixed SNPs found in other inversions and supports the idea that the inversions have been generated more than once on different haplotypes (Aguado et al. 2014). Genotyping errors have been ruled out because all inversion genotypes of individuals carrying haplotypes that suggest recurrence were double-checked in an independent experiment. Also, phasing errors in haplotype reconstruction are unlikely because the same effect (no fixed SNPs and SNP shared between arrangements) is observed without SNP phasing, and in HapMap data haplotype phasing is assisted by the trio information. Gene conversion of DNA segments between *Std* and *Inv* chromosomes could also generate the observed patterns of nucleotide variation (Aguado et al. 2014), but this process should affect equally all inversions, and no shared SNPs have been found within the inverted regions in any of the inversions without IRs at the breakpoints,

indicating that this phenomenon does not seem to have a big impact in inverted regions with sizes like the ones studied here. Consistent with this, the three recurrent inversions have the largest IRs, all of them with a high identity, which suggests that longer identical repeats may be more likely to pair and recombine by NAHR causing the inversion of the intervening segment. In support of this idea, some of the inversions with small IRs that show a low level of identity (like inversions mediated by *Alu* elements, Table 4), seem to have a unique origin. Thus, the comparison of the results from inversions generated by different mechanisms performed in this study provides strong support to the idea that IR mediated inversions show a degree of recurrence much higher than previously thought (Aguado et al. 2014).

An additional evidence for recurrence would be the detection of different breakpoints at nucleotide level for the different inversion and re-inversion events. However, in most cases exchange positions between repeat copies have been narrowed down to stretches with 100% identity between IRs, where it is not possible to differentiate diverse recombination events. Besides, we have to keep in mind that the inversions with IRs that do not show a recurrent pattern in CEU could still be recurrent in other populations (in which case, the tag SNPs derived from the CEU population might not be revealing the correct inversion genotypes and frequencies) or when comparing with non-human primates. Although recent studies have found that many inversions recurrent in the human lineage are also polymorphic in other species (Aguado et al. 2014), it does not seem to be the case for these 17 inversions, although a larger sample size would be needed to increase the chances of detecting polymorphism in non-human primates. However, we have detected one example that suggests that the inversion may have been generated a second time in one of these species apart from the human lineage: HsInv0055, which shows a different orientation in chimpanzee and gorilla genomes. Also, multiple recurrent events occurring on the same haplotype and in the same direction (from *Std* to *Inv* or viceversa) would not be uncovered by this analysis. Finally, inversions with no fixed or shared SNPs at all, have are considered unique, since this null hypothesis can not be discarded based on available data. It is then possible that we are still missing recurrence events.

On the other hand, most inversions without IRs have tag SNPs in perfect linkage disequilibrium with the inversion indicating that the inverted and standard haplotypes have a unique origin and have diverged for some time. These inversions are generated by mechanisms that do not require homology, like NHEJ, even though several of them may be mediated by microhomology. Remarkably, eight out of the nine inversions without IRs (88.9%) have indels at the breakpoint junctions completely linked to the inversion alleles. This association is confirmed by the direct PCR product sizes and by BreakSeq analysis, where alternative probes with no deletion failed to detect sequences in the 1000GP data. Taking into account the ancestral state determined from non-human primates, the allele with missing sequences at the breakpoints corresponds in all cases to the derived allele.

This indicates that the complete linkage between deletion-inversion alleles is probably due to a simultaneous generation of both variants by mechanisms like FoSTeS (Slack et al. 2006; Lee et al. 2007), which would play a larger role in the generation of these complex rearrangements than previously thought. Only in one case, HsInv0006, we find insertions at the breakpoints of the derived allele (*Std* or HG18) due to the generation of the inversion by the incorrect repair of staggered single-strand breaks (Ranz et al. 2007). This creates duplications of 41 bp in BP1 and 39 bp in BP2 in the derived allele, which were not present in the ancestral chromosome (*Inv* or HuRef), raising the possibility that the derived allele could revert to the ancestral allele, while the opposite is not possible due to the lack of repeated sequences (Figure 6). In comparison, only one inversion with IRs at its breakpoints, HsInv0030, shows a deletion within one of the repeats, and it is not related to the inversion generation since it is found only in some *Inv* ancestral chromosomes. Overall, the mechanisms of inversion generation of true polymorphic inversions are in agreement with those reported by Pang et al. (2013), with approximately 50% originated by NAHR mediated by repeats and another 50% not associated to repeated sequences (Figure 7), although it is not clear to what extent this is a representative sample of all the inversions in the human genome.

Inversion frequencies in 14 worldwide-distributed human populations have been examined for 12 inversions using at least one of two methods: global tag SNPs (9 inversions) or breakpoint detection by BreakSeq (7 inversions), with 4 inversions analyzed by both methods. For these last 4 inversions, global tag SNPs have been identified taking into account BreakSeq results, so the values obtained by the two analyses are extremely similar and most likely represent real frequencies (Tables 2 and S9). If frequencies are not exactly the same it is only because a higher number of individuals are genotyped by tag SNPs (indirect detection of the inversion allele) compared to those genotyped by BreakSeq (only feasible when reads covering the inversion breakpoints are available). The frequencies derived from BreakSeq results exclusively (3 inversions) are also based on data from individuals from the different populations, so their frequencies should also be reliable. On the other hand, frequency values for the 5 inversions genotyped only based on tag SNPs identified in CEU individuals should be taken with more caution, because the fact that these tag SNPs are in complete LD with the inversion in one population does not necessarily mean that they are linked also in all the other groups. For the CEU population, we also tested that the inversion frequencies calculated based on PCR genotyping (60 individuals) and on tag SNPs (1000GP CEU, 85 individuals) were not significantly different by comparing inversion frequency in the 50 1000GP individuals not genotyped by PCR to the frequency in the 35 individuals used to identify tag SNPs with a Fisher exact test. Only the deletion of the whole region in HsInv0052 showed different frequencies ($P = 0.038$) when comparing PCR vs. tag SNPs results, which might suggest that the tag SNPs for the deletion might not be completely linked to the deleted allele. If this is the case, *Std* and *Inv* allele frequencies might also be compromised since the 1000GP individuals have been

genotyped based on the inversion tag SNP once the deleted alleles had been taken into account.

All polymorphic inversions analyzed in this work show a geographic distribution comprising all continents, with no inversions specific of certain populations. Considering the ancestral state and allele frequency, we detect a higher global frequency of the derived allele (Tables 1 and 2) for three inversions (HsInv0003, HsInv0004 and HsInv0063), which may indicate positive selection although it can also be a consequence of demographic or random processes. For inversions HsInv0003 and HsInv004 the increase of the derived allele (*Inv* and *Std*, respectively) is observed in all populations, which would suggest a global advantage for the individuals carrying this allele. In the case of HsInv0063 we see an increased frequency of the derived *Inv* allele all over the world except in Africa (also reported previously in Pang et al. 2013), where *Std* remains the most common allele (70%). The out-of-Africa bottleneck can not be discarded as a cause of these differences, although they may also be a consequence of adaptation to new environments. We have also detected a certain degree of population differentiation in five inversions with F_{st} values >0.1 (Tables 2 and S9). All differences among populations seem to be caused by differences among continents. For example, HsInv0068 ancestral *Inv* allele frequencies range from being fixed in Asian populations to a frequency ~8% in American populations. Multiallelic locus HsInv0052 also shows differences among populations for the two derived alleles (deleted and *Inv*), with a higher frequency of the deletion accompanied by a lower presence of *Inv* in both African and Asian populations. However, the geographic distribution and frequency of these inversions could just be achieved by stochastic processes. Interestingly, inversions HsInv0059 and HsInv0006 show an increased frequency of the derived allele exclusively in one continent: Asia and Africa, respectively. While the increase in Asia can be explained by demographic reasons, a high frequency of the derived allele in Africa (90% in Africa vs. ~40% in the other continents) constitutes a pattern indicative of natural selection in African populations (Xue et al. 2009). HsInv0006 is the smallest inversion characterized here (83 bp) but it is located in the first intron of gene *DSTYK* (which encodes a serine/threonine and tyrosine protein kinase that may function as a regulator of cell death) and very close to the first exon (1,973 bp away from the transcription start site) and regulatory regions of this gene (1,209 bp away from a CpG island overlapping the first exon), which suggests that the inversion of this small region may be influencing its expression (Figure 6).

Another inversion, HsInv0059, is also located within the first intron of gene *GABRR1* encoding a gamma-aminobutyric acid (GABA) A receptor. It is interesting that, as well as HsInv0006, this inversion is also located close to the first exon (3,554 bp away from the transcription start site) in most of the known isoforms. Four more inversions can be located within introns of transcribed sequences but in transcripts annotated with different degrees of confidence including pseudogenes, lincRNAs or ESTs (Table S12). There are also two inversions (HsInv0030 and HsInv0069) that disrupt genes located within the IRs

that mediated the generation of the inversion. The most striking case is that of HsInv0030 because the inversion exchanges the first exons of the two copies of chymotrypsinogen found in the 1.5-SDs flanking the inverted sequence (Figure 4). Even though these two first exons are not identical, they both seem to encode a signal peptide and, since most of the coding part of the gene is found within the SDs, both genes are reconstructed and functional in the inverted arrangement. Hybrid mRNAs have been detected in GenBank database confirming gene expression of the rearranged chromosomes. Therefore, in spite of the genomic alteration involving gene disruption, no functional consequences are expected from this inversion. The remaining 10 polymorphic inversions are intergenic and located at different distances from genes, so if they influence gene expression it should be through regulatory elements able to act across longer distances. The most isolated inversion is HsInv0052, located in a 3.47 Mb region in chromosome 3q26.1 with no protein-coding genes and only a long non-coding RNA and some snoRNA or miRNA genes. This location of HsInv0052 in a low gene-density region probably makes possible the high frequency of the 114.2-kb polymorphic deletion that removes the whole inversion region but no genes (Figure 4). Thus, in this work, we have detected several inversions located close to genes or transcripts that are the best candidates to be studied in search for gene expression changes associated to inversion alleles, especially those with particular population distributions, such as HsInv0006. However, functional and phenotypic consequences of inversions remain difficult to study since samples of those tissues where genes adjacent to the inversion are expressed and functional are necessary, and knowledge about gene function is also incomplete in many instances, like for lincRNAs.

In summary, in this study we have validated 18 human polymorphic inversions and uncovered a high proportion of assembly mistakes both in the reference genome and in HuRef in the comparison of these two genomes. We have demonstrated that inversions are hard to predict and methods adapted to the detection of this specific type of variant, as well as experimental validation, are essential. We have also seen that the comparison of sequenced genomes only uncovers small inversions since inversions with long complex BPs are found in regions difficult to assemble. Half of the validated polymorphic inversions are mediated by repeats and generated by NAHR, and some of them seem to be recurrent in the human lineage. The other half are originated by non-homologous mechanisms usually involving loss of DNA content at the breakpoint junctions, raising the possibility that mechanisms like FoSTeS are more common than previously thought. Finally, we have identified a few inversions that are candidates to be targets of selection and that deserve a more detailed functional analysis. However, knowledge about human polymorphic inversions remains incomplete, especially because their effects on gene expression are yet difficult to explore, although they are key to understand the role of inversions in human variation, disease and evolution.

MATERIALS AND METHODS

Sequence alignment and breakpoint definition

HG18 inversion coordinates for the 90 inverted regions predicted in the J. Craig Venter genome (HuRef) by Levy et al. (2007) were obtained from the HuRef Project web site (http://huref.jcvi.org), and sequences corresponding to each inverted region plus 5 kb of flanking sequence at each side were recovered from the HG18 reference genome in the UCSC genome browser (https://genome.ucsc.edu/). Each HG18 sequence was then used as a query in a Blastn search (Altschul et al. 1990) against other human genomes to identify the corresponding sequences within the HuRef genome. The HuRef coordinates obtained from this Blastn search were then used to retrieve the sequences of the putatively inverted regions from the assembled J. Craig Venter chromosomes available at NCBI (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.3/Assembled_chromosomes/).

Next, for each inversion, both genome sequences (HG18 and HuRef) were aligned using Blast2seq (Altschul et al. 1990) to distinguish those parts that align in a direct orientation (flanking sequences) from those that align in inverted orientation (inverted region). Once aligned, we checked that the sequences fit the structure A-B-C-D for the standard reference genome and A-C-B-D for the inverted HuRef genome, and inversion breakpoints were annotated. All breakpoints were defined as a range of chromosomal positions to accommodate easily those located within repeated sequences where the exact breakpoint can not be delimited with single-nucleotide precision. In those inversions with IRs at the breakpoints, the repeats from both *Std* and *Inv* orientations from HG18, HuRef and other available sequences were aligned using MUSCLE (Edgar 2004) to identify those nucleotide changes that differentiate the two repeat copies in the *Std* arrangement, as well as the point where these variants get exchanged between the two repeats in the inverted chromosome. In these cases, the breakpoint boundaries were defined between the last three variable positions that indicate similarity between the two IRs at the same breakpoint junction in both genome assemblies, and three consecutive changes indicating an exchange between both IRs in HuRef inverted genome with respect to HG18 (that is, after the breakpoint, the IR in BP1 junction in the *Inv* chromosome becomes more similar to the IR in BP2 in the *Std* chromosome) (Figure 2A).

DNA samples

A total of 96 samples from the HapMap project (The International HapMap Consortium 2005) were used in this study (Table S6). These samples include 90 individuals of European origin (CEU population) corresponding to 30 parent-child trios, four independent individuals of African origin (YRI population), and two independent Asian

individuals (CHB and JPT populations). DNA from individual NA15510, previously analyzed by fosmid paired-end mapping (Kidd et al. 2008), was also used. High molecular weight genomic DNA from most of these samples was obtained from Epstein-Barr virus-transformed B-lymphoblastoid cell lines of each individual (Coriell Cell Repositories, Camden, New Jersey, USA) as previously described (Aguado et al. 2014). Identity of all the DNAs was confirmed using the MSK microsatellite kit (Coriell Cell Repositories, Camden, NJ, USA). DNA of the remaining cell lines and J. Craig Venter (HuRef) DNA was acquired from Coriell Cell Repositories (Camden, New Jersey, USA). BAC clones used in the human reference genome assembly were obtained from the CHORI BACPAC Resources Center (Oakland, California, USA), the RIKEN Bioresource Center DNA Bank (Ibaraki, Japan) and Source BioScience (Nottingham, UK). Bacteria were grown in LB agar plates with 30 μg/ml chloramphenicol or kanamycin depending on the BAC clone, and BAC DNA was either directly amplified from single colonies or isolated using the Plasmid Mini Kit (Qiagen). Four chimpanzee and two gorilla DNA samples, including a father-son pair in each species, were also used (Aguado et al. 2014). Genomic DNA from chimpanzee N457/03 and two gorillas (Z01/03 and Z02/03) were isolated from frontal cortex brain tissue obtained from the Banc de Teixits Animals de Catalunya (BTAC, Bellaterra, Barcelona, Spain). The three additional chimpanzee DNAs (PTR1211, PTR1213 and PTR1215) were extracted from Epstein-Barr virus-transformed B-lymphoblastoid cell lines generated from blood of three individuals from the Barcelona Zoo. All procedures that involved the use of human and non-human primate samples were approved by the Animal and Human Experimentation Ethics Committee (CEEAH) of the Universitat Autònoma de Barcelona.

PCR

Primers flanking each breakpoint were used to determine the orientation of the polymorphic inverted fragments in each analyzed individual. Primers were designed with Primer 3 (Rozen and Skaletsky 2000) and checked against both HG18, HuRef and dbSNP database to avoid including variable positions in the human genome. PCR reactions were prepared in a total volume of 25 μl containing 1x buffer, 1.5 mM $MgCl_2$, 0.2 μM of each dNTP, 0.4 μM of each primer, 1.5 U of Taq polymerase (BioTherm) and 50-100 ng of genomic DNA. In multiplex PCRs, a final concentration of 0.8 μM was used for the primer shared by the standard and inverted amplicons and 0.4 μM for the other two primers. An initial denaturation step of 5 min at 95 ºC was followed by 30-35 cycles at 95 ºC for 30 seconds, 59-62 ºC for 30 seconds as annealing step, and 72 ºC for 30-120 seconds as extension step, with a final extension at 72 ºC for 7 minutes. Multiplex PCR with three primers testing both orientations simultaneously was performed whenever possible to avoid genotyping errors caused by failed amplification reactions in one of the orientations if they were carried out separately. For long-range PCR (5-10 kb products), 100-200 ng of genomic DNA and 2.5 units of Pfu Turbo DNA polymerase (Stratagene)

were used. Cycling conditions were 92 ºC for 2 min, 35 cycles at 92 ºC for 10 sec, 60-66 ºC for 30 sec and 68 ºC for 10-15 min, and 68 ºC for 10 min. PCR products were analyzed by gel electrophoresis on 1.5-2% agarose gels stained with ethidium bromide (0.8-1% agarose for long PCR products). PCR amplifications from BAC clones were performed without a previous DNA isolation by resuspending a single colony in 100 µl of TE (10:0.1 Tris:EDTA proportions) and using 2 µl as a template. When needed, PCR products were directly sequenced by Sanger sequencing (Macrogen, Seoul, Korea).

Inverse PCR (iPCR)

Genomic DNA was digested with staggered-end restriction enzymes (Table S1) able to cut both within the inverted segment and outside but not within the IRs at the breakpoints. Typically, 150 ng of genomic DNA were digested overnight at 37 ºC (unless otherwise recommended) with 3 U of restriction enzyme. The enzyme was then heat-inactivated and the digested DNA was circularized at 25 ºC for 3 h in a ligation reaction containing 1x ligase buffer and 400 units of T4 DNA ligase (New England Biolabs) in a total volume of 175 µl. Ligase was heat-inactivated for 10 min at 65 ºC and 10 µl of the ligation product (≈8.6 ng of ligated DNA) were used as a template in a PCR reaction with primers spanning the restriction-ligation point.

Inversion genotyping using whole-genome sequencing data

We used the 1000 Genome Project Phase I data (Abecasis et al. 2012) to genotype the orientation of the inverted fragments by identifying reads that span simple breakpoints harboring no repeats. A total of seven inversions (Table 2) were analyzed in 1092 individuals belonging to 14 different worldwide human populations. A breakpoint library file with 100-bp sequences surrounding each breakpoint (50 bp at each side) was prepared. For each inversion four such sequences were generated corresponding to the two breakpoints in the two arrangements (*Std* and *Inv*) (Table S7). Additional sequences were added if insertions or deletions had been detected at the inversion breakpoints to include all possible combinations. Reads of the released individuals in the 1000GP (Abecasis et al. 2012) were downloaded from its ftp server. If the reads had already been aligned to the reference genome, only unmapped reads and reads mapped in the breakpoint regions were downloaded in SAM format with SAMtools (Li et al. 2009), and then converted to fastq. Otherwise, the raw fastq files were downloaded, avoiding color-space sequences and exome reads. The downloaded reads were processed with a slightly modified version of BreakSeq (Lam et al. 2010) as follows. First, we mapped the reads to the breakpoint library using Bowtie2 (instead of Bowtie, in the original BreakSeq). Reads overlapping at least 10 bases in either side of a breakpoint were retained (regardless of their length), and then mapped to the whole reference genome, in

order to check if they were unique. Only reads mapping on the reference conformation of a breakpoint (*Std* alleles) were expected to find also a hit in the reference genome at the breakpoint location. Reads mapping uniquely to an *Inv* or *Std* breakpoint were retained, and counted as allele observations, the rest being deleted.
Given the low coverage of most genomes analyzed (average 4x), allele observations were insufficient to determine individual genotypes with accuracy. Thus, to estimate the inverted allele frequencies directly from these counts, we used the svgem program (Lucas-Lledó et al. 2014), which implements an expectation-maximization algorithm, accounts for the genotype uncertainty, and delivers maximum likelihood estimates of allele frequency. A key parameter for svgem is the allele observation bias, namely how much more probable it is to observe the reference (*Std*) than the alternative (*Inv*) allele from a heterozygous genotype. The presence of repetitive or simple sequences around the breakpoints of one of the alleles can introduce important differences in their detectability. We estimated the allele observation bias of each inversion by generating from the breakpoint sequences all possible segments between 36 and 100 bp long that could possibly be observed as reads overlapping the breakpoints. Then, we mapped the simulated reads with the BreakSeq pipeline, as described above, and we used the ratio between reference and alternative allele observations as an estimate of the bias. After checking that individuals known to be homozygous by PCR assays did not show any spurious count of the absent allele, we assumed that erroneous observations were negligible, and set the corresponding parameter to 1.0E-05. With this information, and without assuming Hardy-Weinberg equilibrium, we run the svgem program on the counts of allele observations of all inversions, either in the whole sample, or in specific populations, to estimate the respective frequencies of the inverted alleles.

In order to calculate the linkage disequilibrium between the inversions and the surrounding SNPs in the different 1000GP populations, we used a likelihood framework that accounts for the uncertainty of the genotypes both at the inversion and the SNP loci. First, for each inversion we downloaded a vcf file containing the likelihoods of the genotypes of all the individuals in all SNPs within 10 kb from the inversion breakpoints from the 1000 Genomes Project website (www.1000genomes.org). Then, we used the svgem program (Lucas-Lledó et al. 2014) to calculate the genotype likelihoods at the inversion loci in all individuals, from the counts of reads that specifically overlapped either allele of the inversion's breakpoints, obtained before with BreakSeq (Lam et al. 2010). For these calculations, we determined the most suitable value of the lambda parameter (odds of observing the reference allele from a heterozygous genotype) for each inversion from simulated reads, because the exact sequences at the breakpoints are known. Once we had the genotype likelihoods of the inversions, we added them to the vcf file in the positions of the breakpoints as a pair of perfectly linked single nucleotide variants. Finally, we used bcftools (Li 2011) to estimate the r^2 among SNPs and breakpoints from the genotype likelihoods. This way, we optimally use the data at hand, and avoid the use of imputed genotypes, which would bias the estimation of the linkage

disequilibrium. The SNP with high linkage ($r^2 > 0.98$) in the 1000GP populations, and located the closest to the inversion breakpoints was considered a tag SNP for the inversion alleles and was used to genotype the whole set of 1,092 individuals. For those inversions not included in the BreakSeq analysis, the tag SNP with $r^2 = 1$ in the CEU population (see below) located closest to one of the inversion breakpoints was used to genotype the inversion in the remaining populations.

Nucleotide variation analysis

Ancestral state was estimated for polymorphic inversions both by aligning the human reference sequence (HG18 assembly) with the chimpanzee (panTro4), gorilla (gorGor3) and Rhesus macaque (macRhe3) genomes using Blast2seq (Altschul et al. 1990), and experimentally by PCR or iPCR in four chimpanzees and two gorillas (see above). Nucleotide variation associated to inversions was investigated for the 17 inversions genotyped in the 90 CEU individuals (HapMap PT01). SNP data for the inversion region plus 10 kb of flanking sequence at either side were collected from both the 1000GP Phase I (Abecasis et al. 2012) and HapMap project release 27 (Altshuler et al. 2010). Inversion genotype data was available for 35 unrelated individuals in the 1000GP and for 60 independent individuals in the HapMap project (even though less SNPs are genotyped for each individual). SNPs included within IRs or indels were excluded from the analysis.

Shared polymorphisms between *Inv* and *Std* chromosomes were estimated based on the presence of polymorphic SNPs in *Std*/*Std* and *Inv*/*Inv* homozygotes, of *Std*/*Inv* heterozygotes homozygous for both alleles of a SNP, or of a SNP allele polymorphic in one orientation and not in the other one (Aguado et al. 2014). Linkage disequilibrium between SNPs and polymorphic inversions was assessed using Haploview v4.2 software (Barrett et al. 2005) and r^2 statistic. We defined as tag SNPs in the European population those with an r^2 value of 1 and therefore in complete LD with the inversion. To estimate haplotypes for both orientations we used Phase software v2.1.1 (Stephens et al. 2001) and median-Joining haplotype networks were constructed using Network 4.612 software (Bandelt et al. 1999). In the case of HsInv0052, with three alleles including the complete deletion of the inverted region (Figure 4A), first we identified a tag SNP for the deletion in the 1000GP data. Then the individuals homozygous for the deletion were removed from the analysis and in individuals heterozygous for the deletion only one chromosome was taken into account.

Data

Detailed information about all inversions and inversion predictions described here can be found at the INVFEST database (http://invfestdb.uab.cat/).

AKNOWLEDGEMENTS

We thank Marta Morell, Raquel Rubio-Acero, Francisca Garcia, and Francisco Cortés for their help with lymphoblastoid cell lines cultures, and Lorraine Toji from the Coriell Institute for help with the confirmation of cell lines genotype. We are also grateful to the Coriell Institute, the Barcelona Zoo, the Banc de Teixits Animals de Catalunya (BTAC), CHORI BACPAC Resources Center, the RIKEN Bioresource Center DNA Bank and Source BioScience for providing the human cell lines, primate blood samples, primate tissue samples, and BAC clones used in this study.

FUNDING

This work was supported by the European Research Council (ERC) Starting Grant 243212 (INVFEST) under the European Union Seventh Research Framework Programme (FP7) to M.C., a MAEC-AECI doctoral fellowship from the Ministerio de Asuntos Exteriores y Cooperación (Spain) to A.M.F., a Beatriu de Pinós Postdoctoral fellowship from the Generalitat de Catalunya (Spain) to J.I.L.LL, and a research PRIC grant from the Barcelona Zoo (Ajuntament de Barcelona, Spain) to A.R.H.

AUTHORS CONTRIBUTIONS

Conceived and designed the experiments: DVS, MP, CA, MC. Performed the experiments: DVS, DI, CA. Analyzed the data: DVS, MP, MGV, AMF, JILL, MC. Contributed reagents/materials/analysis tools: XE, ARH. Wrote the paper: DVS, MP, JILL, MC.

REFERENCES

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467: 1061–73.

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

Aguado C, Gayà-Vidal M, Villatoro S, Oliva M, Izquierdo D, Giner-Delgado C, Montalvo V, García-González J, Martínez-Fundichely A, Capilla L, et al. 2014. Validation and genotyping of multiple human polymorphic inversions mediated by inverted repeats reveals a high degree of recurrence. ed. G.S. Barsh. PLoS Genet 10: e1004208.

Ahn S-M, Kim T-H, Lee S, Kim D, Ghang H, Kim D-S, Kim B-C, Kim S-Y, Kim W-Y, Kim C, et al. 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res 19: 1622–9.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet 12: 363–76.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215: 403–10.

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Bonnen PE, de Bakker PIW, Deloukas P, Gabriel SB, et al. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–8.

Alves JM, Lopes AM, Chikhi L, Amorim A. 2012. On the structural plasticity of the human genome: chromosomal inversions revisited. Curr Genomics 13: 623–32.

Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, Eichler EE. 2009. Characterization of six human disease-associated inversion polymorphisms. Hum Mol Genet 18: 2555–66.

Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16: 37–48.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–5.

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen H-C, Agarwala R, McLaren WM, Ritchie GRS, et al. 2011. Modernizing reference genome assemblies. PLoS Biol 9: e1001091.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464: 704–12.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–7.
Feuk L. 2010. Inversion variants in the human genome: role in disease and genome architecture. Genome Med 2: 11.

Feuk L, MacDonald JR, Tang T, Carson AR, Li M, Rao G, Khaja R, Scherer SW. 2005. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. PLoS Genet 1: e56.

González JR, Cáceres A, Esko T, Cuscó I, Puig M, Esnaola M, Reina J, Siroux V, Bouzigon E, Nadif R, et al. 2014. A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. Am J Hum Genet 94: 361–72.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. Nat Rev Genet 10: 551–64.

Hoffmann AA, Rieseberg LH. 2008. Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? Annu Rev Ecol Evol Syst 39: 21–42.

Imsland F, Feng C, Boije H, Bed'hom B, Fillon V, Dorshorst B, Rubin C-J, Liu R, Gao Y, Gu X, et al. 2012. The Rose-comb mutation in chickens constitutes a structural rearrangement causing both altered comb morphology and defective sperm motility. ed. D. Burt. PLoS Genet 8: e1002775.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484: 55–61.

Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L, et al. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature 477: 203–6.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56–64.

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell 143: 837–47.

Kirkpatrick M. 2010. How and why chromosome inversions evolve. PLoS Biol 8.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. Science 318: 420–6. http://www.sciencemag.org/content/318/5849/420.abstract (Accessed April 28, 2014).

Lam HYK, Mu XJ, Stütz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB. 2010. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol 28: 47–55.

Lee JA, Carvalho CMB, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131: 1235–47.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al. 2007. The diploid genome sequence of an individual human. PLoS Biol 5: e254.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–93.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–9.

Lowry DB, Willis JH. 2010. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. PLoS Biol 8.

Lucas Lledó JI, Cáceres M. 2013. On the power and the systematic biases of the detection of chromosomal inversions by paired-end genome sequencing. PLoS One 8: e61292.

Lucas-Lledó JI, Vicente-Salvador D, Aguado C, Cáceres M. 2014. Population genetic analysis of bi-allelic structural variants from low-coverage sequence data with an expectation-maximization algorithm. BMC Bioinformatics 15: 163.

Martin J, Han C, Gordon LA, Terry A, Prabhakar S, She X, Xie G, Hellsten U, Chan YM, Altherr M, et al. 2004. The sequence and analysis of duplication-rich human chromosome 16. Nature 432: 988–94.

Martínez-Fundichely A, Casillas S, Egea R, Ràmia M, Barbadilla A, Pantano L, Puig M, Cáceres M. 2014. InvFEST, a database integrating information of polymorphic inversions in the human genome. Nucleic Acids Res 42: D1027–32.

Martínez-Fundichely A, Oliva M, Vicente-Salvador D, Aguado C, Izquierdo D, Villatoro S, Novoa A, Estivill X, Lucas-Lledó JI, Puig M, et al. Accurate characterization of inversions in the human genome from paired-end mapping data with the GRIAL algorithm. Prep.

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 19: 1527–41.

McVey M, Lee SE. 2008. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet 24: 529–38.

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, et al. 2013. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res 23: 749–61.

Onishi-Seebacher M, Korbel JO. 2011. Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond. Bioessays 33: 840–50.

Pang AWC, Migita O, Macdonald JR, Feuk L, Scherer SW. 2013. Mechanisms of formation of structural variation in a fully sequenced human genome. Hum Mutat 34: 345–54.

Puig M, Cáceres M, Ruiz A. 2004. Silencing of a gene adjacent to the breakpoint of a widespread Drosophila inversion by a transposon-induced antisense RNA. Proc Natl Acad Sci U S A 101: 9013–9018.

Ranz JM, Maurin D, Chan YS, von Grotthuss M, Hillier LW, Roote J, Ashburner M, Bergman CM. 2007. Principles of genome evolution in the Drosophila melanogaster species group. ed. M.A.F. Noor. PLoS Biol 5: e152.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. Nature 444: 444–54.

Rozen S, Skaletsky and HJ. 2000. Primer3 on the WWW for general users and for biologist programmers. In Bioinformatics Methods and Protocols: Methods in Molecular Biology (eds. S. Krawetz and S. Misener), pp. 365–386, Humana Press, Totowa, NJ.

Salm MPA, Horswell SD, Hutchison CE, Speedy HE, Yang X, Liang L, Schadt EE, Cookson WO, Wierzbicki AS, Naoumova RP, et al. 2012. The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome Res 22: 1144–53.

Slack A, Thornton PC, Magner DB, Rosenberg SM, Hastings PJ. 2006. On the mechanism of gene amplification induced under stress in Escherichia coli. PLoS Genet 2: e48.

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, et al. 2005. A common inversion under selection in Europeans. Nat Genet 37: 129–37.

Steinberg KM, Antonacci F, Sudmant PH, Kidd JM, Campbell CD, Vives L, Malig M, Scheinfeldt L, Beggs W, Ibrahim M, et al. 2012. Structural diversity and African origin of the 17q21.31 inversion polymorphism. Nat Genet 44: 872–80.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68: 978–89.

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437: 1299–320.

Thomas JW, Cáceres M, Lowman JJ, Morehouse CB, Short ME, Baldwin EL, Maney DL, Martin CL. 2008. The chromosomal polymorphism linked to variation in social behavior in the white-throated sparrow (Zonotrichia albicollis) is a complex rearrangement and suppressor of recombination. Genetics 179: 1455–68.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, et al. 2008. The diploid genome sequence of an Asian individual. Nature 456: 60–5.

Xue Y, Zhang X, Huang N, Daly A, Gillson CJ, Macarthur DG, Yngvadottir B, Nica AC, Woodwark C, Chen Y, et al. 2009. Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. Genetics 183: 1065–77.

SUPPORTING INFORMATION

Table S1. Inversion classification. The results of each step of the analysis are shown in different columns. For those inversions tested by inverse PCR (iPCR) the restriction enzyme used is also indicated in the Experiment column.

Inversion	Position	Sequence analysis	PEM analysis	Experiment	Result	Additional references
HsInv0001	chr1:2475133-2489144	Candidate region	HG18 error candidate	Corrected in HG19	HG18 error	
HsInv0002	chr1:26840409-26845806	Candidate region	HG18 error candidate	Long-range PCR	HG18 error	
HsInv0003	chr1:185733101-185733350	Candidate region	Unclassified	PCR	Polymorphic	
HsInv0004	chr1:196023846-196024607	Candidate region	Polymorphic candidate	PCR	Polymorphic	Pang et al. 2012
HsInv0005	chr1:200051763-200051865	Candidate region	HuRef error candidate	PCR	HuRef error	
HsInv0006	chr1:203445294-203445397	Candidate region	Unclassified	PCR	Polymorphic	
HsInv0007	chr1:225748162-225750091	Comparison error	-	-	-	
HsInv0008	chr1:245357729-245358039	Comparison error	-	-	-	
HsInv0009	chr1:245400251-245400562	Comparison error	-	-	-	
HsInv0010	chr1:246733501-246734062	Comparison error	-	-	-	
HsInv0011	chr1:246735283-246736959	Comparison error	-	-	-	
HsInv0012	chr1:246749189-246758249	Candidate region	Unclassified	Not analyzed	-	
HsInv0013	chr1:246773620-246774406	Comparison error	-	-	-	
HsInv0014	chr10:4994874-5036786	Candidate region	Polymorphic candidate	Not analyzed	-	
HsInv0015	chr10:13145105-13145278	Candidate region	Unclassified	Not analyzed	-	
HsInv0016	chr10:58927667-58927951	Candidate region	Unclassified	Not analyzed	-	
HsInv0017	chr10:93191655-93197436	Candidate region	Polymorphic candidate	Not analyzed	-	
HsInv0018	chr11:1871848-1893342	Comparison error	-	-	-	
HsInv0019	chr11:49693128-49706324	Candidate region	HG18 error candidate	Long-range PCR	HG18 error	
HsInv0020	chr11:61619532-61624597	Candidate region	Polymorphic candidate	Not analyzed	-	
HsInv0021	chr12:12436246-12437784	Candidate region	HG18 error candidate	PCR	HG18 error	
HsInv0022	chr12:13441349-13441835	Candidate region	HG18 error candidate	PCR	HG18 error	
HsInv0023	chr12:17815364-17902973	Candidate region	HG18 error candidate	iPCR-NsiI	HG18 error	
HsInv0024	chr12:79370385-79381831	Candidate region	HG18 error candidate	iPCR-NsiI	HG18 error	
HsInv0025	chr12:85764376-85777047	Candidate region	HG18 error candidate	PCR	HG18 error	
HsInv0026	chr13:49412159-49412577	Comparison error	-	-	-	
HsInv0027	chr14:94474662-94474828	Comparison error	-	-	-	
HsInv0028	chr15:28834320-28834859	Comparison error	-	-	-	
HsInv0029	chr16:1227125-1241677	Candidate region	HG18 error candidate	iPCR-EcoRI	HG18 error	
HsInv0030	chr16:73797599-73814159	Candidate region	Polymorphic candidate	PCR	Polymorphic	Pang et al. 2012
HsInv0031	chr16:83746237-83747302	Candidate region	Unclassified	iPCR-EcoRI	Polymorphic	Pang et al. 2012
HsInv0032	chr17:5826739-5827291	Candidate region	Polymorphic candidate	Not analyzed	-	
HsInv0033	chr17:40566233-40567384	Candidate region	HuRef error candidate	PCR	HuRef error	
HsInv0034	chr17:55552838-55556395	Comparison error	-	-	-	
HsInv0035	chr17:57999778-58000250	Candidate region	HuRef error candidate	PCR	HuRef error	
HsInv0036	chr18:12134389-12137214	Candidate region	Polymorphic candidate	Long-range PCR	Polymorphic	
HsInv0037	chr19:4690220-4690480	Comparison error	-	-	-	
HsInv0038	chr19:43956215-43971463	Candidate region	HG18 error candidate	iPCR-NsiI	HG18 error	
HsInv0039	chr19:56177914-56178263	Comparison error	-	-	-	
HsInv0040	chr2:138721419-138725673	Candidate region	HG18 error candidate	iPCR-HindIII	Polymorphic	
HsInv0041	chr2:225001224-225001326	Candidate region	Unclassified	PCR	Polymorphic	
HsInv0042	chr2:234136102-234151235	Candidate region	HG18 error candidate	Corrected in HG19	HG18 error	
HsInv0043	chr21:14350147-14350946	Comparison error	-	-	-	
HsInv0044	chr21:26296029-26296570	Candidate region	HG18 error candidate	PCR	HG18 error	
HsInv0045	chr21:26942554-26943508	Candidate region	Polymorphic candidate	iPCR-SacI	Polymorphic	
HsInv0046	chr21:40321769-40331695	Candidate region	HG18 error candidate	iPCR-NsiI	HG18 error	
HsInv0047	chr22:22622523-22632218	Comparison error	-	-	-	
HsInv0048	chr3:33130423-33131955	Comparison error	-	-	-	
HsInv0049	chr3:44716017-44717266	Candidate region	HG18 error candidate	PCR	HG18 error	
HsInv0050	chr3:50900409-50910036	Candidate region	HG18 error candidate	Corrected in HG19 + iPCR/PCR	HG18 error	
HsInv0051	chr3:57362070-57362262	Candidate region	Unclassified	Not analyzed	-	
HsInv0052	chr3:164028056-164030335	Candidate region	Unclassified	iPCR-HindIII	Polymorphic	Pang et al. 2012
HsInv0053	chr3:188618642-188626799	Candidate region	HG18 error candidate	iPCR-NsiI	HG18 error	
HsInv0054	chr4:128001907-128002158	Comparison error	-	-	-	
HsInv0055	chr5:63800180-63811001	Candidate region	Polymorphic candidate	iPCR-BamHI	Polymorphic	

HsInv0056	chr5:93929739-93930732	Comparison error	-	-	-	
HsInv0057	chr5:178997079-179016954	Candidate region	HG18 error candidate	iPCR-KpnI	HG18 error	
HsInv0058	chr6:31117201-31118074	Candidate region	Polymorphic candidate	PCR	Polymorphic	
HsInv0059	chr6:89980353-89980661	Candidate region	Unclassified	PCR	Polymorphic	
HsInv0060	chr6:94797796-94798133	Candidate region	Unclassified	Not analyzed	-	
HsInv0061	chr6:107275899-107277573	Candidate region	HG18 error candidate	iPCR-HindIII	Polymorphic	
HsInv0062	chr6:168835529-168836601	Candidate region	HG18 error candidate	PCR	HG18 error	Pang et al. 2012
HsInv0063	chr7:70064121-70076815	Candidate region	Polymorphic candidate	PCR	Polymorphic	Pang et al. 2012
HsInv0064	chr7:106846308-106850371	Candidate region	HG18 error candidate	PCR	HG18 error	
HsInv0065	chr7:143147370-143593175	Comparison error	-	-	-	
HsInv0066	chr8:6142791-6144697	Candidate region	HG18 error candidate	PCR	HG18 error	
HsInv0067	chr8:48377484-48388781	Candidate region	HG18 error candidate	iPCR-ApaI	HG18 error	
HsInv0068	chr9:76087960-76088209	Candidate region	Unclassified	PCR	Polymorphic	
HsInv0069	chr9:114913782-114914991	Candidate region	Unclassified	iPCR-NsiI	Polymorphic	
HsInv0070	chrX:6169320-6169676	Comparison error	-	-	-	
HsInv0071	chrX:28821431-28821628	Comparison error	-	-	-	
HsInv0072	chrX:45433293-45435559	Candidate region	Polymorphic candidate	iPCR-HindIII	Polymorphic	
HsInv0073	chrX:46695748-46715632	Candidate region	HG18 error candidate	PCR	HG18 error	Pang et al. 2012
HsInv0074	chrX:48900567-48906237	Candidate region	HG18 error candidate	PCR	HG18 error	
HsInv0075	chrX:51445467-51451530	Comparison error	-	-	-	
HsInv0076	chrX:51456609-51460366	Candidate region	Polymorphic candidate	Not analyzed	-	
HsInv0077	chrX:52919709-52924693	Comparison error	-	-	-	
HsInv0078	chrX:53868900-53870123	Candidate region	HuRef error candidate	PCR	HuRef error	
HsInv0079	chrX:62321770-70858721	Comparison error	-	-	-	
HsInv0080	chrX:72139321-72140074	Comparison error	-	-	-	
HsInv0081	chrX:78809884-78810019	Candidate region	Unclassified	Not analyzed	-	
HsInv0082	chrX:96081284-96082949	Candidate region	HuRef error candidate	PCR	HuRef error	
HsInv0083	chrX:98726343-98726752	Comparison error	-	-	-	
HsInv0084	chrX:103191036-119081776	Comparison error	-	-	-	
HsInv0085	chrX:141103924-141104312	Comparison error	-	-	-	
HsInv0086	chrX:149322178-149335950	Candidate region	HG18 error candidate	PCR	HG18 error	
HsInv0087	chrX:153276592-153279714	Comparison error	-	-	-	
HsInv0088	chrY:15212538-15213321	Comparison error	-	-	-	
HsInv0089	chrY:18573501-19260222	Comparison error	-	-	-	
HsInv0090	chrY:21632695-21635926	Candidate region	HG18 error candidate	iPCR-NsiI	HG18 error	

Table S2. Errors in assembly comparison. Associated inversions in the regions of the inversion predictions are shown for mapping mistakes within segmental duplications (SD). Data about these associated polymorphic inversions are given in the Position, Inversion size and Inverted repeats (IRs) columns.

Inversion prediction	Alignment	Gaps [1]	PEM support	HG18 sequence elements in mapping position	Associated inversion	Position	Inversion size (bp)	IR Size (bp)	IR Identity (%)
Mapping mistakes within inverted SD pairs									
Hsinv0007	No inverted alignment	No	No	SD1	Hsinv0229	chr1:225722441-225789475	39307	13220 / 10577	92.30
Hsinv0008	No inverted alignment	No	No	SD1	Hsinv0180	chr1:245357627-245400877	41947	2792 / 2795	99.00
Hsinv0009	No inverted alignment	No	No	SD2					
Hsinv0010	No inverted alignment	Yes	No	SD1					
Hsinv0011	No inverted alignment	Yes	No	SD1 (gap in HuRef)	Hsinv0012 [2]	chr1:246748189-246759249	9061	59480 / 59611	98.80
Hsinv0013	No inverted alignment	Yes	No	SD2					
Hsinv0028	No inverted alignment	Yes	No	SD1	Hsinv0358 [3]	chr15:28768745-29805179	1025946	103182 / 105628	97.70
Hsinv0034	No inverted alignment	Yes	No	SD2	Hsinv0371 [3]	chr17:55465239-55554129	80469	16438 / 16453	99.70
Hsinv0047	No inverted alignment	No [4]	No	SD1	Hsinv0047 [4]	chr22:22621523-22633218	9696	29617 / 29623	98.50
Hsinv0075	No inverted alignment	Yes	No	SD1 (gap in HuRef)	Hsinv0076-391 [2]	chrX:51422445-51493352	44037	25745 / 25360	99.70
Hsinv0077	No inverted alignment	Yes	No	SD1	Hsinv0819	chrX:52938670-53011944	48591	47968 / 42737	99.00
Hsinv0080	No inverted alignment	Yes	No	SD1	Hsinv0896 [3]	chrX:72132652-72223499	81700	9496 / 9479	99.70
Hsinv0056	No inverted alignment	Yes	No	SD1	-	-	5478104	2519 / 2521	90.60
Hsinv0085	No inverted alignment	Yes	No	SD2	-	-	266712	8583 / 8621	97.10
Inverted duplications only in one assembly									
Hsinv0018	Inverted and direct alignment	Yes	No	-	-				
Hsinv0026	Inverted and direct alignment	Yes	No	-	-				
Hsinv0037	Inverted and direct alignment	Yes	No	-	-				
Hsinv0048	Inverted and direct alignment	Yes	No	-	-				
Hsinv0070	Inverted and direct alignment	Yes	No	-	-				
Hsinv0087	Inverted and direct alignment	Yes	No	-	-				
Polymorphic indels of repetitive sequences									
Hsinv0039	No inverted alignment	No	No	MSR1 satellite sequence	-				
Hsinv0027	No inverted alignment	Yes	No	(TG)n simple repeat	-				
Hsinv0043	No inverted alignment	Yes	No	(CTAGGG)n simple repeat	-				
Hsinv0054	No inverted alignment	Yes	No	L1PA3 (LINE element)	-				
Hsinv0071	No inverted alignment	Yes	No	HERVH (ERV element)	-				
Hsinv0083	No inverted alignment	Yes	No	L1PB (LINE element)	-				
Hsinv0088	No inverted alignment	Yes	No	HERV9 (ERV element)	-				
Unknown/unclear cause									
Hsinv0079		Yes	No	-	-				
Hsinv0084		No	No	-	-				
Hsinv0089		No	No	-	-				
Hsinv0065		No	No	-	-				

[1] Gaps should be relevant for the error to occur

[2] Putative polymorphic inversions with PEM support (Martinez-Fundichely et al. In preparation)

[3] Contained within inversion experimentally validated by Antonacci et al. 2009

[4] 37-kb deletion supported by fosmid PEM and present in HuRef that removes one copy of a SD pair.

[5] Inversion experimentally validated by Aguado et al. 2014

Table S3. Breakpoint definition in the 59 polymorphic candidates. Breakpoints defined by Levy et al. 2007 are compared to our results.

		Levy et al. 2007			Present work					Differences			
Inversion	**Chr.**	**Inversion start**	**Inversion end**	**Size (bp)**	**BP1**	**BP2**	**Inversion start**	**Inversion end**	**Size (bp)**	**Start (bp)**	**End (bp)**	**Size (bp)**	**BP definition**
HsInv0001	chr1	2475133	2489144	14012	2474382-2475133	2489145-2489896	2475134	2489144	14011	1	0	-1	Accurate
HsInv0002	chr1	26840409	26845806	5398	26839407-26840265	26845951-26846809	26840266	26845950	5685	-143	144	287	Imprecise
HsInv0019	chr11	49693128	49706324	13197	49691860-49692868	49706585-49707593	49692869	49706584	13716	-259	260	519	Imprecise
HsInv0021	chr12	12436246	12437784	1539	12436093-12436153	12437832-12437892	12436154	12437831	1678	-92	47	139	Imprecise
HsInv0022	chr12	13441349	13441835	487	13441349-13441350	13441835-13441836	13441351	13441834	484	2	-1	-3	Accurate
HsInv0023	chr12	17815364	17902973	87610	17813889-17814515	17903823-17904449	17814516	17903822	89307	-848	849	1697	Imprecise
HsInv0024	chr12	79370385	79381831	11447	79369050-79370199	79382018-79383167	79370200	79382017	11818	-185	186	371	Imprecise
HsInv0025	chr12	85764376	85777047	12672	85763520-85764349	85777075-85777904	85764350	85777074	12725	-26	27	53	Imprecise
HsInv0029	chr16	1227125	1241677	14553	1221550-1228054	1238397-1244871	1228055	1238396	10342	930	-3281	-4211	Imprecise
HsInv0038	chr19	43956215	43971463	15249	43954882-43956117	43971562-43972797	43956118	43971561	15444	-97	98	195	Imprecise
HsInv0042	chr2	234136102	234151235	15134	234134828-234135583	234151755-234152510	234135584	234151754	16171	-518	519	1037	Imprecise
HsInv0044	chr21	26296029	26296570	542	26296023-26296029	26296571-26296577	26296030	26296570	541	1	0	-1	Accurate
HsInv0046	chr21	40321769	40331695	9927	40317223-40317736	40331906-40332419	40317737	40331905	14169	-4032	210	4242	Imprecise
HsInv0049	chr3	44716017	44717266	1250	44715995-44716017	44717267-44717289	44716018	44717266	1249	1	0	-1	Accurate
HsInv0050	chr3	50900409	50910036	9628	50899680-50900357	50910049-50910726	50900358	50910048	9691	-51	12	63	Imprecise
HsInv0053	chr3	188618642	188626799	8158	188614227-188615326	188628199-188629298	188615327	188628198	12872	-3315	1399	4714	Imprecise
HsInv0057	chr5	178997079	179016954	19876	178993587-178996987	179015092-179018173	178996988	179015091	18104	-91	-1863	-1772	Imprecise
HsInv0062	chr6	168835529	168836601	1073	168835082-168835529	168836602-168837049	168835530	168836601	1072	1	0	-1	Accurate
HsInv0064	chr7	106846308	106850371	4064	106845789-106846308	106850372-106850891	106846309	106850371	4063	1	0	-1	Accurate
HsInv0066	chr8	6142791	6144697	1907	6141940-6142692	6144797-6145549	6142693	6144796	2104	-98	99	197	Imprecise
HsInv0067	chr8	48377484	48388781	11298	48376601-48377415	48393421-48394235	48377416	48393420	16005	-68	4639	4707	Imprecise
HsInv0073	chrX	46695748	46715632	19885	46695638-46695777	46715616-46715725	46695778	46715615	19838	30	-17	-47	Imprecise
HsInv0074	chrX	48900567	48906237	5671	48900144-48900536	48906269-48906661	48900537	48906268	5732	-30	31	61	Imprecise
HsInv0086	chrX	149322178	149335950	13773	149322178-149322179	149336021-149336022	149322180	149336020	13841	2	70	68	Imprecise
HsInv0090	chrY	21632695	21635926	3232	21619042-21619643	21635019-21635970	21619644	21635018	15375	-13051	-908	12143	Imprecise
HsInv0005	chr1	200051763	200051865	103	200051521-200051763	200051866-200052108	200051764	200051865	102	1	0	-1	Accurate
HsInv0033	chr17	40566233	40567384	1152	40565329-40566222	40567385-40567635	40566223	40567384	1162	-10	0	10	Imprecise
HsInv0035	chr17	57999778	58000250	473	57999876-57999896	58000197-58000526	57999897	58000196	300	119	-54	-173	Imprecise
HsInv0078	chrX	53868900	53870123	1224	53868477-53868762	53870313-53870595	53868763	53870312	1550	-137	189	326	Imprecise
HsInv0082	chrX	96081284	96082949	1666	96081228-96081284	96082949-96083089	96081285	96082948	1664	1	-1	-2	Accurate
HsInv0003	chr1	185733101	185733350	250	185733099-185733101	185733353-185733355	185733102	185733352	251	1	2	1	Accurate
HsInv0004	chr1	196023846	196024607	762	196023411-196023414	196024607-196024608	196023415	196024606	1192	-431	-1	430	Imprecise
HsInv0006	chr1	203445294	203445397	104	203445230-203445294	203445378-203445457	203445295	203445377	83	1	-20	-21	Imprecise
HsInv0030	chr16	73797599	73814159	16561	73797041-73797599	73814160-73814718	73797600	73814159	16560	1	0	-1	Accurate
HsInv0031	chr16	83746237	83747302	1066	83746215-83746259	83747296-83747305	83746260	83747295	1036	23	-7	-30	Imprecise
HsInv0036	chr18	12134389	12137214	2826	12131460-12135081	12136523-12140144	12135082	12136522	1441	693	-692	-1385	Imprecise
HsInv0040	chr2	138721419	138725673	4255	138720715-138721469	138725059-138725814	138721470	138725058	3589	51	-615	-666	Imprecise
HsInv0041	chr2	225001224	225001326	103	225001193-225001195	225001332-225001334	225001196	225001331	136	-28	5	33	Imprecise
HsInv0045	chr21	26942554	26943508	955	26942303-26942558	26943499-26943755	26942559	26943498	940	5	-10	-15	Imprecise
HsInv0052	chr3	164028056	164030335	2280	164028056-164028057	164030337-164030338	164028058	164030336	2279	2	1	-1	Accurate
HsInv0055	chr5	63800180	63811001	10822	63797584-63802465	63808718-63813599	63802466	63808717	6252	2286	-2284	-4570	Imprecise
HsInv0058	chr6	31117201	31118074	874	31117199-31117201	31118075-31118075	31117202	31118074	873	1	0	-1	Accurate
HsInv0059	chr6	89980353	89980661	309	89980347-89980353	89980662-89980668	89980354	89980661	308	1	0	-1	Accurate
HsInv0061	chr6	107275899	107277573	1675	107275246-107275899	107277574-107278229	107275900	107277573	1674	1	0	-1	Accurate
HsInv0063	chr7	70064121	70076815	12695	70064121-70064122	70076816-70076823	70064123	70076815	12693	2	0	-2	Accurate
HsInv0068	chr9	76087960	76088209	250	76087959-76087960	76088210-76088213	76087961	76088209	249	1	0	-1	Accurate
HsInv0069	chr9	114913782	114914991	1210	114907364-114913787	114914988-114921411	114913788	114914987	1200	6	-4	-10	Imprecise
HsInv0072	chrX	45433293	45435559	2267	45432027-45433183	45435670-45436826	45433184	45435669	2486	-109	110	219	Imprecise
HsInv0012	chr1	246749189	246758249	9061	246748189-246750189	246757249-246759249	246750190	246757248	7059	1001	-1001	-2002	Imprecise
HsInv0014	chr10	4994874	5036786	41913	5015247-5015248	5016554-5016555	5015249	5016553	1305	20375	-20233	-40608	Imprecise
HsInv0015	chr10	13145105	13145278	174	13145044-13145074	13145275-13145283	13145075	13145274	200	-30	-4	26	Imprecise
HsInv0016	chr10	58927667	58927951	285	58926947-58927668	58927953-58927989	58927669	58927952	284	2	1	-1	Accurate
HsInv0017	chr10	93191655	93197436	5782	93190197-93193232	93195859-93198894	93193233	93195858	2626	1578	-1578	-3156	Imprecise
HsInv0020	chr11	61619532	61624597	5066	61619145-61619444	61624686-61624985	61619445	61624685	5241	-87	88	175	Imprecise
HsInv0032	chr17	5826739	5827291	553	5826333-5826588	5827443-5827698	5826589	5827442	854	-150	151	301	Imprecise
HsInv0051	chr3	57362070	57362262	193	57361931-57361932	57362346-57362347	57361933	57362345	413	-137	83	220	Imprecise
HsInv0060	chr6	94797796	94798133	338	94797796-94797796	94798134-94798141	94797797	94798133	337	1	0	-1	Accurate
HsInv0076	chrX	51456609	51460366	3758	51427692-51452745	51460138-51485162	51452746	51460137	7392	-3863	-229	3634	Imprecise
HsInv0081	chrX	78809884	78810019	136	78808544-78809888	78810016-78812735	78809889	78810015	127	5	-4	-9	Imprecise

Table S4. Genotypes for the experimentally analyzed inversions in a 10-individual panel. Origin of samples: CEU (NA12156 and NA12878), YRI (NA18507, NA18517, NA19129 and NA19240), CHB (NA18555), JPT (NA18956) and unknown (NA15510).

Inversion	NA12156	NA12878	NA15510	NA18507	NA18517	NA18555	NA18956	NA19129	NA19240	HuRef	Method	Result
Polymorphic candidates by PEM												
HsInv0003	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	PCR	Polymorphic
HsInv0004	Std/Inv	Std/Inv	Std/Inv	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	Std/Inv	PCR	Polymorphic
HsInv0006	Std/Inv	Inv/Inv	Std/Inv	Std/Std	Std/Std	Inv/Inv	Inv/Inv	Std/Inv	Std/Std	Inv/Inv	PCR	Polymorphic
HsInv0030	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	PCR	Polymorphic
HsInv0031	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	iPCR-EcoRI	Polymorphic
HsInv0036	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Long-range PCR	Polymorphic
HsInv0041	Std/Inv	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Std/Std	Std/Inv	Std/Inv	Std/Std	Inv/Inv	PCR	Polymorphic
HsInv0045	Std/Std	Std/Std	Std/Std	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	iPCR-SacI	Polymorphic
HsInv0052	Std/Inv	Inv/Del	Del/Del	Std/Del	Std/Del	Del/Del	Del/Del	Std/Std	Std/Del	Inv/Inv	iPCR-HindIII	Polymorphic
HsInv0055	Std/Std	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Inv/Inv	iPCR-BamHI	Polymorphic
HsInv0058	Std/Inv	Inv/Inv	Std/Std	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	PCR	Polymorphic
HsInv0059	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Std	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	PCR	Polymorphic
HsInv0063	Std/Inv	Inv/Inv	Std/Inv	Std/Std	Std/Std	Inv/Inv	Inv/Inv	Std/Std	Std/Std	Inv/Inv	PCR	Polymorphic
HsInv0068	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	PCR	Polymorphic
HsInv0069	Std/Inv	Inv/Inv	Std/Std	Std/Std	Std/Inv	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	iPCR-NsiI	Polymorphic
HsInv0072	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-	Inv/Inv	Inv/Inv	Inv/Inv	Std/Std	Inv/Inv	Inv/-	iPCR-HindIII	Polymorphic
HG18 error candidates by PEM												
HsInv0021	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	PCR	HG18 error
HsInv0022	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	PCR	HG18 error
HsInv0023	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	iPCR-NsiI	HG18 error
HsInv0024	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	iPCR-NsiI	HG18 error
HsInv0029	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	-	Inv/Inv	Inv/Inv	-	iPCR-EcoRI	HG18 error
HsInv0038	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	iPCR-NsiI	HG18 error
HsInv0040	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Std	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	iPCR-HindIII	Polymorphic
HsInv0044	-	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	PCR	HG18 error
HsInv0046	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	iPCR-NsiI	HG18 error
HsInv0049	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	PCR	HG18 error
HsInv0050	-	-	-	-	-	-	-	-	-	-	Corrected in HG19 + iPCR/PCR	HG18 error
HsInv0053	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	iPCR-NsiI	HG18 error
HsInv0061	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	iPCR-HindIII	Polymorphic
HsInv0062	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	PCR	HG18 error
HsInv0066	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	PCR	HG18 error
HsInv0067	Inv/Inv	Inv/Inv	Inv/Inv	-	-	-	-	-	-	-	iPCR-ApaI	HG18 error
HsInv0073	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-	PCR	HG18 error
HsInv0074	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-	PCR	HG18 error
HsInv0086	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-	-	-	-	Inv	-	Inv/-	PCR	HG18 error
HsInv0090	-	-	-	Inv/-	-	-	-	-	-	Inv/-	iPCR-NsiI	HG18 error
HuRef error candidates by PEM												
HsInv0005	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	PCR	HuRef error
HsInv0033	-	-	-	-	-	-	-	-	-	Std/Std	PCR	HuRef error
HsInv0035	-	-	-	-	-	-	-	-	-	Std/Std	PCR	HuRef error
HsInv0078	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	Std/Std	PCR	HuRef error
HsInv0082	-	-	-	-	-	-	-	-	-	Std/Std	PCR	HuRef error

Table S5. HG18 errors.

Inversion	Chr.	BP1	BP2	Size (bp)	HG18 Clone	Experiment	Sample	PCR products tested	Genotype	Corrected
Hsinv0001	chr1	2473962-2475133	2489145-2490316	14011	RP3-395M20	-	-	-	-	HG19+HG38
Hsinv0002	chr1	26839407-26840265	26845951-26846809	5685	RP4-785J11	Long-range PCR	BAC clone	AB/AC	*Inv*	No
Hsinv0019	chr11	49691860-49692868	49706585-49707593	13716	RP11-707M1	Long-range PCR	BAC clone	AB/BD	*Inv*	No
Hsinv0021	chr12	12436093-12436153	12437832-12437892	1678	RP11-253I19	PCR	BAC clone	AB/AC	*Inv*	No
Hsinv0022	chr12	13405560-13437631	13441214-13475149	3582	RP11-18J10	PCR	BAC clone	CD/BD	*Inv*	No
Hsinv0023	chr12	17813889-17814515	17903823-17904449	89307	RP11-633O13	iPCR-NsiI	BAC clone	AB/AC	*Inv*	No
Hsinv0024	chr12	79369050-79370199	79382018-79383167	11818	RP11-288D9	iPCR-NsiI	BAC clone	AB/AC	*Inv*	No
Hsinv0025	chr12	85763520-85764349	85777075-85777904	12725	RP11-27M21	PCR	BAC clone	CD/BD	*Inv*	No
Hsinv0029	chr16	1221550-1228054	1238397-1244871	10342	RP11-616M22	iPCR-EcoRI	BAC clone	CD/BD	*Inv*	No
Hsinv0038	chr19	43954882-43956117	43971562-43972797	15444	CTD-2540F13	iPCR-NsiI	BAC clone	AB/AC	*Inv*	No
Hsinv0042	chr2	234134828-234135583	234151755-234152510	16171	RP11-289A15	-	-	-	-	HG19+HG38
Hsinv0044	chr21	26296023-26296029	26296571-26296577	541	CMP21-S491	PCR	P1 plasmid clone	CD/BD	*Inv*	No
Hsinv0046	chr21	40317223-40317736	40331906-40332419	14169	RP1-31P10	iPCR-NsiI	BAC clone	AB/AC	*Inv*	No
Hsinv0049	chr3	44716234-44716235	44717052-44717053	816	RP11-348P10	PCR	BAC clone	AB/AC	*Inv*	No
Hsinv0050	chr3	50899680-50900357	50910049-50910726	9691	RP11-73I17	iPCR/Southern	BAC clone		*Inv*	HG19+HG38
Hsinv0053	chr3	188614227-188617860	188625665-188629298	7804	RP11-560F18	iPCR-NsiI	BAC clone	CD/BD	*Inv*	No
Hsinv0057	chr5	178993587-178996665	179015092-179018173	18426	RP11-1379J22	iPCR-KpnI	BAC clone	CD/BD	*Inv*	No
Hsinv0062	chr6	168834983-168835529	168836602-168837148	1072	RP1-125N5	PCR	BAC clone	AB/AC	*Inv*	HG38
Hsinv0064	chr7	106845789-106846308	106850372-106850891	4063	CTB-20D2	PCR	BAC clone	AB/AC	*Inv*	No
Hsinv0066	chr8	6141940-6142692	6144797-6145549	2104	RP11-1I11	PCR	BAC clone	AB/BD	*Inv*	No
Hsinv0067	chr8	48376601-48377415	48393421-48394235	16005	RP11-113O13	iPCR-ApaI	BAC clone	AB/AC	*Inv*	HG38
Hsinv0073	chrX	46695638-46695777	46715616-46715725	19838	RP1-306D1	PCR	BAC clone	CD/BD	*Inv*	HG38
Hsinv0074	chrX	48900144-48900536	48906269-48906661	5732	LL0XNC01-11 8A22* LL0XNC01-6 M132*	PCR	HapMap PT01	CD/BD	Inv (92 chromosomes)	HG38
Hsinv0086	chrX	149321613-149321673	149336280-149336340	14606	A121971*	PCR	HapMap PT01	AB/AC	Inv (92 chromosomes)	HG38
Hsinv0090	chrY	21622786-21623776	21636305-21637295	12528	RP11-65G9	iPCR-NsiI	BAC clone	CD/BD	*Inv*	No

[1] Cosmid clone
[2] Number of independent chromosomes analyzed
* Not available

Table S6. Orientation of polymorphic inversions regions in chimpanzee (Pan troglodytes), gorilla (Gorilla gorilla), and Rhesus macaque (Macaca mulatta) genomes. The results from both genome analysis (genome assemblies are indicated in the column heading) and genotyping experiments are shown. The ancestral allele is indicated in the last column. The total number of successful experimental genotyping in each non-human primate species can be found in the Total line.

	Genome analysis			Experimental analysis		
Inversion	**Chimp (PanTro4)**	**Gorilla (GorGor3)**	**Rhesus (RheMac3)**	**Chimp**	**Gorilla**	**Ancestral**
Inversions with IRs						
HsInv0030	Inv	Std	Inv	Inv	Inv	Inv
HsInv0031	Inv	Inv	ND	Inv	Inv	Inv
HsInv0040	Inv	Inv	Inv	Inv	Inv	Inv
HsInv0045*	ND	ND	Std	ND	Std	Std
HsInv0055	ND	ND	Inv	Inv	Std	ND
HsInv0061	Inv	Inv	Inv	Inv	Inv	Inv
HsInv0069	ND	Std	ND	ND	Inv	Inv
HsInv0072	Std	Std	ND	Inv	Inv	Inv
Inversions without IRs						
HsInv0003	Std	Std	Std	Std	Std	Std
HsInv0004	Inv	Inv	Inv	Inv	Inv	Inv
HsInv0006	Inv	Inv	Inv	Inv	Inv	Inv
HsInv0041	Std	Std	Std	Std	Std	Std
HsInv0052	Std	Std	ND	Std	Std	Std
HsInv0058	Inv	Inv	Inv	Inv	Inv	Inv
HsInv0059	Inv	Inv	Inv	Inv	Inv	Inv
HsInv0063	Std	Std	Std	Std	Std	Std
HsInv0068	Inv	Inv	Inv	Inv	Inv	Inv
Total				15	17	

* Whole inversion region deleted in available chimpanzee genome.

Table S7. Genotypes of 17 validated inversions for 90 HapMap individuals of European origin. Genotyped DNAs correspond to CEU samples from HapMap Plate 01. Individuals are grouped by parent-child trios.

Individual	Sex	Family	Relationship	HsInv0003	HsInv0004	HsInv0006	HsInv0030[1]	HsInv0031	HsInv0040	HsInv0041	HsInv0045	HsInv0052[2]
NA06985	Female	1341	maternal grandmother	Std/Inv	Std/Inv	Std/Inv	Std/InvDel	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Del
NA06991	Female	1341	mother	Inv/Inv	Inv/Inv	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Del
NA06993	Male	1341	maternal grandfather	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Std	Inv/Inv	Inv/Inv	Std/Inv	Std/Del
NA07034	Male	1341	paternal grandfather	Std/Inv	Std/Std	Std/Inv	Inv/Inv	ND	Inv/Inv	Std/Inv	Std/Inv	Std/Del
NA07048*	Male	1341	father	Std/Inv	Std/Std	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Std	Std/Inv	Std/Del
NA07055	Female	1341	paternal grandmother	Inv/Inv	Std/Std	Std/Inv	Inv/InvDel	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Del
NA06994*	Male	1340	paternal grandfather	Std/Std	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Del
NA07000*	Female	1340	paternal grandmother	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Std	Std/Del
NA07029	Male	1340	father	Std/Inv	Std/Inv	Std/Std	Std/Inv	Std/Inv	ND	Std/Std	Std/Std	Inv/Del
NA07019	Female	1340	mother	Std/Inv	Std/Std	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Del
NA07022	Male	1340	maternal grandfather	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Std	Std/Inv	Std/Inv	Inv/Inv	Std/Del
NA07056*	Female	1340	maternal grandmother	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Std	Del/Del
NA07345	Female	1345	maternal grandmother	Std/Std	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Inv/Inv	Std/Std	Inv/Inv	Inv/Del
NA07348	Female	1345	mother	Std/Inv	Std/Inv	Std/Std	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Std/Inv	Inv/Del
NA07357*	Male	1345	maternal grandfather	Inv/Inv	Std/Inv	Std/Std	Inv/Inv	Std/Std	Std/Inv	Std/Inv	Std/Inv	Del/Del
NA10830	Male	1408	father	Inv/Inv	Std/Std	Inv/Inv	Inv/Inv	Std/Std	Inv/Inv	Std/Std	Std/Std	Std/Del
NA12154*	Male	1408	paternal grandfather	Std/Inv	Std/Std	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Std/Std	Std/Del
NA12236	Female	1408	paternal grandmother	Std/Inv	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Del
NA10831	Female	1408	mother	Inv/Inv	Std/Inv	Std/Std	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Std	Inv/Inv
NA12155*	Male	1408	maternal grandfather	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Inv/Inv	Std/Std	Std/Std	Std/Std	Std/Inv
NA12156	Female	1408	maternal grandmother	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	ND	Std/Inv	Std/Std	Std/Inv
NA10835	Male	1416	father	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Del
NA12248	Male	1416	paternal grandfather	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Std	Inv/Inv
NA12249*	Female	1416	paternal grandmother	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Del
NA10838	Male	1420	father	Std/Inv	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Std	Std/Inv	Std/Del
NA12003*	Male	1420	paternal grandfather	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Std/Inv	Del/Del
NA12004*	Female	1420	paternal grandmother	Std/Std	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Std
NA10839	Female	1420	mother	Inv/Inv	Std/Std	Std/Std	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Std/Inv	Inv/Del
NA12005	Male	1420	maternal grandfather	Inv/Inv	Std/Inv	Std/Std	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Del/Del
NA12006*	Female	1420	maternal grandmother	Inv/Inv	Std/Std	Std/Std	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Std/Std	Inv/Del
NA10846	Male	1334	father	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Std	Std/Inv	Std/Inv
NA12144*	Male	1334	paternal grandfather	Std/Inv	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Std	Std/Inv	Std/Inv
NA12145	Female	1334	paternal grandmother	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Std	Inv/Inv	Std/Std	Std/Inv	Inv/Inv
NA10847*	Female	1334	mother	Inv/Inv	Std/Std	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv
NA12146	Male	1334	maternal grandfather	Std/Inv	Std/Std	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv
NA12239	Female	1334	maternal grandmother	Inv/Inv	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Del
NA10851*	Male	1344	father	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Del/Del
NA12056	Male	1344	paternal grandfather	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Std	Inv/Del
NA12057	Female	1344	paternal grandmother	Inv/Inv	Std/Std	Std/Inv	Inv/InvDel	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Del/Del
NA10854	Female	1349	mother	Inv/Inv	Std/Std	Std/Std	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Std	Std/Inv
NA11839	Male	1349	maternal grandfather	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Std	Inv/Inv
NA11840	Female	1349	maternal grandmother	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Del
NA10856	Male	1350	father	Inv/Inv	Std/Inv	Inv/Inv	Inv/InvDel	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv
NA11829*	Male	1350	paternal grandfather	Inv/Inv	Std/Inv	Inv/Inv	Inv/InvDel	Std/Inv	Inv/Inv	Inv/Inv	Std/Std	Std/Std
NA11830*	Female	1350	paternal grandmother	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Inv
NA10855	Female	1350	mother	Std/Inv	Std/Std	Std/Std	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Std	Std/Inv
NA11831*	Male	1350	maternal grandfather	Inv/Inv	Std/Std	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Std	Std/Std	Inv/Inv
NA11832	Female	1350	maternal grandmother	Std/Inv	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Del
NA10857	Male	1346	father	Inv/Inv	Std/Std	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Std	Std/Inv	Std/Inv
NA12043*	Male	1346	paternal grandfather	Inv/Inv	Std/Std	Inv/Inv	Inv/InvDel	Std/Std	Std/Inv	Std/Inv	Std/Inv	Std/Std
NA12044*	Female	1346	paternal grandmother	Inv/Inv	Std/Std	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Std	Inv/Inv	Inv/Inv
NA10859	Female	1347	mother	Inv/Inv	Std/Std	Inv/Inv	Inv/Inv	Std/Inv	Std/Std	Inv/Inv	Std/Inv	Std/Inv
NA11881	Male	1347	maternal grandfather	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Del
NA11882	Female	1347	maternal grandmother	Inv/Inv	Std/Std	Inv/Inv	Inv/InvDel	Std/Inv	Std/Std	Std/Inv	Std/Std	Inv/Inv
NA10860	Male	1362	father	Std/Inv	Std/Std	Inv/Inv	Inv/InvDel	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Inv/Inv
NA11992*	Male	1362	paternal grandfather	Std/Std	Std/Std	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv
NA11993*	Female	1362	paternal grandmother	Inv/Inv	Std/Inv	Std/Inv	Inv/InvDel	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv
NA10861	Female	1362	mother	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Std	Std/Std	Inv/Del
NA11994*	Male	1362	maternal grandfather	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Del
NA11995*	Female	1362	maternal grandmother	Inv/Inv	Std/Std	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv
NA10863	Female	1375	mother	Std/Inv	Std/Std	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Del
NA12234	Female	1375	maternal grandmother	Std/Inv	Std/Std	Std/Std	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Del
NA12264	Male	1375	maternal grandfather	Std/Inv	Std/Std	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Std/Std	Inv/Inv	Del/Del
NA12707	Male	1358	father	Inv/Inv	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Std	Inv/Inv
NA12716*	Male	1358	paternal grandfather	Std/Inv	Std/Std	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Del
NA12717*	Female	1358	paternal grandmother	Inv/Inv	Std/Std	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Inv
NA12740	Female	1444	mother	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Std	Std/Std	Std/Inv	Std/Std	Del/Del
NA12750*	Male	1444	maternal grandfather	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Del/Del
NA12751*	Female	1444	maternal grandmother	Std/Inv	Std/Inv	Std/Inv	Inv/InvDel	Std/Inv	Std/Inv	Std/Inv	Std/Std	Inv/Del
NA12752	Male	1447	father	Std/Std	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Std	Inv/Inv	Del/Del
NA12760	Male	1447	paternal grandfather	Std/Std	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Del
NA12761*	Female	1447	paternal grandmother	Std/Inv	Std/Std	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Std	Std/Inv	Std/Del
NA12753	Female	1447	mother	Inv/Inv	Std/Std	Std/Inv	Inv/InvDel	Std/Std	Inv/Inv	Std/Inv	Std/Std	Std/Inv
NA12762	Male	1447	maternal grandfather	Inv/Inv	Std/Std	Std/Std	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Std
NA12763*	Female	1447	maternal grandmother	Std/Inv	Std/Std	Inv/Inv	Inv/InvDel	Std/Inv	Inv/Inv	Std/Std	Std/Inv	Inv/Inv
NA12801	Male	1454	father	Inv/Inv	Std/Std	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Std	Std/Inv	Del/Del
NA12812*	Male	1454	paternal grandfather	Std/Inv	Std/Std	Std/Inv	Std/InvDel	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/Del
NA12813	Female	1454	paternal grandmother	Std/Inv	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Std	Std/Std	Std/Del
NA12802	Female	1454	mother	Inv/Inv	Std/Std	Std/Std	Std/Inv	Inv/Inv	Inv/Inv	Std/Std	Std/Inv	Std/Del
NA12814*	Male	1454	maternal grandfather	Inv/Inv	Std/Std	Std/Std	Std/Std	Inv/Inv	Inv/Inv	Std/Std	Std/Std	Std/Del
NA12815*	Female	1454	maternal grandmother	Inv/Inv	Std/Std	Std/Std	Inv/Inv	Inv/Inv	Std/Inv	Std/Std	Std/Inv	Std/Del
NA12864	Male	1459	father	Std/Inv	Std/Std	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv
NA12872*	Male	1459	paternal grandfather	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Del
NA12873*	Female	1459	paternal grandmother	Std/Inv	Std/Std	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Std
NA12865	Female	1459	mother	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/Del
NA12874*	Male	1459	maternal grandfather	Inv/Inv	Std/Std	Inv/Inv	Inv/InvDel	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Del/Del
NA12875	Female	1459	maternal grandmother	Inv/Inv	Std/Std	Std/Std	Inv/InvDel	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/Del
NA12878	Female	1463	mother	Std/Inv	Std/Inv	Inv/Inv	Std/InvDel	Inv/Inv	Inv/Inv	Std/Std	Std/Std	Inv/Del
NA12891	Male	1463	maternal grandfather	Std/Inv	Std/Std	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Std	Std/Inv	Del/Del
NA12892	Female	1463	maternal grandmother	Inv/Inv	Inv/Inv	Std/Inv	InvDel/InvDel	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Del

Individual	Sex	Family	Relationship	HsInv0055[3]	HsInv0058	HsInv0059	HsInv0061	HsInv0063	HsInv0068	HsInv0069	HsInv0072[4]
NA06985	Female	1341	maternal grandmother	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv
NA06991	Female	1341	mother	Std/Inv	Std/Std	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv
NA06993	Male	1341	maternal grandfather	Std/Inv	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/-
NA07034	Male	1341	paternal grandfather	Std/Std	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/-
NA07048*	Male	1341	father	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/-
NA07055	Female	1341	paternal grandmother	ND	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv
NA06994*	Male	1340	paternal grandfather	Inv/Inv	Std/Inv	Std/Std	Inv/Inv	Inv/Inv	Std/Std	Std/Inv	Inv/-
NA07000*	Female	1340	paternal grandmother	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv
NA07029	Male	1340	father	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/-
NA07019	Female	1340	mother	ND	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv
NA07022	Male	1340	maternal grandfather	Std/Inv	ND	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/-
NA07056*	Female	1340	maternal grandmother	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv
NA07345	Female	1345	maternal grandmother	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Inv/Inv
NA07348	Female	1345	mother	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Std	Inv/Inv
NA07357*	Male	1345	maternal grandfather	Inv/Inv	ND	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/-
NA10830	Male	1408	father	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/-
NA12154*	Male	1408	paternal grandfather	Inv/Inv	Std/Inv	Std/Std	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/-
NA12236	Female	1408	paternal grandmother	ND	Std/Inv	Inv/Inv	Inv/Inv	Std/Std	Std/Inv	Std/Inv	Inv/Inv
NA10831	Female	1408	mother	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv
NA12155*	Male	1408	maternal grandfather	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/-
NA12156	Female	1408	maternal grandmother	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv
NA10835	Male	1416	father	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/-
NA12248	Male	1416	paternal grandfather	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Std	Std/Inv	Inv/-
NA12249*	Female	1416	paternal grandmother	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Inv/Inv
NA10838	Male	1420	father	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/-
NA12003*	Male	1420	paternal grandfather	ND	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/-
NA12004*	Female	1420	paternal grandmother	Std/Inv	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Std	Inv/Inv
NA10839	Female	1420	mother	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv
NA12005	Male	1420	maternal grandfather	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Std	Inv/-
NA12006*	Female	1420	maternal grandmother	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv
NA10846	Male	1334	father	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/-
NA12144*	Male	1334	paternal grandfather	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/-
NA12145	Female	1334	paternal grandmother	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv
NA10847*	Female	1334	mother	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv
NA12146	Male	1334	maternal grandfather	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/-
NA12239	Female	1334	maternal grandmother	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv
NA10851*	Male	1344	father	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/-
NA12056	Male	1344	paternal grandfather	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/-
NA12057	Female	1344	paternal grandmother	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Inv
NA10854	Female	1349	mother	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv
NA11839	Male	1349	maternal grandfather	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/-
NA11840	Female	1349	maternal grandmother	ND	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv
NA10856	Male	1350	father	ND	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/-
NA11829*	Male	1350	paternal grandfather	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-
NA11830*	Female	1350	paternal grandmother	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv
NA10855	Female	1350	mother	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Inv
NA11831*	Male	1350	maternal grandfather	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/-
NA11832	Female	1350	maternal grandmother	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Inv
NA10857	Male	1346	father	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Std	Inv/Inv	Inv/Inv	Inv/-
NA12043*	Male	1346	paternal grandfather	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Std/Inv	Inv/Inv	Inv/-
NA12044*	Female	1346	paternal grandmother	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv
NA10859	Female	1347	mother	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Std	Inv/Inv
NA11881	Male	1347	maternal grandfather	Inv/Inv	Std/Std	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/-
NA11882	Female	1347	maternal grandmother	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv
NA10860	Male	1362	father	Std/Std	Std/Std	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/-
NA11992*	Male	1362	paternal grandfather	Std/Inv	Std/Inv	Std/Std	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/-
NA11993*	Female	1362	paternal grandmother	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv
NA10861	Female	1362	mother	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv
NA11994*	Male	1362	maternal grandfather	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/-
NA11995*	Female	1362	maternal grandmother	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv
NA10863	Female	1375	mother	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv
NA12234	Female	1375	maternal grandmother	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/Inv
NA12264	Male	1375	maternal grandfather	Inv/Inv	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-
NA12707	Male	1358	father	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Inv/-
NA12716*	Male	1358	paternal grandfather	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Std	Inv/-
NA12717*	Female	1358	paternal grandmother	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Inv/Inv
NA12740	Female	1444	mother	Std/Std	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv
NA12750*	Male	1444	maternal grandfather	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-
NA12751*	Female	1444	maternal grandmother	Std/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Std	Std/Inv	Std/Inv	Inv/Inv
NA12752	Male	1447	father	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-
NA12760	Male	1447	paternal grandfather	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	ND	Inv/Inv	Std/Inv	Inv/-
NA12761*	Female	1447	paternal grandmother	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv
NA12753	Female	1447	mother	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv
NA12762	Male	1447	maternal grandfather	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-
NA12763*	Female	1447	maternal grandmother	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Std	Inv/Inv	Std/Inv	Inv/Inv
NA12801	Male	1454	father	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Std	Inv/-
NA12812*	Male	1454	paternal grandfather	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	ND	Inv/Inv	Std/Inv	Inv/-
NA12813	Female	1454	paternal grandmother	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv
NA12802	Female	1454	mother	Inv/Inv	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv
NA12814*	Male	1454	maternal grandfather	Inv/Inv	Std/Std	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-
NA12815*	Female	1454	maternal grandmother	Std/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv
NA12864	Male	1459	father	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-
NA12872*	Male	1459	paternal grandfather	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/-
NA12873*	Female	1459	paternal grandmother	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv
NA12865	Female	1459	mother	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Std/Inv	Std/Inv	Inv/Inv
NA12874*	Male	1459	maternal grandfather	Std/Inv	Inv/Inv	Std/Inv	Inv/Inv	Std/Inv	Std/Inv	Std/Inv	Inv/-
NA12875	Female	1459	maternal grandmother	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv
NA12878	Female	1463	mother	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv
NA12891	Male	1463	maternal grandfather	Inv/Inv	Std/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	Inv/-
NA12892	Female	1463	maternal grandmother	Inv/Inv	Inv/Inv	Inv/Inv	Inv/Inv	ND	Std/Inv	Std/Inv	Inv/Inv

[*] Individuals in the 1000GP used for haplotype analysis.
[1] InvDel corresponds to an allele carrying a 574-bp deletion in SD1 that occurred in the ancestral arrangement (Inv in this case since HG18 carries the derived STD allele) independently of the inversion itself.
[2] Some individuals carry a polymorphic 114.2-kb deletion that removes the whole region including the inverted segment. This deletion has been genotyped with primers flanking the deletion point that produce specific PCR products confirming the absence or presence of this sequence. Inversion status can only be genotyped in those individuals that do not have the deletion.
[3] SNP polymorphisms rs2591414 generates in some chromosomes a new BamHI restriction site that changes the size of PCR products containing this region (55 bp shorter than expected according to HG18 sequence). SNP polymorphism rs6894002 eliminates BamHI site in region D in some chromosomes. An additional primer D2 close to the next BamHI site (6,1 kb downstream) was added to the multiplex PCR to allow amplification in these cases.
[4] Inversion located in the X chromosome.

Table S8. Breakpoint library used by BreakSeq algorithm to identify reads from each arrangement. In red are shown the breakpoint sequences. When the breakpoint is located between two nucleotide positions it is indicated by an arrow. Blue sequences denote those nucleotides deleted in the alternative allele. A positive sign in the Indel column indicates the presence of a sequence fragment that is absent from the probes denoted with a negative sign.

Locus	Allele	Breakpoint	Indel	Sequence
HsInv0003	Inv	BP1	-	TAGCTTAATTTACAGAGAGCTGACATTGTCACAATATTGAGTATTACATT↓AATATGCTTT TGTTTTTGGGTGTCACTCTCAATTTTTTTGAAACAATAC
	Inv	BP2	-	TTCAGCCTAATTTTTTAGATATTACTACCAATAGAATGGTCTCCTGGATTTCCTTGTGAGAGAGATAGTCTTAATATTACTAACTTCAAGAATGAGAGAA
	Std	BP1	+	AATGTGTTTTTTAATGAATGGAAAGATAAGAAATAGCAAAGATATTGAGGAAATCCAGGAGACCATTCTATTGGTAGTAATATCTAAAAAATTAGGCTGA
	Std	BP2	-	ATTGTTTCAAAAAAATTGAGAGTGACACCCAAAAACAGAAAGCATATTTACCTTGTGAGAGAGATAGTCTTAATATTACTAACTTCAAGAATGAGAGAAA
HsInv0006*	Inv	BP1a	-	TTTCTGGACGTAAGATGTCAAAGACACAACTTCAGAAAAACTAGTCATCCCCCTCCTTTTATGTTACAAGGTAAAATAAGGCTTTTTAAATTCCTATCCA
	Inv	BP1b	-	TACAAGTCCATTCCAGGACAGTATATTTCTGGACGTAAGATGTCAAAGACACAACTTCAGAAAAACTAGTCATCCCCCTCCTTTTATGTTACAAGGTAAA
	Inv	BP2a	-	TTTTAAATTCCTATCCAAACCAACTAAATGGAACGGTATCATTTAGGACCCTGTCATGATCTTAGTATCTTATTTTTATCACCATTTCTGGATTCCCAGT
	Inv	BP2b	-	TTTAGGACCCTGTCATGATCTTAGTATCTTATTTTTATCACCATTTCTGGATTCCCAGTCGGATGTAGGATATTAAGCGCAAAACTTTGGGAATATAAGA
	Std	BP1	+	ATTTCTGGACGTAAGATGTCAAAGACTAGTTTCAGAAAAACTAGTCATCCCCAGAAATGGTGATAAAAAATAAGATACTAAGATCATGACAGGGTCCTAAA
	Std	BP2	+	TAAAAGGAGGGGGATGACTAGTTTTTCTGAAACTACTAGTTTTTCTGATACTGTCATGATCTTAGTATCTTATTTTTATCACCATTTCTGGATTCCCAGT
HsInv0041	Inv	BP1	-	CAAGTCATAATGTAAGGATCTGGGGAGGAATTAGGTGCTGATAAGAAGCGTCTCAAAAAAAAAAAACAAAAAAAAGACAGACAGAGAGAATGAATGAATGAA
	Inv	BP2	-	AATGAATGTGCTGGAAGTGAGATAGGATAAAAACAGCCATCAGAGCCGGAC↓CACCCTTCTCCTCCCTGGTCTGTGCCCTGAGACTGGCCTCTTGCCCTTTC
	Std	BP1	+	AAGTCATAATATAAGGATCTGGGGAGGAATTAGGTGCTGATAAGAAGCGTGCTGTGTGGGCACTTTGCTTAGGAGGGCAGTCCGGCTCTGATGGCTGTCT
	Std	BP2	+	tcattcattcattcTCTCTGTCTGTCttttttgttttttttttgagacggagtgtcgctctgtcgcccaggctggagtgcagtggcgctatctcgg
HsInv0058	Inv	BP1	+	CTTGAGCCTGGGAGGTTGCAGTGAGTCAAGATTGTGCCACTGCACTCCAGCCTGGGCAACAGAGCAAGACCCTGTCTCAAAAAAACAGAAAACAGGCCGG
	Inv	BP2	+	GTGACCAGCGTGAATCACTGTGTATAGAATGGGAAGCGATGGGCGGCACACCATTCCATACCATTCCCACCAGGCAGATGTGATAGGCGTACGTAACTCC
	Std	BP1	-	aaggatttagtacaagaagtAGATGTTTATAAAACCACTCGAAAAGGTGGTGTGCCGCCCATTGCTTCCCATTCTATACACAGTGATTCACGCTGGTCAC
	Std	BP2	-	ccggcctgtttctgtttttgagacagggtcttgctctgttgcccaggc↓CCTTGAACTATTCCACTTCCCACTCCTCCCCAGGCCCTTCCTAATGGGA
HsInv0059	Inv	BP1	NA	TTTTGGTAAATTTAAATAATAGTTATATGTGTCTTTCATAACTAAAGAAATGAAAGCTTTTAAAAGTAAAATTACAAAGAAATTTTATTTCATTCTCTAA
	Inv	BP2	+	TACAATAAAGTTGCAGATTCAATTCAGCATGCAAAAATCAGAAGTGTTTCTGTACACTAGCAACAAACTACCTAAAAAAAAGAAATTTAAAAAAAATCCAAT
	Std	BP1	NA	ttttggtaaatttaaataatagttatatgtgtctttcataactaaagaaacagaaacacttctgatttttgcatgctgaattgaatctgcaactttat g
	Std	BP2	-	aTGAAATAAAATTTCTTTGTAATTTTACTTTTAAAAGCTTTCatttcttttgatatacaagaatcaactcaaaatggattaaagacttaagtgtaagacc
HsInv0063	Inv	BP1	-	TGCTAAAGAGAAAGTTAAGCTTGGAAAACTGATTCAGTAAAACTACCTTCAGCTAAGTTAAAAAAAAAAAAAATTTATAGCGATGAGACCTCACTGTGTTA
	Inv	BP2	NA	TGCCCATAATTCTTGACTGTCTGTTTCCCTTTGGTATAACTGGCCTTCAGTTAGCTGAGTGGTGCTTGCCTGTGGTCCCAGCTACTTGGGAGGCTGAAGC
	Std	BP1	+	tcccttacagcgtgtctgcacttgaaaaaagctaaaggaagtcctttatactgaaggccagttataccaaagggaaacagacagtcaagaattatgggca
	Std	BP2	NA	cctgggtaacacagtgaggtctcatcactataaatttttttttttaacttagctgagtggtgcttgcctgtggtcccagctacttgggaggctgaagc
HsInv0068	Inv	BP1	+	AGCAATGAGTACACACGGACATGCAGAATAGAATAATAGACATTGCAGGCTACAGTGGGCAGGAGGGTGAGAGAGGGGTGAGGGGTGAAATGCTATTGAT
	Inv	BP2	+	TGACCCAAGTCAGGGTGTGACGCCAGGCTTTTATAAGTTCCCACTTCCTGCTCCCACCTACATCTACCCCTGAAAAGAACAGAACCGTGGACACATCAGC
	Std	BP1	-	AAGAGATAGATTGGTTTTCTCACATTATACATACATCACCTATTAACTAGCAGGAAGTGGGAACTTATAAAAGCCTGGCGTCACACCCTGACTTGGGTCA
	Std	BP2	-	atcaatagcatttcacccctcacccctctctcaccctcctgccCACTGTAGCCTAAGTGACATCATCAATACGAAGAAAGCATCATTCCTGATCACAAGC

* Two versions of each breakpoint were used to detect the *Inv* allele.

Table S9. Polymorphic inversion frequencies in different populations by tag SNPs and/or BreakSeq/svgem analysis.

																												F_{st}	
					European					African				Asian					American										
Inversion	Method	SNP	ALL	CEU	GBR	FIN	TSI	IBS	EUR	YRI	LWK	ASW	AFR	CHB	CHS	JPT	ASN	CLM	MXL	PUR	AMR	Populations	Continents
Hsinv0003	Global tag SNP	rs2303648	0.83	0.82	0.77	0.73	0.70	0.82	0.75	0.79	0.71	0.80	0.76	0.99	0.99	0.98	0.99	0.81	0.88	0.82	0.84	0.07	0.09
	BreakSeq /svgem	-	0.84 (569)	0.84 (61)	0.74 (37)	0.78 (39)	0.70 (67)	NA (1)	0.76 (225)	0.78 (69)	0.82 (12)	0.80 (23)	0.79 (95)	1.00 (63)	0.98 (49)	1.00 (51)	0.99 (165)	0.75 (17)	0.86 (32)	0.77 (26)	0.81 (75)	0.08	0.09
	Global tag SNP	rs6731222	0.46	0.37	0.34	0.38	0.42	0.43	0.38	0.66	0.76	0.61	0.69	0.44	0.36	0.40	0.40	0.40	0.44	0.41	0.42	0.06	0.08
Hsinv0041	BreakSeq / svgem	-	0.48 (575)	0.46 (50)	0.30 (38)	0.44 (36)	0.44 (65)	NA (1)	0.42 (218)	0.68 (62)	0.65 (21)	0.62 (29)	0.66 (112)	0.49 (66)	0.41 (50)	0.44 (56)	0.45 (172)	0.40 (16)	0.47 (37)	0.58 (20)	0.49 (73)	0.02	0.04
	Global tag SNP	rs382113	0.71	0.84	0.67	0.80	0.85	0.79	0.83	0.90	0.91	0.89	0.90	0.38	0.44	0.34	0.34	0.73	0.64	0.78	0.71	0.21	0.25
Hsinv0059	BreakSeq / svgem	-	0.69 (589)	0.76 (53)	0.90 (41)	0.79 (45)	0.81 (79)	NA (0)	0.81 (218)	0.87 (66)	0.82 (19)	0.83 (33)	0.85 (112)	0.36 (66)	0.47 (54)	0.42 (32)	0.41 (172)	0.67 (21)	0.72 (36)	0.70 (30)	0.72 (87)	0.15	0.20
	Global tag SNP	rs6460609	0.56	0.61	0.61	0.74	0.59	0.54	0.63	0.30	0.28	0.35	0.30	0.70	0.71	0.69	0.70	0.53	0.63	0.55	0.57	0.09	0.12
Hsinv0063	BreakSeq / svgem	-	0.65 (532)	0.71 (61)	0.72 (35)	0.74 (31)	0.68 (83)	NA (0)	0.71 (210)	0.34 (60)	0.25 (14)	0.30 (25)	0.31 (99)	0.68 (66)	0.63 (47)	0.73 (32)	0.74 (165)	0.67 (12)	0.79 (29)	0.62 (17)	0.72 (58)	0.11	0.15
Hsinv0006	BreakSeq / svgem	-	0.51 (545)	0.63 (52)	0.75 (39)	0.73 (40)	0.58 (69)	NA (0)	0.67 (200)	0.02 (67)	0.09 (23)	0.27 (30)	0.10 (120)	0.52 (57)	0.63 (52)	0.56 (31)	0.57 (160)	0.39 (18)	0.58 (31)	0.72 (16)	0.61 (65)	0.22	0.25
Hsinv0058	BreakSeq / svgem	-	0.60 (585)	0.62 (63)	0.67 (36)	0.55 (35)	0.66 (90)	NA (1)	0.63 (225)	0.67 (62)	0.58 (20)	0.56 (28)	0.63 (110)	0.56 (68)	0.39 (54)	0.57 (56)	0.52 (178)	0.71 (16)	0.72 (37)	0.61 (21)	0.70 (74)	0.01	0.01
Hsinv0068	BreakSeq / svgem	-	0.71 (544)	0.62 (51)	0.44 (27)	0.35 (24)	0.74 (67)	NA (0)	0.62 (189)	0.89 (40)	NA (5)	NA (2)	0.77 (47)	1.00 (52)	1.00 (20)	1.00 (17)	1.00 (69)	NA (3)	NA (9)	NA (7)	0.08 (19)	0.25	0.32
Hsinv0004	Global tag SNP	rs1775463	0.14	0.21	0.25	0.19	0.18	0.14	0.20	0.01	0.03	0.07	0.03	0.14	0.11	0.14	0.13	0.25	0.15	0.20	0.20	0.03	0.05
Hsinv0031	Global tag SNP	rs993231	0.64	0.69	0.71	0.67	0.72	0.39	0.69	0.64	0.54	0.65	0.60	0.62	0.58	0.55	0.58	0.70	0.68	0.66	0.68	0.01	0.02
Hsinv0040	Global tag SNP	rs4300008	0.79	0.76	0.80	0.80	0.62	0.68	0.74	0.74	0.79	0.74	0.76	0.89	0.89	0.90	0.89	0.71	0.83	0.75	0.77	0.03	0.03
Hsinv0045	Global tag SNP	rs7293610	0.52	0.41	0.51	0.50	0.56	0.43	0.49	0.61	0.42	0.58	0.53	0.55	0.62	0.55	0.57	0.47	0.47	0.55	0.49	0.01	0.00
	Deletion tag SNP	rs206276	0.53	0.28	0.30	0.21	0.36	0.36	0.29	0.64	0.66	0.43	0.59	0.90	0.91	0.88	0.89	0.35	0.34	0.42	0.37	0.26	0.30
Hsinv0052	Inversion tag SNP	rs13073727	0.22	0.39	0.39	0.53	0.33	0.43	0.41	0.04	0.05	0.14	0.07	0.02	0.01	0.07	0.03	0.33	0.38	0.36	0.36	0.19	0.22
		N (1000GP)	1092	85	89	93	98	14	379	88	97	61	246	97	100	89	296	60	66	55	181		

All frequencies correspond to the *Inv* allele (alternative allele to the one in the reference HG18 genome). The number of individuals analyzed by tag SNPs in each 1000GP population is shown in the last line of the table. The number of individuals for each inversion and population with reads spanning the BPs detected by BreakSeq and used by svgem to calculate allele frequency is shown in parentheses next to the frequency values. Fst values >0.1 are shown in boldcase. NA (Not Applicable) indicates that svgem frequency calculation is based on data from less than 10 individuals for a particular inversion and population. See Abecasis et al. 2012 for the meaning of the initials of each population.

Table S10. Fixed and shared SNPs between *Std* and *Inv* alleles. The total number of tag SNPs shown in the last column for both 1000GP and HapMap data corresponds to the sum of fixed SNPs found outside (±10 kb) and inside the inverted sequence.

	1000 Genomes Project Data					HapMap Project Data					
	Outside inversion		Within inversion		tag SNPs	Outside inversion		Within inversion		tag SNPs	Origin
Inversion	Fixed	Shared	Fixed	Shared		Fixed	Shared	Fixed	Shared		
Inversions with IRs											
HsInv0030	0	35	0	13	0	0	15	0	9	0	Recurrent
HsInv0055	0	9	0	1	0	0	1	0	0	0	Recurrent
HsInv0069	0	41	0	4	0	0	19	0	2	0	Recurrent
HsInv0031	5	39	5	0	10	4	17	3	0	7	Unique
HsInv0040	15	6	3	0	18	7	0	1	0	8	Unique
HsInv0045	0	94	1	0	1	0	27	0	0	0	Unique
HsInv0061	0	0	0	0	0	0	0	0	0	0	Unique
HsInv0072	0	0	0	0	0	0	0	0	0	0	Unique
Inversions without IRs											
HsInv0003	24	5	0	0	24	8	0	0	0	8	Unique
HsInv0004	47	1	5	0	52	14	0	2	0	16	Unique
HsInv0006	10	6	0	0	10	6	1	0	0	6	Unique
HsInv0041	11	61	1	0	12	1	5	0	0	1	Unique
HsInv0052	37	1	3	0	40	0	0	0	0	0	Unique
HsInv0058	16	12	3	0	19	6	19	2	0	8	Unique
HsInv0059	1	5	0	0	1	1	3	0	0	1	Unique
HsInv0063	15	5	4	0	19	6	0	1	0	7	Unique
HsInv0068	10	1	1	0	11	5	2	1	0	6	Unique

Table S11. Tag SNPs associated to Std and Inv alleles. Tag SNPs (r^2=1) resulting from combining our experimental genotypes with either HapMap SNPs (60 individuals available) or 1000GP (35 individuals available) are shown in the first two columns for all inversions. For inversions with simple breakpoints that could be analyzed by BreakSeq, svgem results for the CEU population and all 14 1000GP populations are also given. Inversion breakpoints (BP1 and BP2) are shown in grey in the relative position with respect to the surrounding SNPs for each inversion. Highlighted in orange are the SNPs used to infer genotypes in the 1092 individuals of the 1000GP (since no data from all populations are available for the inversions with IRs, the SNP located within the inverted sequence and closer to the breakpoints was used). Only r^2 values >0.95 are shown in this table. SNPs not found in the HapMap collection have blank spaces. Those SNPs located within the inverted segments are marked with an asterisk.

BreakSeq/svgem inversions

Inversion	SNP	HapMap 60 CEU	1000GP 35 CEU	svgem 1000GP CEU[1]	svgem 1000GP All
HsInv0003	rs6425098		1.00	1.00	0.97
	rs6425099		1.00	-	-
	rs6425100		1.00	1.00	0.99
	rs6691810		1.00	0.95	0.97
	rs10753006		1.00	1.00	0.98
	rs59529645			1.00	-
	rs10753007		1.00	1.00	0.99
	rs6668027		1.00	-	0.98
	rs6692630		1.00	-	0.96
	rs6692806	1.00	1.00	-	0.97
	rs6671314	1.00	1.00	0.97	0.97
	rs6425101		1.00	-	1.00
	rs6425102	1.00	1.00	0.96	-
	rs6425103		1.00	0.96	0.98
	rs4465159		1.00	0.95	0.98
	rs10912079	1.00	1.00	1.00	0.99
	rs10753008	1.00	1.00	1.00	0.99
	rs10753009		1.00	-	0.98
	rs10753010	1.00	0.92	-	0.98
	rs7554573	1.00	1.00	-	0.99
	rs2383645	1.00	1.00	1.00	0.99
	rs2383646	1.00	0.92	-	0.97
	rs2383647		1.00	0.96	0.99
	rs2383648		**1.00**	**1.00**	**1.00**
	rs10753011		1.00	-	-
	BP1				
	BP2				
	rs10753012		1.00	-	0.96
	rs4233129	1.00	1.00	-	-
HsInv0006	rs10900470		1.00	-	-
	rs34119249		1.00	1.00	-
	rs4333898	1.00	1.00	1.00	-
	rs11240383		1.00	1.00	-
	rs11240384	1.00	1.00	-	-
	rs11240385	1.00	1.00	1.00	-
	rs10900471		1.00	0.97	-
	rs6682400	1.00	1.00	0.98	-
	rs11240387		1.00	0.97	-
	rs11240388 [2]	1.00		1.00	-
	BP1				
	BP2				
	rs11240390	1.00	1.00	0.98	-
	rs10900472			1.00	-
	rs6664706			1.00	-
	rs12137294			1.00	-

Other inversions

Inversion	SNP	HapMap 60 CEU	1000GP 35 CEU
HsInv0031	rs2937141	1.00	1.00
	rs34235919		1.00
	rs9939451	1.00	1.00
	rs7191362		1.00
	rs11149714	1.00	1.00
	BP1		
	rs9933231*	**1.00**	**1.00**
	rs9923783*	1.00	1.00
	rs9935376*		1.00
	rs9935396*	1.00	1.00
	rs8056715*		1.00
	BP2		
	rs8057854	1.00	1.00
HsInv0040	rs6430776		1.00
	rs149768985		1.00
	rs9287496	1.00	1.00
	BP1		
	rs7425791*		1.00
	rs4954685*	1.00	1.00
	rs4350808*		**1.00**
	BP2		
	rs10199066		1.00
	rs10199067		1.00
	rs6716175	1.00	1.00
	rs6743792	1.00	1.00
	rs7565340		1.00
	rs7571194	1.00	1.00
	rs4266068	1.00	1.00
	rs4258857		1.00
	rs5002026		1.00
	rs4477977	1.00	1.00
	rs6741774	1.00	1.00
	rs4499476		1.00
HsInv0045	*BP1*		
	rs7283610*		**1.00**
	BP2		
HsInv0004	rs1747812		1.00
	rs1775446		1.00
	rs1775445		1.00
	rs1747813		1.00
	rs1342694	1.00	1.00
	rs2477069	1.00	1.00
	rs1775469	1.00	1.00
	rs1747825	1.00	1.00
	rs1775468	1.00	1.00

HsInv0041	rs10190507		1.00	1.00	0.99
	rs10166620		1.00	0.99	-
	rs6710450	1.00	1.00	1.00	0.95
	rs4674898		1.00	0.97	-
	BP1				
	rs11898198*		1.00	-	-
	BP2				
	rs6733222		**1.00**	**1.00**	**1.00**
	rs6718446			1.00	-
	rs4674899		1.00	-	-
	rs4674900		1.00	0.97	-
	rs9784048		1.00	0.97	-
	rs9784050		1.00	1.00	1.00
	rs9288590		1.00	1.00	-
	rs11904082		1.00	-	-
	rs6734149			0.99	-
	rs142872425			1.00	-
	rs10804321			1.00	-
	rs74876662			1.00	-
	rs6709124			1.00	-
HsInv0058	rs150278551		1.00	-	-
	rs115545454		1.00	-	-
	rs115762310 [3]	1.00	1.00	-	-
	rs115652646 [3]	1.00	1.00	0.99	-
	rs114909794		1.00	-	-
	rs141877547		1.00	0.96	-
	rs138095131		1.00	-	-
	rs149731696		1.00	-	-
	rs116070819 [3]	1.00	1.00	-	-
	rs115708357 [3]	1.00	1.00	-	-
	rs9262567	1.00	No data	No data	No data
	BP1				
	rs10947128*	1.00	No data	No data	No data
	rs115901304*[3]	1.00	1.00	-	-
	rs114547037*		1.00	0.97	-
	rs143980660*		1.00	-	-
	BP2				
	rs115187244 [3]	1.00	1.00	0.98	-
	rs115277645		1.00	-	-
	rs116448331		1.00	-	-
	rs116230904		1.00	-	-
	rs114946731		1.00	-	-
	rs147091126		1.00	-	-
HsInv0059	rs4707529	1.00	1.00	-	-
	BP1				
	BP2				
	rs7773203		-	-	0.98
	rs282113		**0.91**	-	**1.00**
	rs43993		0.91	-	1.00
HsInv0063	rs7805662	1.00	1.00		-
	rs146848561		1.00		-
	rs7782419		1.00		-
	rs62460368		1.00		-
	rs1581552		1.00		-
	rs57364974		1.00		-
	rs58332498		1.00		-
	rs1568866	1.00	1.00		-
	rs62460371		1.00		-
	rs4719030	1.00	1.00		-
	rs4717571		1.00		-
	rs1568864			1.00	-
	rs1464853 [2]	1.00			-
	rs1464851 [2]	1.00			-
	esv2662405	-	1.00		0.98
	BP1				
	rs58830691*		1.00		-
	rs4323369*		1.00		-
	rs1525303*	1.00	1.00		0.96
	rs6460609*		1.00		**0.98**
	BP2				
	rs56231320		1.00		-
	rs10269258	1.00	1.00		-
	rs12698965		1.00		-
	rs1775467	1.00	1.00		
	rs1775466	1.00	1.00		
	rs1775465	1.00	1.00		
	rs1775464	1.00	1.00		
	rs1747823	1.00	1.00		
	rs1747822	1.00	1.00		
	rs1747821		1.00		
	rs1747820		1.00		
	rs2497863		1.00		
	BP1				
	rs1627999*	1.00	1.00		
	rs1623752*	1.00	1.00		
	rs1622904*		1.00		
	rs1622740*		1.00		
	rs1775463*		**1.00**		
	BP2				
	rs1747828		1.00		
	rs1775462		1.00		
	rs1747824		1.00		
	rs1775461		1.00		
	rs1775460		1.00		
	rs1775459		1.00		
	rs1775458		1.00		
	rs2649557		1.00		
	rs2497862	1.00	1.00		
	rs2488402		1.00		
	rs2488403		1.00		
	rs2488404		1.00		
	rs2488405		1.00		
	rs35191270		1.00		
	rs2454642		1.00		
	rs2649556	1.00	1.00		
	rs1578718		1.00		
	rs2797147		1.00		
	rs2266035		1.00		
	rs2266034		1.00		
	rs2454641		1.00		
	rs2265664		1.00		
	rs2266033		1.00		
	rs1592221		1.00		
	rs1592220		1.00		
	rs2454640	1.00	1.00		
	rs1578719		1.00		
	rs34451672		1.00		
	rs55873039		1.00		
HsInv0052	rs12185924		1.00		
	rs147091126		1.00		
	rs11916098		1.00		
	rs11916978		1.00		
	rs12185935		1.00		
	rs12185939		1.00		
	rs11923230		1.00		
	rs13066346		1.00		
	rs13086592		1.00		
	rs13091715		1.00		
	rs12490361		1.00		
	rs13098441		1.00		
	rs13098325		1.00		
	rs35173683		1.00		
	rs34394037		1.00		
	rs13096689		1.00		
	rs13069624		1.00		
	rs13090052		1.00		
	rs11927690		1.00		
	BP1				
	rs12486709*		1.00		
	rs2193271*		1.00		
	rs13073727*		**1.00**		
	BP2				
	rs13079279		1.00		
	rs13079616		1.00		
	rs137944185		1.00		
	rs12488446		1.00		

HsInv0068	rs513987	1.00	1.00	-	-
	rs2781744		1.00	-	-
	rs4013967	1.00	1.00	-	-
	rs7859799	1.00	1.00	-	-
	BP1				
	rs5002587*	**1.00**	**1.00**	-	-
	BP2				
	rs2604263		1.00	-	-
	rs641494		1.00	-	-
	rs2604264	1.00	1.00	-	-
	rs2604265		1.00	-	-
	rs2781746		1.00	-	-
	rs2604266	1.00	1.00	-	-

	rs13085625		1.00
	rs13085867		1.00
	rs13091922		1.00
	rs12493705		1.00
	rs114862500		1.00
	rs13060757		1.00
	rs13060924		1.00
	rs10446338		1.00
	rs10446339		1.00
	rs13067540		1.00
	rs11927628		1.00
	rs13096150		1.00
	rs13076086		1.00
	rs13097182		1.00

[1] The number of individuals included in this analysis is different for each inversion. See the number in parenthesis in the CEU population in Table S8.

[2] A single individual has different SNP genotype in HapMap compared to 1000GP data.

[3] SNPs with different identifier in 1000GP and HapMap data.

Table S13. Genes located adjacent to inversions. Genes in RefSeq and GENCODE v19 annotations are shown. Non-coding RNAs and pseudogenes have also been included. For regions upstream and downstream of the inversion, all genes found between the inversion and the next protein-coding gene are shown regardless of the distance where this last gene is found. The distance is measured from the closest end of the gene to the nearest inversion breakpoint. RefSeq names have been used for genes when possible, if not avaialable, GENECODE IDs are included instead. Intron numbers are counted relative to the majority of the different isoforms.

	Inversion region		Upstream genes	Downstream genes
Inversion	**Genes**	**Relative position**	**Name**	**Name**
Inversions within introns				
HsInv0006	DSTYK	Inversion within intron 1	RBBP5 *RP11-383G10.3 (processed pseudogene)*	TMCC2
HsInv0059	GABRR1	Inversion within intron 1	PM20D2	GABRR2
HsInv0061	*LOC100422737 (lincRNA)*	Inversion within intron 6	*RP1-60O10.2 (lincRNA)* QRSL1	*RNU6-117P (snRNA)* *RP1-60O10.1 (lincRNA)* *MIR587 (miRNA)* C6orf203
HsInv0055	*AC016561.1 (unitary pseudogene)*	Inversion within intron 1	RNF180	RGS7BP
HsInv0031	7 spliced ESTs*	Inversion within intron	*LOC400548 (lincRNA)* FAM92B	CTC-786C10.1 *LINC00311 (lincRNA)*
HsInv0052	*BC019327, BC073807 (lincRNA)**	Inversion within intron 1	OTOL1	*LINC01192 (lincRNA)*
Genes within IRs at inversion breakpoints				
HsInv0030	CTRB2 CTRB1	Genes partially overlapping inversion BPs	ZFP1	*LOC100506281* BCAR1
HsInv0069	*FAM225B (lincRNA)* *FAM225A (lincRNA)*	Genes totally overlapping inversion BPs	ZFP37	SLC31A2
Intergenic inversions				
HsInv0003	-	-	*RP11-445P19.3 (lincRNA)* *LINC01037 (lincRNA)*	*RP11-445P19.2 (processed pseudogene)* *FDSP1 (processed pseudogene)*
HsInv0004	-	-	DENND1B *FAM204BP (processed pseudogene)*	C1orf53
HsInv0036	-	-	ANKRD62	*RP11-64C12.7 (processed pseudogene)* *C18orf61* CIDEA
HsInv0040	-	-	*AC069394.1 (lincRNA)* HNMT	*AC097523.1 (processed pseudogene)* *AC097523.2 (processed pseudogene)* *AC097523.3 (processed pseudogene)* *AC097721.1 (processed pseudogene)* *AC097721.2 (lincRNA)* SPOPL
HsInv0041	-	-	FAM124B	CUL3
HsInv0045	-	-	CYYR1	ADAMTS1
HsInv0058	-	-	MUC22	*HCG22 (lincRNA)* *RNU6-1133P (snRNA)*
HsInv0063	-	-	AUTS2	WBSCR17
HsInv0068	-	-	-	*LOC101927358 (lincRNA)*
HsInv0072	-	-	*KRT8P14 (processed pseudogene)* *LINC01204 (lincRNA)*	*LOC392452 (processed pseudogene)* *MIR221 (microRNA)* *MIR222 (microRNA)* *RP6-99M1.2 (lincRNA)*

* Not included in RefSeq or GENCODE v19 gene annotations.

3.2 Participación en otros estudios

En este apartado se intenta resumir cómo se ha contribuido a otros estudios a partir de la validación y descarte de errores de este conjunto de inversiones. En ese aspecto se han realizado dos colaboraciones principales, la primera está relacionada con el desarrollo de la aplicación estadística svgem [Lucas-Lledó et al. 2014], y la segunda con la genotipación masiva de inversiones en individuos de poblaciones de todo el mundo mediante una técnica que se basa en el *MLPA* [S.Villatoro and M.Cáceres. unpublished data] y que forma parte de un análisis a nivel global en el genoma humano de todas las inversiones polimórficas validadas.

La participación en el desarrollo de svgem fue indirecta a través de la genotipación bioinformática que se realizó mediante la pipeline *BreakSeq* [Lam et al. 2010] en 1092 individuos de 14 poblaciones correspondientes al proyecto de los 1000 Genomas [1000 Genomes Project Consortium. 2012] para 7 inversiones polimórficas detectadas en *HuRef,* además de otras inversiones entre las que se encuentra la inversión *HsInv0201* utilizada en el estudio de Lucas-Lledó y colaboradores para analizar el funcionamiento de svgem sobre datos de genotipación reales. Tal y como hemos mostrado en el apartado de materiales y métodos, *BreakSeq* permite genotipar los individuos mediante el alineamiento de *reads* provenientes de su secuenciación sobre una librería de puntos de rotura de las variantes estructurales a detectar y sobre el genoma de Referencia. Con el objetivo de usar *BreakSeq* para genotipar los 1092 individuos en cuestión, se modificó su código para descargar los datos para un sólo individuo, de manera que se analizasen, se guardasen los resultados y se eliminasen los datos iniciales, y esto para todos los individuos. Esta modificación surgió de la imposibilidad de almacenar todos los datos de secuenciación para todos los individuos en el servidor del grupo. Mi participación en este proceso consistió en solucionar problemas relacionados con la descarga de los datos, detección de errores y sobretodo con la gestión del funcionamiento del programa *BreakSeq*. Se realizó un gran número de pruebas en las que se controló el inicio y el final del programa *BreakSeq* y se automatizó su funcionamiento en los momentos en que se dispuso de máximo ancho de banda para la descarga de los datos. También fui responsable de parte del análisis del funcionamiento a través de la comparación de los resultados de la genotipación bioinformática y de la genotipación experimental de las inversiones en *HuRef*.

En cuanto a la participación en el estudio global de las inversiones, se dio a través de la aportación de las condiciones usadas para la validación experimental de las inversiones polimórficas en *HuRef,* así como la localización de las regiones con mejores condiciones para el diseño de cebadores. Además se aportaron los resultados de la genotipación por *PCR* y *PCR* inversa para la comparación con los obtenidos en el nuevo método basado en *MLPA*. Finalmente se aportaron los genotipos y demás análisis de algunas inversiones no presentes en la genotipación por *MLPA* para el estudio global.

Además de estos dos estudios, se ha participado en 3 estudios más, y en la siguiente tabla (**Tabla 3.1**) se muestra un resumen de los estudios a los que ha contribuido cada inversión.

Tabla 3.1. Participación en otros estudios.

Identificador InvFEST	Artículo o estudio	Objetivo del estudio	Contribución / colaboración	Estado del estudio
HsInv0003, *HsInv0004*, *HsInv0031*, *HsInv0040*, *HsInv0041*, *HsInv0045*, *HsInv0055*, *HsInv0058*, *HsInv0059*, *HsInv0061*, *HsInv0063*, *HsInv0068*, *HsInv0072*, *HsInv0095**	Large-scale genotyping of inversions in human populations.	Modificación de la técnica MLPA para la genotipación masiva de individuos para inversiones polimórficas y análisis global de las inversiones en el genoma humano	Se han proporcionado 14 inversiones polimórficas demostradas experimentalmente junto con las condiciones y cebadores tanto para el protocolo de PCR estándar como el protocolo de PCR inversa. Además se han proporcionado los genotipos de 90 individuos Europeos procedentes de la placa 01 del proyecto HapMap obtenidos de la genotipación experimental, para la puesta a punto de la técnica.	No publicado
*HsInv0201***	Population genetic analysis of bi-allelic structural variants from low-coverage sequence data with an expectation-maximizati on algorithm.	Aplicación de svgem, una aplicación estadística que permite evitar el sesgo en la detección de los alelos en variantes polimórficas.	La aplicación del software Breakseq para la genotipación bioinformática de los 1092 individuos correspondientes a 14 poblaciones en el proyecto de los 1000 Genomas, ha sido clave para el desarrollo de la aplicación estadística svgem.	Publicado
*HsInv0102***, *HsInv0379***	Diferentes estudios que han usado BreakSeq y svgem para genotipar bioinformáticamente individuos de varias poblaciones.	Varios objetivos pero uno común es la genotipación de individuos de varias poblaciones.	Se han genotipado 1092 individuos procedentes del proyecto de los 1000 Genomas mediante el uso del software Breakseq y la aplicación estadística svgem.	No publicado
HsInv0004, *HsInv0031*, *HsInv0061*, *HsInv0063*, *HsInv0095**	GRIAL, a program specifically designed to predict accurately inversions from PEM data.	Predicción de inversiones polimórficas en el Genoma Humano a partir de datos existentes producidos por la técnica de detección de Paired-end Mapping.	Se han proporcionado 5 inversiones polimórficas demostradas experimentalmente junto con las condiciones y cebadores para su reproducción por PCR y genotipación. Además se han proporcionado los genotipos de 90 individuos Europeos correspondientes a la placa 01 del proyecto HapMap.	No publicado

* *La inversión HsInv0095 se detectó en HuRef en el estudio de Pang et al. 2010, por lo que no pertenece al conjunto analizado.*

** *Las inversiones HsInv0102, HsInv0201 y HsInv0379 no fueron descubiertas en HuRef, por lo que no pertenecen al conjunto analizado.*

4. DISCUSIÓN

4. DISCUSIÓN

4.1 Análisis de las diferencias en la definición de los puntos de rotura

El análisis manual de la secuencia en ambos genomas de las 90 regiones supuestamente invertidas nos permitió mejorar la definición de los puntos de rotura de la mayoría de inversiones y que se descartase un porcentaje importante de falsos positivos. Más concretamente se redefinieron los puntos de rotura de 59 inversiones. Los puntos de rotura de 18 de estas inversiones estaban aceptablemente bien definidos, a pesar de existir una diferencia total (sumando ambos puntos de rotura) de entre 1 y 3 bases respecto los definidos en nuestro estudio. Las 41 inversiones restantes tenían sus puntos de rotura definidos con un error de más de 3 nucleótidos, 27 de ellas con una diferencia total de menos de 1 Kb y 14 inversiones con una diferencia total mayor a 1 Kb. Estas diferencias en la definición de los puntos de rotura provienen principalmente del análisis automatizado de los alineamientos sin tener en cuenta las variantes estructurales o los elementos que se encuentran en la secuencia donde se van a localizar. Por ejemplo, en muchas de las zonas donde Levy y colaboradores definieron los puntos de rotura, hay pequeñas inserciones/deleciones que afectaron a la precisión de su definición. En cuanto a los falsos positivos relacionados con la comparación genómica, se descartaron 31 en total. En el análisis del alineamiento de ambas secuencias encontramos diferentes casos en que no se encontró un alineamiento invertido único. La situación más repetida fue que la secuencia en *HuRef* alineaba en orientación invertida con la secuencia del genoma de Referencia pero este alineamiento no era único ya que al añadir secuencia flanqueante extra encontramos un segundo alineamiento, por lo que estas regiones no corresponden a inversiones sino a duplicaciones invertidas. La secuencia de *HuRef* presentaba varios *gaps* en la región, y en muchos casos correspondían a una duplicación segmental. En otros casos, además de la presencia de *gaps*, la secuencia era completamente repetitiva. Otra combinación que encontramos fue la presencia de inserciones y deleciones junto a los *gaps*. En el resto de regiones no encontramos alineamiento invertido de la secuencia, incluso al añadir secuencia flanqueante extra, y se atribuyen a errores en el mapeo de regiones invertidas adyacentes a la secuencia entre ambos genomas en el proceso de comparación, aunque en casos puntuales no se hallaron elementos que se pudiesen relacionar con el error en la comparación. En conjunto, 31 errores fueron resultado de la comparación entre ambos genomas, un porcentaje importante que corrobora que el proceso de comparación de ambos genomas ensamblados no tuvo como objetivo principal la detección de inversiones cromosómicas.

Nuestros resultados sobre el análisis de la comparación genómica se explican por las diferencias entre los objetivos del estudio de Levy y colaboradores y del nuestro. En el primer vistazo que le dimos a las coordenadas de las 90 inversiones publicadas [Levy et al. 2007], pudimos observar que no correspondían a ningún rango o intervalo, sino que los puntos de rotura fueron definidos en su extensión mínima, un nucleótido, y aparentemente delimitaban los extremos internos de la inversión. Por lo tanto, no se incluyeron las duplicaciones segmentales en las que se localizan los puntos de rotura en muchas inversiones; a pesar de que al tratarse de inversiones detectadas en todo el genoma era poco probable que no hubiese ninguna de este tipo. Además, en un estudio anterior de detección de variantes estructurales por comparación genómica [Feuk et al. 2005] ya se correlacionó la variación estructural con las repeticiones con bajo número de copias, *LCRs*, también conocidas como duplicaciones segmentales. Este hecho evidenció que la secuencia de los puntos de rotura no había sido analizada de una manera específica para cada inversión. Las 90 inversiones putativas forman parte de los resultados globales de una comparación en que se analizó con mayor detalle la variación nucleotídica y las inserciones/deleciones de menor tamaño. En conjunto, el planteamiento del estudio sugiere que el objetivo no fue analizar la variación estructural de una manera precisa, especialmente en el caso de las inversiones, que tan solo fueron detectadas a gran escala.

Idealmente, la comparación de dos genomas perfectamente ensamblados sin errores daría como resultado la detección de todas las inversiones sin incluir falsos positivos, siempre y cuando la comparación estuviese dirigida a la detección de estas variantes estructurales. No es el caso del estudio de Levy y colaboradores [Levy et al. 2007]. La comparación se planteó como un alineamiento local de ambas secuencias y eso provocó una parte importante de los falsos positivos detectados en nuestro estudio. Los alineamientos continuos de máxima identidad y misma orientación formaron grupos donde se permitieron las discontinuidades como inserciones, deleciones o *gaps*, pero no cambios en la orientación. Estos grupos formaron bloques en los que sí se permitieron discontinuidades en la orientación de los alineamientos. Entre otras cosas, esta estrategia hizo que no se pudiese diferenciar bien las duplicaciones segmentales invertidas de las inversiones reales, ya que quedaron aisladas de sus parejas en bloques diferentes.

Por el contrario, el planteamiento de nuestro estudio fue analizar manualmente los puntos de rotura y las secuencias flanqueantes. Por eso se definieron los puntos de rotura como intervalos para incluir los localizados en repeticiones invertidas y delimitar bien las secuencias implicadas en la reorganización. La inspección detallada de las secuencias permitió determinar mejor la presencia de regiones de micro-homología implicadas en la generación de las inversiones con puntos de rotura no localizados en repeticiones invertidas (RIs). Además se incluyó cualquier elemento que afectase al punto de rotura como son las inserciones/deleciones. En el caso de las inversiones con puntos de rotura en RIs, la definición de sus puntos de rotura no fue trivial y tampoco siempre es posible. En nuestro estudio se definieron mediante el alineamiento múltiple de las RIs en las dos

orientaciones, que permitió detectar en qué zona de la duplicación o elemento repetitivo se produjo la inversión a partir del intercambio de cambios nucleotídicos entre éstas. Se acotó así la zona donde están localizados, contribuyendo a la mejora de la definición de este conjunto de inversiones.

En el caso de variantes estructurales balanceadas como los son las inversiones, una detección a gran escala como la que se realizó en el estudio de Levy y colaboradores [Levy et al. 2007] implica la inclusión de una gran cantidad de falsos positivos. El análisis manual de la secuencia nos permitió descartarlos. También era esperable que no se hubiesen detectado otras muchas inversiones debido a la estrategia poco dirigida a la detección de inversiones usada en el proceso de comparación. Precisamente en un segundo estudio sobre la variación estructural presente en *HuRef* [Pang et al. 2010] se usó la información generada en la secuenciación *Sanger* en forma de extremos apareados, para detectar por *PEM* las inversiones no identificadas por comparación genómica. El resultado fue de 105 inversiones detectadas por *PEM*, 79 nuevas inversiones y 26 inversiones entre las detectadas por comparación genómica. Este estudio evidenció así que la comparación genómica realizada no detectó correctamente todas las inversiones, ya que al menos la parte de las 79 nuevas inversiones correspondiente a las inversiones reales se tendría que haber detectado en el primer estudio de Levy y colaboradores. Por lo tanto en relación al primer estudio las consideramos falsos negativos y falsos positivos.

Claramente el conjunto de 90 inversiones publicadas [Levy et al. 2007] no representa una aproximación al total de inversiones en el genoma de J. Craig Venter respecto al genoma de Referencia, pero la naturaleza no sesgada del método de detección por comparación genómica hace pensar que sí puede tratarse de un conjunto representativo del total. Por otra parte, el estudio de Pang y colaboradores [Pang et al. 2010] amplió el número de inversiones conocidas en *HuRef*, aunque la naturaleza sesgada del método de detección por *PEM* hacia la detección del alelo estándar en las inversiones con puntos de rotura localizados en repeticiones invertidas y la dificultad de alinear los extremos apareados en zonas duplicadas, hace pensar que una parte importante de las nuevas inversiones detectadas pueden ser falsos positivos y que otra parte resta sin detectarse. Por lo tanto, es difícil saber hasta que punto ambos estudios cubren las inversiones de *HuRef;* lo que sí sabemos es que el conjunto publicado por Levy y colaboradores no contiene la totalidad de inversiones pero que por el método de detección usado, los puntos de rotura de las inversiones detectadas tienen diferentes características y pueden representar los diferentes tipos de inversiones que se encuentran en el genoma humano.

Nuestro estudio se ha centrado en la eliminación de los falsos positivos para tener un conjunto fiable de las inversiones en *HuRef*, aunque evidentemente no se trate del total de inversiones que existen en este genoma. Esto ha requerido un análisis manual de la secuencia y una redefinición de los puntos de rotura en mucho casos, puntos además importantes para el diseño de los experimentos de *PCR*. Finalmente los resultados han

contribuido a la generación de un catálogo no redundante de inversiones fiables que se encuentra en la base de datos *InvFEST* [Martínez-Fundichely et al. 2013] y además junto con los datos mostrados sobre detección demuestran que las inversiones, por sus características balanceadas, no sólo requieren un análisis dirigido para poder diferenciarlas de falsos positivos y definir sus puntos de rotura con precisión; sino que también requieren un método de detección que tenga en cuenta sus características para intentar evitar los falsos negativos.

4.2 La repetitividad del genoma humano como fuente de errores

El proceso de ensamblaje de un genoma tiene como objetivo lograr la secuencia más completa y fiable posible y para ello se han de ordenar y orientar las secuencias formadas por fragmentos secuenciados. En genomas con poca cantidad de secuencia repetitiva es más fácil ponerles orden que en genomas con un porcentaje alto como es el genoma humano, donde se estima que el 50% de la secuencia es repetitiva. En estas regiones es más difícil conocer el orden de los fragmentos debido a que no existe un alineamiento único, y las convierte en susceptibles a generar errores de ensamblaje [Bailey et al. 2001]. La dificultad aumenta cuando se trata de duplicaciones segmentales, por su tamaño superior a 1 Kb y su alta similitud. En uno de los primeros estudios que se realizaron con la secuencia borrador del genoma humano, se alineó la secuencia consigo misma para detectar este tipo de repeticiones y se estableció que representan el 5% del genoma, cifra muy superior a la esperada en aquel momento. También se demostró que las secuencias duplicadas están sobrerrepresentadas en los *contigs* sin ordenar o asignar [Bailey et al. 2001]. Se analizó la cobertura de estas regiones en el genoma mediante ensayos de *FISH* sobre los cromosomas, en los que se usaron *BACs* que contienen duplicaciones intercromosómicas y además se realizó un búsqueda por *BLAST* de las regiones duplicadas en sus cromosomas correspondientes en el genoma de Referencia. Los resultados mostraron que sólo el 47% de los cromosomas con señal en el experimento de *FISH* tenía la secuencia de la duplicación intracromosómica correspondiente en el genoma de Referencia [Bailey et al. 2001]. Se demostró así que se trataba de regiones mal ensambladas, mal asignadas o con una cobertura baja. En otro estudio realizado sobre el ensamblaje público *NCBI Build 30* del genoma humano, se detectaron mediante *BLAST* 38.9 Mb de secuencia involucrada en errores de ensamblaje, correspondiente a un 1.28% del genoma [Cheung et al. 2003]. Por lo tanto, teniendo en cuenta ambos estudios, no sólo se demostró que el genoma de Referencia contiene errores de ensamblaje, sino que además representan un porcentaje importante de la secuencia.

Evidentemente se está trabajando para corregirlos y tener un genoma de Referencia de la máxima calidad posible. El organismo encargado de esta tarea es el Consorcio del Genoma de Referencia, *GRC* [http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/]. Desde la

publicación del genoma humano se han generado 38 versiones incluyendo el primer ensamblaje y una parte de estos errores han sido corregidos, pero cabe esperar que algunas regiones donde las duplicaciones segmentales son muy grandes y similares y la complejidad de las repeticiones es alta, no puedan ser resueltas al menos por ahora. En ese sentido y como muestran nuestros resultados, los estudios de detección de variantes son una manera de identificar estos errores ya que aparecen como variantes estructurales (falsos positivos). Por ejemplo las técnicas como *PEM* que en algunos estudios se usan para analizar múltiples genomas, los identifican porque se muestran como variantes estructurales en todos los individuos excepto el genoma de Referencia. Además aunque el número de errores de ensamblaje en el genoma humano no sea muy alto, en este tipo de estudios pueden representar una parte importante de las variantes estructurales detectadas.

Todos estos datos indican lo que ya sabemos, que es primordial que el genoma de Referencia de cualquier especie tenga la máxima calidad posible, ya que de él dependen todos los estudios de genómica comparativa tanto interespecíficos como intraespecíficos y las conclusiones a las que se llegan con ellos. En ese sentido es común pensar en el genoma humano de Referencia como un proyecto terminado, como una secuencia terminada al 100%. Por el contrario, nuestro estudio confirma que más de una década después de su publicación los errores de ensamblaje relacionados con las duplicaciones segmentales siguen presentes y están detrás de un porcentaje importante de falsos positivos en los estudios de genómica comparada.

En nuestro estudio se realizó un esfuerzo muy importante por colaborar con la mejora del genoma de referencia. Al tratarse de un análisis de las inversiones entre *HuRef* y el genoma de Referencia, sólo pudimos detectar los errores de ensamblaje en ambos genomas que conllevan falsos positivos, es decir, las secuencias en orientación errónea que se han detectado como inversiones. Por lo tanto, es posible que haya una parte de errores de ensamblaje comunes en ambos genomas que no hemos detectado y que consideramos falsos negativos. Además, estos errores no se pudieron distinguir de las inversiones reales mediante un análisis manual de la secuencia de los puntos de rotura, por lo que fueron necesarios experimentos de *PCR* sobre el ADN. En el caso del genoma de Referencia, este punto requirió un esfuerzo extra respecto a cualquier validación experimental de una inversión polimórfica. El genoma de Referencia se secuenció a partir de clones de librerías genómicas de distintos individuos de identidad desconocida, por lo que no es posible obtener la muestra de ADN. A cambio se obtuvieron los *BACs* de la secuenciación para cada región potencialmente errónea del genoma, lo que es complicado ya que se encuentran repartidos entre los diferentes centros de secuenciación que participaron en el macro-proyecto. Además hay que comentar que para la demostración de un error de ensamblaje ha de usarse el mismo *BAC* del mismo individuo que se usó en el momento de la secuenciación, ya que otro *BAC* de la misma zona generado a partir de la muestra de ADN de otro individuo (o incluso del mismo que pudiese provenir del otro alelo) no resolvería el problema. Esta es la única manera de asegurarse que no se trata de

una inversión a una frecuencia muy baja y que es realmente un error en el ensamblaje. En algunos casos, el *GRC* ha substituido las regiones con una orientación errónea por la secuencia proveniente de un *BAC* diferente pero eso no soluciona el problema, simplemente substituye la secuencia por la de un *BAC* que puede no representar el mismo alelo provenga o no el *BAC* del mismo individuo.

Los errores en ambos genomas se detectaron a partir de la utilización de la información de los fósmidos usados por Kidd y colaboradores en el año 2008 [Kidd et al. 2008] para detectar variación estructural en 9 individuos mediante *PEM* (8 individuos más 1 extra); en los casos en que todos los individuos presentaban solamente fósmidos discordantes en la región putativamente invertida, se clasificaron las regiones como potencialmente erróneas en el genoma de Referencia. Se realizaron experimentos de *PCR* sobre las muestras de ADN correspondientes a dichos individuos para comprobar la orientación de la secuencia y en el caso de confirmarse que todos tenían una orientación invertida respecto al genoma de Referencia se procedió a la búsqueda y adquisición de los *BACs* correspondientes. El experimento final de *PCR* sobre el ADN del *BAC* nos desveló la existencia o no de los errores en la orientación del genoma de Referencia y de esta manera se demostraron 25 errores de ensamblaje en la versión *NCBI36/hg18* del genoma humano. Para estas regiones se estableció la orientación real del *BAC* y se informó al *GRC*, que está en el proceso de cambiar la orientación de estas regiones en el nuevo ensamblaje del genoma. Hay que comentar que los errores que hemos detectado coinciden con la gran mayoría de errores detectados por *PEM* en distintos estudios.

Hay que tener en cuenta que la secuencia del actual genoma de Referencia es el resultado del proyecto de secuenciación en el que se han invertido más recursos hasta el momento y de una estrategia basada en mapas físicos de marcadores moleculares conocidos, pero evidentemente no todos los genomas secuenciados tienen la misma calidad, y este tipo de errores son comunes a todos los genomas ensamblados. Por lo tanto, es de esperar que a menor calidad del ensamblaje y con estrategias de ensamblaje menos precisas mayor sea el contenido de errores. En concreto la estrategia de secuenciación *shotgun* tiene más dificultades con las zonas duplicadas, debido a que se fragmenta todo el genoma sin tener referencias sobre el orden de los fragmentos, a diferencia de la estrategia basada en mapas físicos. Por este motivo se elimina una parte de las secuencias duplicadas para permitir el ensamblado de los fragmentos y esto afecta a la secuencia que contiene menos duplicaciones segmentales. En un estudio sobre la estrategia de secuenciación *shotgun* se demostró que las duplicaciones grandes de más de 15 Kb y de una identidad mayor del 97% no se resuelven bien en este tipo de ensamblajes [She et al. 2004]. Las duplicaciones pueden contener genes duplicados que se pierden y con ellos una parte de la variación genética. En el caso específico de *HuRef* se cuantificó esta pérdida de regiones duplicadas en un 42.8% de las duplicaciones segmentales anotadas en el ensamblaje *HG18* del genoma de Referencia [Levy et al. 2007]; aunque una parte de las diferencias en contenido de duplicaciones corresponde a la variación entre individuos que se ha

estimado alrededor del 25% en este tipo de regiones. Todo esto se resume a nivel global en una visión simplificada de los genomas secuenciados por *shotgun*, especialmente en las regiones pericentroméricas y subteloméricas que son zonas con abundantes duplicaciones segmentales [She et al. 2004].

En nuestro estudio se demostraron 5 errores en el ensamblaje de *HuRef* siguiendo el mismo procedimiento que con los errores en el genoma de Referencia, aunque con la gran diferencia que *HuRef* proviene de un único individuo, J. Craig Venter, del que se puede obtener la muestra de su ADN. En todos los casos se relacionan con la presencia de elementos repetitivos y *gaps* en la región, que explicarían una mala orientación del fragmento. El hecho de que se hayan detectado 5 errores en *HuRef* frente a los 25 en el genoma humano de Referencia refleja la baja representación de las regiones más difíciles de ensamblar en los genomas secuenciados por *shotgun*. Evidentemente no debe ser confundido con una mayor calidad de ensamblaje de *HuRef*.

En conjunto un tercio de las putativas inversiones son falsos positivos por errores en el ensamblaje de ambos genomas. Es una evidencia directa de que la calidad de los genomas ensamblados afecta a los estudios comparados. Además, el hecho de que las zonas duplicadas sean más proclives a contener errores afecta a los resultados de la comparación genómica, lo cual no implica que la detección de variantes estructurales por comparación genómica no sea un método preciso.

4.3 Las características de los puntos de rotura dividen las inversiones cromosómicas

El análisis de los puntos de rotura dio como resultado que la mitad de las inversiones polimórficas tienen sus puntos de rotura localizados en repeticiones invertidas y la otra mitad tienen puntos de rotura sencillos, aunque contienen otros elementos como inserciones y deleciones. Los puntos de rotura localizados en RIs no sólo se encuentran en *LCRs* sino que también se han encontrado en elementos repetitivos *SINEs* y *LINEs*, por lo que los resultados concuerdan con estudios anteriores [Gu et al. 2008]. Estas características de los puntos de rotura de las inversiones vienen determinadas directamente por el mecanismo responsable de su formación, y en consecuencia en estudios anteriores se concluyó que las inversiones de mayor tamaño suelen tener los puntos de rotura localizados en RIs mientras que las de menor tamaño no [Lam et al. 2010]. Los resultados de nuestro estudio muestran esa tendencia (**Tabla 4.1**). Aunque las inversiones más grandes de uno y otro grupo no parecen tener un tamaño tan distinto, debido a una inversión con puntos de rotura no localizados en RIs que tiene un tamaño de unas 12 Kb; si nos fijamos en el tamaño mínimo y en el número de inversiones de tamaño menor a 500 pb o 1 Kb se aprecia claramente que las inversiones de menor tamaño no tienen sus puntos de rotura en RIs. Por otra parte podemos ver que los puntos de rotura localizados en RIs no se pudieron definir con la misma precisión que los localizados fuera

de éstas y por eso tienen un tamaño mayor. Son indicativos de ello el tamaño máximo y el tamaño promedio.

Tabla 4.1: Comparación del tamaño de inversiones con puntos de rotura localizados y no localizados en RIs.

	PR no localizados en RIs	PR localizados en RIs
Tamaño máximo de las inversiones	12693 pb	16560 pb
Tamaño mínimo de las inversiones	83 pb	940 pb
Tamaño promedio de las inversiones	2007 pb	3909 pb
Total Inversiones	9	9
Nº Inversiones con tamaño menor a 2 Kb	7	5
Nº Inversiones con tamaño menor a 1 Kb	6	1
Nº Inversiones con tamaño menor a 500 pb	5	0
Tamaño máximo de los puntos de rotura	79 pb	6423 pb
Tamaño mínimo de los puntos de rotura	Entre dos bases	9 pb
Tamaño promedio de los puntos de rotura	10 pb	2037 pb

En general entre las inversiones polimórficas validadas no se encuentran inversiones de tamaños mayores a las 16 Kb debido a que algunas inversiones de mayor tamaño no han podido ser validadas experimentalmente por las limitaciones para amplificar productos de tamaño mayor a sus puntos de rotura en el caso del protocolo de *PCR* estándar o a la limitación para encontrar enzimas con dianas de corte fuera de las duplicaciones segmentales y dentro de la región invertida.

No obstante no solo no se han validado experimentalmente inversiones con puntos de rotura localizados en RIs, sino que están representados ambos tipos, ya que algunas inversiones no se validaron por su tamaño menor a 1 Kb y no tener efectos posicionales sobre genes. El resultado de nuestro estudio nos ha dejado 9 inversiones polimórficas de cada tipo. Por otra parte, la validación de la totalidad de las inversiones potencialmente polimórficas nos hubiese generado un mayor conjunto de inversiones fiables de cada tipo. Sabemos que la proporción de inversiones con puntos de rotura en RIs está subestimada, debido a la eliminación de duplicaciones segmentales en la estrategia de secuenciación *shotgun*, por lo que una parte de las inversiones con puntos de rotura en estas regiones no se han descubierto o validado experimentalmente, y esto implica, que al tratarse de las inversiones de mayor tamaño, se estén subestimando también los efectos sobre genes del conjunto de inversiones. Además como ya hemos visto, el conjunto de inversiones original [Levy et al. 2007] no representa la totalidad de las inversiones entre *HuRef* y el genoma de Referencia.

En cuanto a la localización de las inversiones de ambos grupos en los diferentes cromosomas, no hay un patrón visible, como podemos ver en la **Figura 4.1**. Las inversiones con puntos de rotura no localizados en RIs no están determinadas por la homología de la secuencia y tienen más libertad para localizarse en cualquier parte del genoma, mientras que las inversiones con puntos de rotura en RIs se forman mediante duplicaciones segmentales. No obstante, un análisis de la distribución requeriría un conjunto de inversiones que representaran la totalidad de inversiones en el genoma a analizar y no es el caso de *HuRef*. Además, la distribución de las inversiones polimórficas con puntos de rotura en RIs no sería completa aunque se hubiesen validado todas las inversiones, ya que como hemos visto anteriormente, se eliminan regiones duplicadas durante el proceso de ensamblaje de los genomas secuenciados por *shotgun*. Por último, en ambos grupos de inversiones se espera una distribución que cubra todos los cromosomas, porque las inversiones con puntos de rotura localizados en RIs no solo se forman mediante duplicaciones segmentales sino que pueden formarse mediante elementos móviles, que están localizados por todo el genoma. En el siguiente apartado se discute más profundamente sobre los mecanismos de formación.

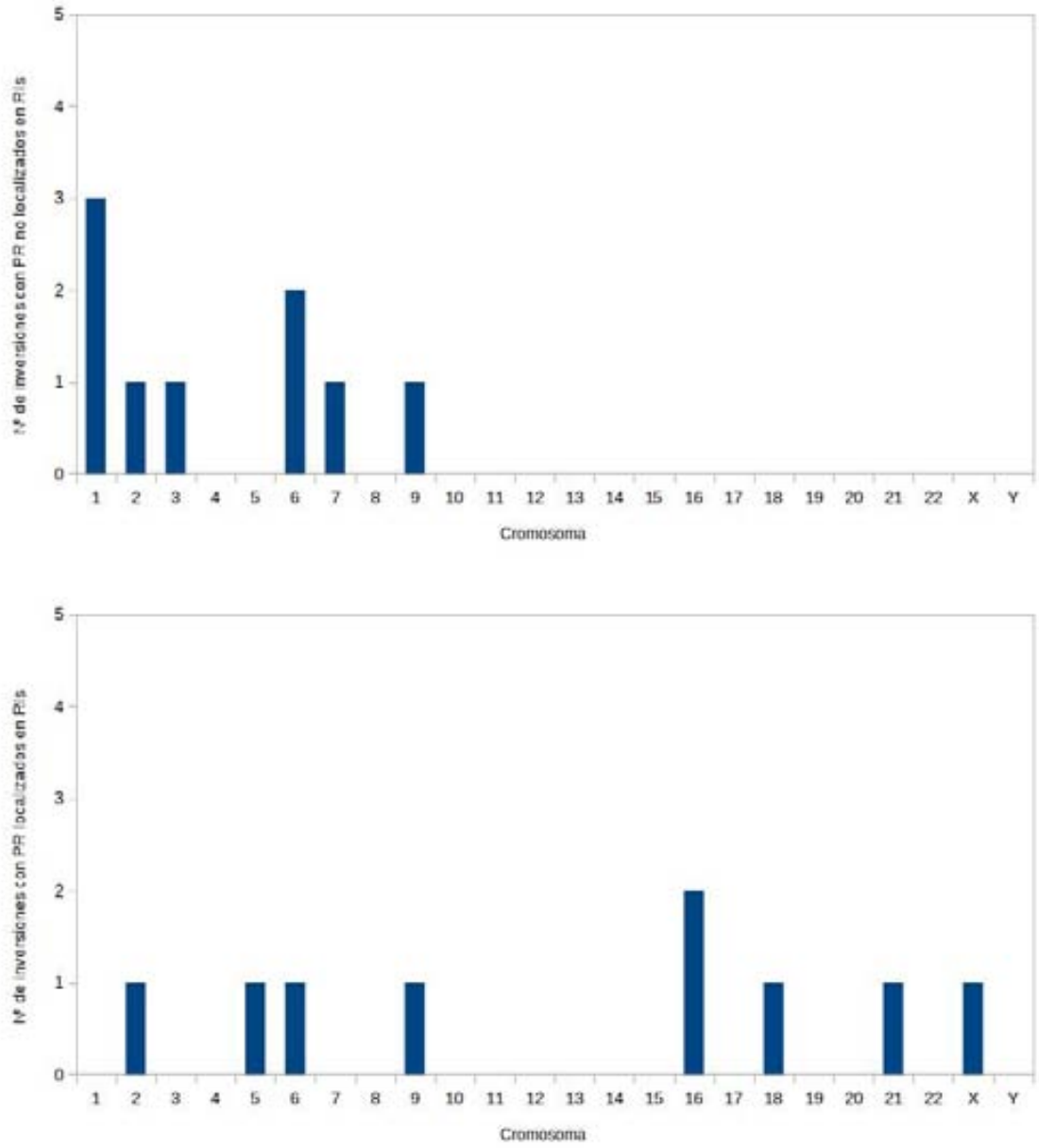

Figura 4.1: Comparación de la localización de inversiones con puntos de rotura localizados y no localizados en RIs. Se muestra el número de inversiones polimórficas localizadas en cada cromosoma. En la parte de arriba se muestran las inversiones con puntos de rotura no localizados en RIs y abajo las inversiones con puntos de rotura localizados en RIs.

Volviendo a la **Figura 4.1**, podemos ver que las inversiones con puntos de rotura no localizados en repeticiones invertidas se han agrupado en los primeros cromosomas. Este hecho aunque no tiene más importancia podría deberse a que los primeros cromosomas tienen un tamaño mayor y si la distribución de estas inversiones es aleatoria, tienen más probabilidades de alojar un número mayor de inversiones. De todas formas dos factores impiden que podamos sacar conclusiones. En primer lugar como ya hemos comentado antes, el número de inversiones es demasiado pequeño para analizar si su localización

está dirigida o es al azar, ya que otras inversiones de este tipo no han sido validadas por tener tamaños menores a 1 Kb y no estar localizadas cerca de genes y en segundo lugar. En segundo lugar, se han seleccionado mediante los criterios anteriores las inversiones a validar y por lo la localización de éstas es un producto de la esa selección.

4.3.1 Mecanismos de formación homólogos vs. no homólogos

Los resultados de nuestro estudio han aportado información sobre los mecanismos de formación de las inversiones. Dejan huellas en los puntos de rotura o las secuencias flanqueantes, y mediante su interpretación se puede determinar el tipo de mecanismo responsable de la formación de cada inversión. En el caso de las repeticiones invertidas, el mecanismo homólogo responsable es *NAHR*. Como hemos visto en la literatura, es el mecanismo más frecuente entre las variantes estructurales de mayor tamaño [Onishi-Seebacher and Korbel. 2011] y nuestros resultados también muestran esa tendencia. Por el contrario, las variantes de menor tamaño suelen ser consecuencia de los mecanismos que no usan secuencias homólogas o bien usan secuencias micro-homólogas. A diferencia de los mecanismos homólogos, es difícil determinar qué procesos exactamente están detrás de cada inversión, ya que las huellas que dejan son comunes entre varios mecanismos, incluso entre los de tipo replicativo como *FoSTeS/MMBIR* y los no replicativos como *NHEJ* o *MMEJ*.

En nuestro estudio intentamos asignar un posible mecanismo de formación a cada inversión polimórfica, basándonos en la presencia o ausencia de secuencias micro-homólogas e inserciones/deleciones. Se atribuyó el mecanismo *NHEJ* a las inversiones sin secuencias micro-homólogas ni inserciones/deleciones, mientras que en presencia de secuencias de micro-homología no se pudo distinguir entre los mecanismos *NHEJ* y *MMEJ*, ya que ambos las pueden usar. En presencia de inserciones/deleciones en los puntos de rotura, se asignó el mecanismo replicativo *FoSTeS/MMBIR* debido a su tendencia a asociarse a la formación de variantes complejas. Por el momento no se ha podido distinguir qué mecanismo es el causante de una variante estructural simple a partir de las secuencias micro-homólogas, que son aparte de las inserciones/deleciones las únicas señales de que disponemos [Hastings et al. 2009]. Los resultados obtenidos ponen de manifiesto la necesidad de investigar las diferentes vías de formación de variantes y las diferencias entre ellas para poder comprender mejor en el futuro cómo afectan estos mecanismos al paisaje genómico.

Tal como hemos dicho, podemos pensar en que los mecanismos homólogos, en concreto *NAHR*, forman variantes de mayor tamaño porque las RIs implicadas están separadas en el genoma, seguramente a consecuencia de su mecanismo de formación, o porque es necesario un espaciamiento para formar un lazo y poder aparear las RIs, como es el caso de *NAHR* intracromátida [Stankiewicz and Lupski. 2002]. Por el contrario, los

mecanismos no homólogos basados en la reparación de *DSBs*, como es el caso de *NHEJ*, reparan fragmentos de ADN mucho más pequeños a partir de la complementariedad de ambas cadenas. Además es lógico pensar que los mecanismos homólogos formen inversiones cromosómicas con mayor frecuencia, ya que para la formación a partir de *NHEJ* o *MMEJ* se requiere que ocurran dos *DSBs* en una misma región y que se dé un error en la reparación de manera que la secuencia entre ambos cambie de orientación; mientras que la recombinación entre secuencias homólogas ocurre como mínimo en cada meiosis ya que por cada cromosoma se forma al menos un entrecruzamiento necesario para su correcta segregación, lo que hace mucho más probable que se dé un error en el apareamiento de RIs. La excepción pueden ser los mecanismos no homólogos replicativos, *FoSTeS/MMBIR*. Se basan en los errores en el apareamiento de secuencias en la horquilla de replicación (bien sea durante la propia replicación o durante la reparación del DNA) y esto hace que ocurran con mayor frecuencia que los dos *DSBs* y el error de reparación necesarios en *NHEJ/MMEJ*. La frecuencia de formación de inversiones de estos mecanismos podría ser comparable a la *NAHR* dado que la tasa de replicación del ADN es proporcional a la tasa de división meiótica o mitótica. Los resultados de nuestro estudio corroboran la importancia de los mecanismos replicativos *FoSTeS/MMBIR*, porque se han descubierto inserciones y deleciones en los puntos de rotura en 7 de las 9 inversiones con puntos de rotura no localizados en RIs. Esto indica que los mecanismos replicativos forman inversiones con mucha más frecuencia de lo que se había pensado hasta ahora. Además, estos mecanismos no sólo forman las variantes complejas de menor tamaño sino que pueden ser responsables de variantes de cualquier tamaño, incluidas las inversiones más grandes con puntos de rotura sencillos [Gu et al. 2008]. En nuestro análisis, la inversión *HsInv0063*, cumple con las características para haber sido formada por *FoSTeS/MMBIR*. En primer lugar, sus puntos de rotura no están localizados en repeticiones invertidas y en segundo lugar se detectó una inserción de 5216 pb y una deleción de 8 pb tomando como referencia el genoma de Referencia, pasando a ser una variante compleja formada por una inversión, una inserción y una deleciones. Su tamaño es de 12.7 Kb, siendo una de las inversiones de mayor tamaño del conjunto y un ejemplo de que los mecanismos no homólogos replicativos pueden formar variantes de todos los tamaños. Finalmente cabe destacar que hemos descartado el mecanismo denominado Cromotripsis ya que algunos autores lo consideran una generación masiva de variantes complejas que se da en procesos cancerígenos por parte de *FoSTeS/MMBIR* [Liu et al. 2011].

4.3.2 La variación estructural comparte los puntos de rotura: ecología genómica

Las variantes estructurales complejas se caracterizan por ser agrupaciones de variantes estructurales que comparten un mismo punto de rotura [Quinlan and Hall. 2012]. Como su nombre indica necesitan de un análisis muy detallado para su detección y validación. Por eso, a pesar de que se detectaron este tipo de agrupaciones junto con las primeras

variantes estructurales, no han sido caracterizadas al mismo nivel. Los mecanismos no homólogos replicativos son los únicos que pueden explicar su formación a partir de los sucesivos errores en el apareamiento de las secuencias en las horquillas de replicación. Los resultados del análisis de los puntos de rotura de las inversiones realizados en este estudio, muestran que la gran mayoría de las inversiones con puntos de rotura no alojados en RIs tienen inserciones/deleciones compartiendo al menos un punto de rotura, por lo que se trata de variantes estructurales complejas. Además en muchos casos estos puntos de rotura contienen secuencias micro-homólogas. Esto hace que en nuestro estudio se hayan analizado más variantes complejas en las que están involucradas inversiones cromosómicas que en ningún otro estudio hasta el momento, y aporta información sobre su frecuencia en el genoma humano. En otro estudio realizado por Conrad y colaboradores en el año 2010, se analizaron 324 puntos de rotura de *CNVs* en todo el genoma y se clasificaron aproximadamente un 5% como correspondientes a variantes complejas [Conrad et al. 2010]. En cambio, los resultados de nuestro estudio muestran una proporción de inversiones con puntos de rotura no localizados en RIs que forma parte de variantes complejas del 78%. Estos resultados hacen pensar que la importancia de las variantes complejas en el genoma humano es mayor de lo esperado y por extensión, los mecanismos no homólogos replicativos que están detrás de su formación.

El agrupamiento de variantes puede dar lugar a efectos fenotípicos múltiples, y evidentemente más difíciles de caracterizar que las variantes simples [Quinlan and Hall, 2012]. Las variantes complejas permiten que se den cambios muy importantes en la secuencia de manera muy rápida, por lo que se da una evolución puntuada de la secuencia [Quinlan and Hall, 2012]. Por ejemplo permiten pasar en un sólo evento mutagénico de una secuencia única con orientación estándar a una secuencia invertida con la duplicación o deleción de las secuencias adyacentes. Por el contrario la acumulación de variantes estructurales a lo largo del tiempo en una zona concreta hace que la secuencia evolucione de una manera más continua.

Por otra parte, la ocurrencia de los errores en los mecanismos no homólogos podría no ser cosa sólo del azar. Algunos autores sugieren que la arquitectura genómica local podría ser un estímulo para que tanto *NHEJ/MMEJ* como *FoSTeS/MMBIR* cometan errores que dan lugar a variantes estructurales [Gu et al. 2008]. Se relacionaría indirectamente con la presencia de palíndromos y estructuras cruciformes. En ese sentido, los resultados de nuestro estudio son un buen punto de partida para analizar la colocalización de las variantes formadas por mecanismos no homólogos con elementos que puedan inducir a errores de reparación y replicación, como pueden ser las RIs. Las repeticiones invertidas también pueden formar estructuras al aparear con sus secuencias homólogas cuando se encuentran en cadena sencilla, por ejemplo durante la replicación. En el caso de encontrar una relación positiva entre ambos, las duplicaciones segmentales y otros tipos de repeticiones no sólo determinarían la localización de las inversiones más grandes formadas por *NAHR*, sino que influirían en la acumulación de inversiones con puntos de

rotura sencillos a su alrededor y la variación estructural se tendería a formar siempre en unas determinadas regiones que son reutilizadas. Además, el número de inversiones que restan por descubrir sería mayor debido a la baja calidad de las regiones duplicadas en los genomas ensamblados.

4.3.3 Implicaciones del origen recurrente de las inversiones

Tradicionalmente se ha considerado que las inversiones cromosómicas tienen un origen único o monofilético, es decir se forman una única vez [Krimbas y Powell 1992]. Sin embargo, nuestro estudio aporta ejemplos de inversiones que se han generado más de una vez en humanos usando los mismos puntos de rotura, lo que implica que la secuencia ha cambiado de orientación con cada evento. Se las denomina inversiones con origen recurrente o polifilético y también se pueden dar en especies diferentes. En nuestro estudio se analizó el origen de las inversiones basándonos en el desequilibrio de ligamiento entre los alelos de la inversión y los alelos de los *SNPs* localizados dentro de la región invertida. También se analizaron los haplotipos formados por ambos alelos para cada orientación y se genotiparon varios chimpancés y gorilas para analizar un posible polimorfismo entre especies como evidencia extra de recurrencia.

En el momento en que se genera una inversión cromosómica, la diferente orientación entre la secuencia no invertida y la invertida impide la recombinación en los individuos heterocigotos, de manera que la secuencia invertida actúa como un haplotipo independiente y captura una combinación de alelos de los *SNPs* localizados dentro, por lo que segregan siempre con la inversión y se genera un bloque de desequilibrio de ligamiento. Con el paso del tiempo, ambos haplotipos acumulan cambios y se diferencian entre sí. Esta huella en la variación nucleotídica y haplotípica caracteriza a las inversiones con origen único. En cambio, en las inversiones recurrentes se reinvierte la secuencia invertida o se invierte de forma independiente de manera que ambas secuencias vuelven a recombinar en su totalidad, por lo que se rompe el desequilibrio de ligamiento entre los alelos de los *SNPs* de la región anteriormente invertida. Es cierto que en las inversiones de gran tamaño se da cierta recombinación lejos de los puntos de rotura, pero el efecto de la recurrencia es mucho mayor porque se restablece la recombinación en toda la región. En estas inversiones no encontramos alelos de *SNPs* propios de cada haplotipo, sino que debido al cambio de orientación que ha sufrido la secuencia, los alelos son compartidos en ambos haplotipos.

En nuestro estudio se detectaron 3 inversiones con origen recurrente y junto con otros estudios en humanos [Antonacci et al. 2009] [Aguado et al. 2014], los resultados dan más importancia a este fenómeno. Por el momento sólo un estudio ha analizado este tipo de inversiones en más de una población humana y concluyó que 5 de las 6 inversiones analizadas con puntos de rotura en *LCRs* son recurrentes [Antonacci et al. 2009]. En el

estudio de Aguado y colaboradores [Aguado et al. 2014] se analizó su origen usando sólo individuos de población Europea y se determinó la recurrencia de 13 de las 17 inversiones analizadas. Al igual que en estos estudios, las 3 inversiones tienen puntos de rotura localizados en RIs, y se forman por recombinación entre ellas, mediante el mecanismo homólogo de *NAHR* [Aguado et al. 2014]. El análisis de la variación nucleotídica y haplotípica nos reveló que estas inversiones tienen características comunes. En primer lugar tienen *SNPs* compartidos entre ambas orientaciones, es decir, se encuentran los mismos alelos de varios *SNPs* en la secuencia con orientación estándar y en la secuencia de orientación invertida. Además no tienen *SNPs* fijados, es decir, *SNPs* en que un alelo segrega siempre junto al alelo estándar de la inversión y el otro alelo junto al alelo invertido. Por último, en el análisis de los haplotipos de este tipo de inversiones se han detectado grupos de haplotipos compartidos donde cromosomas portadores y no portadores de la inversión tienen los mismos cambios nucleotídicos en la región invertida. En nuestro caso, el menor tamaño de las inversiones provoca que el número de alelos de *SNPs* compartidos sea menor que en las inversiones analizadas en el estudio de Aguado y colaboradores.

Evidentemente la formación de una nueva inversión exactamente en el mismo lugar donde se formó la inversión anterior tendría las mismas consecuencias, pero la probabilidad de que se produzcan dos mutaciones independientes en el mismo lugar del genoma es muy baja. El único fenómeno que puede producir que se compartan los alelos de los *SNPs* entre ambas ordenaciones de la secuencia es que se copie la secuencia en los individuos heterocigotos y el mecanismo responsable sería la recombinación. No obstante uno de los efectos más conocidos de las inversiones es la inhibición de la recombinación en la región invertida entre ambas secuencias en los individuos heterocigotos. Aun así las inversiones de gran tamaño pueden formar un bucle o lazo mediante el que se podría dar recombinación en la región invertida lejos de los puntos de rotura por doble entrecruzamiento. En el caso de la inversión en el cromosoma *8p23*, la inversión polimórfica más grande del genoma humano con un tamaño de unas 4.5 Mb, se especula con esa posibilidad [Salm et al. 2012]. Se han realizado estudios con espermatozoides humanos que revelan que la incidencia de la recombinación está relacionada con el tamaño de la región invertida. En estos estudios se ha visto que el tamaño mínimo de las inversiones que permiten que se dé recombinación lejos de sus puntos de rotura que afecta a los gametos es de 100 Mb y que es necesario que la inversión ocupe al menos un 50% del tamaño del cromosoma [Anton et al. 2005]. Por lo tanto podemos descartar la formación de entrecruzamientos dobles en nuestro conjunto de inversiones de tamaños muy inferiores.

Otra forma de copiar la secuencia entre ambas orientaciones podría ser la conversión génica. Se trata de una forma de recombinación en que se copia la información de un cromosoma a otro sin que el primero resulte alterado. En mamíferos se tiene constancia de este fenómeno que en general no copia secuencias de más de 1 Kb [Chen et al. 2007].

Por lo tanto, la conversión génica puede explicar la existencia de alelos de *SNPs* compartidos entre ambas ordenaciones sin necesidad de mediar recurrencia. De hecho en estudio realizado por Aguado y colaboradores en el año 2014 se detectaron alelos compartidos en inversiones que aparentemente tenían un origen único [Aguado et al. 2014]. No obstante, la conversión génica no puede explicar la ausencia de alelos fijados entre ambas orientaciones. Si tomamos como ejemplo una inversión de origen único que fija alelos en ambas orientaciones y suponemos que la conversión génica copiase algunos alelos de la secuencia de ordenación estándar a la secuencia de ordenación invertida, lo que obtendríamos sería una inversión con alelos fijados y alelos compartidos. Por lo tanto la conversión génica no puede explicar la ausencia de alelos fijados en las inversiones con alelos compartidos y eso demuestra que se trata de inversiones recurrentes. Para obtener el patrón de variación nucleotídica de las inversiones recurrentes la conversión génica tendría que copiar todos los alelos de los *SNPs* a lo largo de la región invertida, que en algunos casos se extiende varias kilo bases. Además existen otras evidencias de que se trata de eventos recurrentes, como por ejemplo los resultados del análisis haplotípico. Estas inversiones muestran grupos de haplotipos correspondientes a la orientación invertida que en vez de localizarse en los arboles de haplotipos junto al resto de haplotipos de su misma orientación, por la similitud de sus alelos con los haplotipos correspondientes a la otra orientación se encuentran agrupados junto a ellos. Podemos ver un ejemplo en la **Figura 5** en el apartado de resultados. Todo esto sugiere que se trata de inversiones recurrentes.

Por último, una de las principales ventajas de este trabajo es que hemos podido comparar los patrones de variación nucleotídica y haplotípica de las inversiones con puntos de rotura localizados en RIs y de las inversiones con puntos de rotura no localizados en RIs, que debido a su mecanismo de formación no homólogo tienen un origen único. Ambos grupos de inversiones provienen del mismo método de detección y de la comparación de los mismos genomas. Los resultados contrastan claramente, ya que las inversiones formadas por mecanismos no homólogos no tienen alelos de *SNPs* compartidos dentro de la región invertida y tampoco tienen grupos de haplotipos compartidos entre ordenaciones. Los alelos de *SNPs* en sus regiones invertidas están fijados y los grupos de haplotipos de uno y otro alelo de la inversión claramente diferenciados en los arboles de haplotipos. Este hecho es una evidencia más de la recurrencia de las inversiones con puntos de rotura localizados en RIs. En comparación con el trabajo de Aguado y colaboradores, el tamaño de las inversiones analizadas en nuestro estudio es menor y provoca que el número de alelos de *SNPs* compartidos sea también menor.

Por otra parte, hay que mencionar que se han tomado precauciones antes de designar estas inversiones como recurrentes. En primer lugar nos aseguramos que no se trataba de errores de genotipación de las inversiones porque se repitieron los experimentos de *PCR* para aquellos individuos que indicaron recurrencia y se comprobó que ambos puntos de rotura indicasen el mismo genotipo. En segundo lugar, la identidad de los individuos fue

confirmada por microsatelites para asegurarnos que los genotipos de las inversiones y los *SNPs* provenían de la misma persona. En tercer lugar se repitió el análisis de la variación nucleotídica y haplotípica usando la información de *SNPs* de dos conjuntos de datos diferentes, los correspondientes al proyecto de los 1000 Genomas y los correspondientes al proyecto HapMap, obteniendo los mismos resultados. Finalmente, los resultados del análisis de la variación nucleotídica que se basan en los genotipos de los *SNPs* y los resultados del análisis de los haplotípos que se basan en datos faseados coinciden perfectamente. Además en el caso de los haplotipos construidos a partir de los datos HapMap se usó la información de los tríos familiares, que le dan más robustez al proceso de faseo de los datos. Al igual que en el estudio de Aguado y colaboradores [Aguado et al. 2014], se genotiparon 4 chimpancés y 2 gorilas para determinar si las inversiones también son polimórficas en otras especies o bien muestran diferentes orientaciones, ya que se trataría de otra evidencia de recurrencia porque es poco probable que se mantenga un polimorfismo entre especies distintas durante tanto tiempo. A diferencia del estudio de Aguado y colaboradores, nuestros resultados fueron negativos y no se detectó ninguna inversión polimórfica o con distintos alelos entre especies, lo que indica que los eventos de recurrencia se han producido en la especie humana. No obstante, el tamaño de la muestra es bastante limitado y habría que confirmar estos resultados con más individuos.

Hay que tener en cuenta que el número de inversiones recurrentes podría ser mayor ya que han resultado ser recurrentes 3 de las 9 inversiones analizadas con puntos de rotura en RIs (*HsInv0030*, *HsInv0055 y HsInv0069*), pero debido a que sólo se usaron individuos europeos en su análisis, únicamente se han detectado los eventos de recurrencia ocurridos en esta población y esto constituye una limitación a la hora de concluir cuantas de estas 9 inversiones son recurrentes en humanos. En un estudio que se está realizando sobre la recurrencia de las inversiones polimórficas del genoma humano presentes en la base de datos *InvFEST* [Martínez-Fundichely et al. 2013], entre las cuales se encuentran algunas de las inversiones de nuestro estudio, se han obtenido genotipos para 7 poblaciones humanas de distintos continentes correspondientes al proyecto HapMap mediante una nueva técnica basada en el MLPA [S. Villatoro and M. Cáceres, resultados no publicados]. Por el momento se han detectado eventos de recurrencia de la inversión *HsInv0072* en 3 poblaciones y se ha confirmado la recurrencia de las inversiones *HsInv0030* y *HsInv0055* en al menos 3 poblaciones distintas a la Europea [S.Villatoro y M. Cáceres, resultados no publicados], mientras que la inversión *HsInv0069* no se ha podido analizar aún en estos individuos. En este sentido es necesario un análisis a mayor escala donde se incluyan individuos de todas las poblaciones posibles para poder conocer el impacto global de este tipo de inversiones a nivel del genoma humano.

En el apartado anterior hemos visto como en el genoma se reutilizan los puntos de rotura de las variantes estructurales o las zonas en que estos están ubicados. Siguiendo en la misma línea, las variantes formadas por *NAHR* entre RIs, en este caso inversiones, también reutilizan sus puntos de rotura aunque de forma diferente. Mientras las variantes

estructurales complejas usan un mismo punto de rotura para varias variantes estructurales, las variantes recurrentes los reutilizan para revertir y reaparecer. En cualquier caso, se generan pocos puntos de rotura en zonas nuevas.

Por otro lado, la recurrencia no solo implica que sea más difícil analizar la historia evolutiva de estas variantes, sino que la rotura del desequilibrio de ligamiento implica que no puedan ser genotipadas a partir de *SNPs* y que sea necesaria la genotipación experimental [Aguado et al. 2014]. Además esto conlleva que en los estudios de asociación a nivel de todo el genoma basados en genotipos de *SNPs* se pierden los efectos fenotípicos de las inversiones recurrentes. No solo eso, sino que puede tratarse de un porcentaje importante de las inversiones si tenemos en cuenta que las inversiones de mayor tamaño tienen sus puntos de rotura localizados en RIs y están formadas por *NAHR*, características que las hacen candidatas a ser recurrentes. Por lo tanto, pueden haberse obviado y se siguen obviando los efectos de la mayoría de inversiones de gran tamaño del genoma humano en este tipo de estudios de asociación.

4.4 La frecuencia de las inversiones indica diferencias entre poblaciones

4.4.1 Diferentes factores afectan las frecuencias obtenidas de los distintos métodos de genotipación

En el apartado de resultados se muestran las diferencias entre poblaciones de las frecuencias alélicas que se han obtenido de la genotipación de distintos individuos de varias poblaciones humanas pero hay que tener en cuenta distintos factores específicos de cada método que afectan al proceso de genotipación de los individuos y por lo tanto pueden afectar a la fiabilidad de las frecuencias.

Nuestro estudio ha tenido como objetivo obtener genotipos fiables en el proceso de genotipación, y para ello se han aplicado todas las medidas necesarias dentro de las limitaciones por ejemplo económicas que tiene hacer estudios usando individuos de varias poblaciones. En la genotipación por *PCR* e *iPCR* se diseñaron los experimentos para que sean *multiplex*, es decir, que en una misma reacción de *PCR* se amplifiquen los fragmentos correspondientes a los alelos de ambas orientaciones de la secuencia. De esta manera si falla la reacción no obtendremos ningún fragmento, mientras que un experimento independiente para ambos alelos puede conllevar la genotipación errónea de individuos heterocigotos como homocigotos en caso de que falle una de las reacciones de *PCR*. Lamentablemente el diseño de cebadores en el genoma humano no es trivial debido a la repetitividad de determinadas zonas, que como hemos visto están relacionadas con las inversiones. En aquellos casos en que no fue posible el diseño *multiplex*, se repitió la genotipación de cualquier individuo que no fuese clara. Es decir, se repitieron los genotipos en que hubiese sospecha de no estar amplificando algún producto, por ejemplo

en el caso de amplificación de una banda de intensidad muy baja que no permitiese la genotipación correcta del individuo. Este control se llevó a cabo en la genotipación de los 90 individuos de población Europea correspondientes al proyecto HapMap y permitió obtener unas frecuencias muy fiables para esta población. Evidentemente las frecuencias obtenidas no son definitivas, ya que se genotipó sólo una parte del total de la población Europea, pero sí que nos dan una idea bien aproximada.

En el caso de la genotipación bioinformática también se dan circunstancias que nos pueden llevar a errores en la estima de la frecuencia de las inversiones. En primer lugar, debido a la secuenciación de baja cobertura se dispuso de pocos *reads* que cubrieran las regiones de los puntos de rotura en cada individuo, y a menor número de *reads* mapeados mayor probabilidad de detectar un solo alelo y subestimar el número de individuos heterocigotos para las inversiones. En segundo lugar, existe un sesgo en la detección, ya que ambos alelos no tienen las mismas probabilidades de ser detectados y es más probable detectar el más frecuente. Para evitar que estos factores alterasen la correcta genotipación de los individuos secuenciados en el proyecto de los 1000 Genomas, se usó *svgem* [Lucas-Lledó et al. 2014]. Se trata de un programa estadístico que aplica un algoritmo de esperanza-maximización que permite determinar la incertidumbre de cada genotipo teniendo en cuenta el sesgo de detección de ambos alelos. De esta manera se estima cual es el genotipo más probable de cada individuo. Además usa un algoritmo de máxima verosimilitud para estimar las frecuencias alélicas para las diferentes poblaciones. A pesar de esto, debido a la baja cobertura de secuenciación, muchos individuos no pueden ser genotipados por la falta de *reads* que mapeen en los puntos de rotura. Para que esto no afecte a las frecuencias poblacionales, se estableció un número mínimo de 10 individuos genotipados para considerar una frecuencia poblacional válida. Por lo tanto, la aplicación de *svgem* y de este umbral hacen que las frecuencias obtenidas de la genotipación *in silico* sean fiables.

Por último se han genotipado a través de *SNPs* marcador todos los individuos provenientes de las 14 poblaciones humanas en el proyecto de los 1000 Genomas. En la determinación de estos *SNPs* se usaron los genotipos de los individuos de la población Europea determinados experimentalmente y solo los *SNPs* fijados, con valores de r^2 de 1 y por lo tanto en completo desequilibrio de ligamiento con la inversión en esta población, para obtener los *SNPs* marcador globales para todas las poblaciones. Las frecuencias que se obtuvieron en población Europea son muy fiables ya que los genotipos de los individuos para estos *SNPs* coincidieron con los de las inversiones. En las inversiones genotipadas por mapeo de *reads* en los puntos de rotura, se usaron los genotipos predichos en las distintas poblaciones para asociarlos con los de los *SNPs* marcador en población Europea y así se encontraron los globales de manera manual. En este proceso se detectaron errores esporádicos de genotipación de los *SNPs* en individuos de distintas poblaciones por parte del consorcio encargado del proyecto [1000 Genomes Project Consortium. 2012].

Para el resto de inversiones en las que no disponemos de genotipos para los individuos de las 14 poblaciones del proyecto de los 1000 Genomas, se trató de determinar de la manera más fiable posible los *SNPs* marcadores globales. Para ello se analizó la correlación de los genotipos de los *SNPs* marcador en población Europea por parejas de *SNPs* en los 1092 individuos disponibles. Las parejas de *SNPs* marcador Europeos con valores de r^2 superiores a 0.99 en todas las poblaciones, es decir, con un desequilibrio de ligamiento completo o casi completo entre ellos, se seleccionaron como *SNP* marcadores globales. La idea es que cuando no fue posible encontrar *SNPs* totalmente ligados entre ellos, se seleccionaron los *SNPs* con valores de r^2 más altos posibles, indicando que en un porcentaje muy bajo, del 1% de los individuos no segregan conjuntamente. Hay que comentar que solo se pudo analizar el desequilibrio entre los alelos de *SNPs* y de las inversiones en la población Europea, de la que disponemos de genotipos provenientes de experimentos de *PCR;* por lo que de todas formas no podemos saber si segregan realmente con la inversión o no. Aún así se ha intentado buscar los *SNPs* marcador con el mayor desequilibrio de ligamiento posible con la inversión y no se han considerado como válidos valores de r^2 inferiores a 0.99. Esto contrasta con otros estudios donde se catalogan como *SNPs* marcador aquellos que tienen valores de r^2 superiores a 0.8, donde el 20% de los individuos tienen alelos para un determinado *SNP* que no segregan con el alelo de la inversión y por lo tanto, las frecuencias estimadas a partir de ellos no son fiables porque pueden variar de manera importante dependiendo de los genotipos de ese 20% de individuos [Pang et al. 2013]. A pesar de las medidas que se aplicaron para obtener las frecuencias más fiables posibles, el hecho de no relacionar directamente los genotipos de *SNPs* con los genotipos de las inversiones conlleva el riesgo de que la asociación que encontramos entre parejas de *SNPs*, que son *SNPs* marcadores en población Europea, no se dé porque ambos *SNPs* están asociados con la inversión en todas las poblaciones. En otras palabras, podemos estar seguros de que los *SNPs* marcador en Europeos están asociados con la inversión, pero sin los genotipos para la inversión en el resto de poblaciones, no podemos saber si también se trata de *SNPs* marcador en el resto de poblaciones o bien estos *SNPs* están asociados entre ellos pero no con la inversión. Por lo tanto, que una pareja de *SNPs* marcadores en población Europea estén asociados en el resto de poblaciones no indica necesariamente que se trate de *SNPs* marcadores a nivel global. Por ese motivo se seleccionaron los *SNPs* que por su localización tienen más probabilidades de estar asociados con la inversión. En concreto, se seleccionaron los más cercanos a los puntos de rotura, priorizando los localizados dentro de la inversión, ya que por los efectos de inhibición de la recombinación de las inversiones, es la localización más indicada para contener *SNP* marcador a nivel poblacional y global. En los casos en que estos pares de *SNPs* ligados están localizados lejos de los puntos de rotura, se descartaron, y se utilizaron los *SNPs* marcador en población Europea para obtener frecuencias aproximadas. En este caso también se mantuvieron los criterios de localización dentro de la inversión y cercanía a los puntos de rotura.

Por otra parte, es importante comentar que se produce una variación en las frecuencias poblacionales de las inversiones debido al número de individuos genotipados mediante cada método. En el caso de la genotipación experimental de la población Europea, el número de individuos genotipados es similar al número de individuos con información de *SNPs* marcador. La diferencia está en la genotipación bioinformática, que como ya hemos visto, debido a la baja cobertura de secuenciación no puede establecer un genotipo para muchos individuos. En el resto de poblaciones, la variación de las frecuencias entre las obtenidas por genotipación bioinformática y por *SNPs* marcador es visible en algunos casos, donde la diferencia en el número de individuos genotipados puede ser de hasta unos 80 individuos más en la genotipación por *SNP* marcador. Idealmente, la secuenciación de alta cobertura de los individuos del proyecto de los 1000 Genomas permitiría la genotipación bioinformática de todos los individuos y se corregirían estas diferencias.

Por último cabe destacar que tal y como hemos visto en apartados anteriores, la genotipación a través de *SNPs* marcador solamente es fiable en inversiones de origen único, donde los alelos de los *SNPs* están fijados o asociados con valores de r^2 superiores a 0.95 con los alelos de la inversión. Esto excluye a las inversiones con origen recurrente que por la recombinación que se da durante los eventos de recurrencia, no tienen alelos fijados de *SNPs* y por lo tanto no tienen *SNPs* marcador fiables según nuestro criterio. En ese sentido es importante comentar que puede darse el caso de inversiones que son recurrentes en unas poblaciones mientras que en otras no y por lo tanto pueden tener *SNPs* fijados en las poblaciones donde no se ha producido el evento de recurrencia. En nuestro caso podría ser que algunas de las inversiones con puntos de rotura en RIs que tienen un origen único en la población Europea sean recurrentes en otras poblaciones y por lo tanto la genotipación por *SNPs* marcador sea errónea.

Además las características de los puntos de rotura de estas inversiones no nos permiten la genotipación bioinformática, porque se basa en el alineamiento de *reads* en los puntos de rotura específicos de cada orientación y no pueden alinearse de manera única en las repeticiones invertidas donde se localizan los puntos de rotura de las inversiones recurrentes. Por estos motivos, la genotipación de las inversiones recurrentes o potencialmente recurrentes debido a las características de sus puntos de rotura, ha de ser experimental.

4.4.2 Explicación de la distribución poblacional actual de las inversiones a través de sus posibles efectos funcionales

En este estudio se genotiparon individuos de diferentes poblaciones humanas para todas las inversiones polimórficas. Se usaron tres métodos diferentes. Mediante experimentos de *PCR* e *iPCR* se genotiparon 10 individuos de 3 poblaciones procedentes del proyecto

HapMap, en concreto las poblaciones Europea, Africana y Asiática. También se genotiparon los 90 individuos pertenecientes a la población Europea provenientes del mismo proyecto. Los resultados de la genotipación nos permitieron tener una frecuencia muy aproximada de las inversiones polimórficas a nivel global y más fiable para la población Europea. En segundo lugar, para las inversiones con puntos de rotura sencillos se genotiparon bioinformáticamente individuos de 14 poblaciones humanas procedentes del proyecto de los 1000 Genomas [1000 Genomes Project Consortium. 2012], mediante el mapeo de *reads* provenientes de la secuenciación sobre los puntos de rotura de las inversiones. Además, a partir de la determinación de *SNPs* marcador para las inversiones polimórficas se estimaron sus frecuencias en esas mismas 14 poblaciones. Estos dos últimos métodos de genotipación permitieron obtener frecuencias de poblaciones humanas en todos los continentes y pudimos analizar de esta manera su distribución mundial, aunque sólo se pueden aplicar a una parte de las inversiones.

El resultado de la genotipación nos mostró diferentes frecuencias en las distintas poblaciones para las inversiones polimórficas, y en algunos casos una variación visible que se hace más grande cuando agrupamos las poblaciones por continentes. Para analizar estas diferencias y calcular qué parte de esta variación se debe a cómo las inversiones diferencian las poblaciones, se usó el índice de fijación o *Fst* y se calculó su significación mediante el *p*-valor. Como resultado obtuvimos valores de *Fst* significativos para 9 de las 10 inversiones polimórficas con frecuencias para las 14 poblaciones. Los valores de *Fst* mayores que 0 indican cierto grado de diferenciación interpoblacional, siendo el valor 1 el aislamiento total. En el proyecto de los 1000 Genomas se estimó la diferenciación basal entre poblaciones humanas, por ejemplo entre población Europea y Africana en valores de *Fst* de 0.07, de 0.08 entre población Africana y Asiática, y 0.05 entre población Asiática y Europea [Durbin et al. 2010]. A partir de estas estimas, se considera el valor de *Fst* aproximado de 0.1 como la diferenciación genética basal de las poblaciones y por eso lo usamos para distinguir las inversiones que muestran una diferenciación mayor de la esperada entre las poblaciones, que son las 5 que tienen valores de *Fst* superiores a 0.1. Consecuentemente estas inversiones son candidatas a estar implicadas en procesos adaptativos [Kirckpatrick et al. 2010].

No obstante la determinación de si realmente tienen efectos adaptativos en las poblaciones y si su posible implicación adaptativa se produce a través de los efectos de posición de sus puntos de rotura o bien a través de la inhibición de la recombinación en la región invertida en individuos heterocigotos no es trivial. En la literatura no hay prácticamente evidencias de fenotipos asociados a inversiones polimórficas causados por disrupción de genes por los puntos de rotura. En este aspecto, en el conjunto de inversiones analizado no se encuentran puntos de rotura con efectos de posición drásticos que expliquen la posible selección de las inversiones, pero sí posibles efectos más sutiles, ya que algunas se encuentran en intrones, cerca de algún gen, o invierten genes duplicados o parte de ellos, y podrían tener efectos sobre el *splicing* o la regulación

génica. Además, algunas inversiones que diferencian poblaciones humanas tienen tamaños muy pequeños del orden de 100 pb, de manera que es difícil pensar en que hayan capturado combinaciones de alelos favorables en una región tan pequeña. Por eso no podemos descartar que las inversiones puedan ser seleccionadas por los efectos de posición de sus puntos de rotura. Evidentemente no podemos hablar de selección porque no hemos realizado *tests* para demostrar su implicación. En cambio, sí que podemos decir que de haber selección no esperamos que sea fuerte, ya que no encontramos inversiones con efectos de posición drásticos como por ejemplo que rompan un exón o separen un gen en dos partes. En nuestro estudio hemos realizado el test de Hardy-Weingberg para todas las inversiones en la población Europea y todas las inversiones lo cumplen. Esto sólo nos permite corroborar que no se esperan efectos selectivos fuertes, pero el test de Hardy-Weingberg tiene poco poder para demostrar la ausencia de selección. Por eso no podemos descartar que se dé selección suave sobre las inversiones. A diferencia de especies como *D. melanogaster*, en humanos las inversiones polimórficas analizadas no forman clinas que evidencien su implicación en la adaptación local. Nuestros resultados nos muestran que algunas inversiones pueden estar detrás de la diferenciación entre poblaciones, y algunas son candidatas a haberse propagado gracias a la actuación de la selección natural, cosa que discutiremos en los siguientes apartados.

En este apartado tratamos de explicar cuál es la fuerza evolutiva que actuó o está actuando en la propagación de cada una de las inversiones polimórficas en las poblaciones humanas, a través de las frecuencias del alelo invertido, la estructuración de la población, el alelo ancestral y la especulación de los posibles efectos funcionales de las inversiones. La gran limitación es el no contar con *tests* de selección. Por lo tanto, no podemos sacar conclusiones sobre la implicación adaptativa de estas inversiones; pero sí que podemos analizar los casos interesantes, es decir, las inversiones que por su distribución poblacional pueden diferenciar a las poblaciones, e intentar determinar si la selección natural puede tener un papel en su distribución.

El análisis de la distribución poblacional se ha realizado sobre las frecuencias alélicas calculadas a partir de la genotipación bioinformática y de *SNPs* marcador, y para algunas inversiones se han comparado con las frecuencias obtenidas mediante genotipación experimental en 7 poblaciones de todo el mundo [S. Villatoro y M. Cáceres, resultados no publicados]. Las frecuencias alélicas son muy similares para todas las inversiones en las 6 poblaciones que son comunes a ambos análisis (*CEU*, *TSI*, *CHB*, *JPT*, *LWK* y *YRI*) y no varían por encima del 10%, con la excepción de la inversión *HsInv0068*, para la que no disponemos de frecuencia del alelo invertido en la población originaria de Kenya mientras que si se ha calculado en el estudio de Villatoro y colaboradores además de mostrar una variación mayor del 10% en población Europea; y las inversiones *HsInv0045* y *HsInv0058* que tienen una variación de más del 10% en población *LWK*. Hay que tener en cuenta que el número de individuos genotipados es diferente, lo cual explica que las frecuencias no sean exactamente iguales. Aun así los resultados muy similares obtenidos

en la comparación de ambas aproximaciones a sus distribuciones poblacionales reales, da fiabilidad a los datos obtenidos de manera no experimental.

En primer lugar, en base al índice *Fst* y las frecuencias poblacionales del alelo invertido (diferente al del genoma de Referencia) podemos diferenciar tres tipos de inversiones: las que no muestran diferencias aparentes en la frecuencia del alelo invertido en las poblaciones (una diferencia menor o igual al 10% en la frecuencia del alelo invertido entre cada continente y la frecuencia global) y no provocan una estructuración de la población por encima de la estructura basal, que como hemos visto anteriormente situamos en valores de *Fst* iguales o inferiores a 0.10; las inversiones que no provocan una estructuración de la población por encima de la estructura basal y que tienen diferencias en la frecuencias del alelo invertido entre las poblaciones (una diferencia superior al 10% en la frecuencia del alelo invertido entre cada continente y la frecuencia global); y por último las inversiones que provocan una estructuración por encima del valor basal y tienen diferencias en las frecuencias del alelo invertido entre poblaciones y continentes.

Las frecuencias del alelo invertido similares entre poblaciones y los valores de *Fst* iguales o inferiores a 0.1 que encontramos en el primer grupo de inversiones, nos sugieren a priori que posiblemente se han propagado en las poblaciones humanas de manera neutra, por lo que la fuerza evolutiva responsable sería la deriva genética, y sería la migración de la población humana del continente Africano hacia el resto de continentes lo que ha determinado su distribución actual, a través de los cambios de tamaño de la población. No obstante, al no haber realizado *test* de selección, no podemos descartar que estas inversiones hayan sido seleccionadas. Podría darse por ejemplo selección equilibradora que mantuviese el polimorfismo en las poblaciones y las frecuencias del alelo invertido similares. En este grupo se encuentran las inversiones *HsInv0031*, *HsInv0040* y *HsInv0058*.

En cuanto a los posibles efectos funcionales e implicación en enfermedades de este grupo de inversiones, hay que destacar que no interaccionan directamente con genes, es decir, sus puntos de rotura no están localizados en ningún gen. Aun así, las inversiones *HsInv0031* y *HsInv0058* están localizadas cerca de genes por lo que pueden tener efectos reguladores sobre ellos. En concreto, la inversión *HsInv0031* se encuentra 5665 pb aguas abajo del gen no codificante de proteínas *LOC400548* de función desconocida, por lo que debido a la distancia y posición relativa de ambos parece difícil explicar los posibles efectos de la inversión sobre la regulación del gen. La inversión *HsInv0058* se encuentra 6041 pb aguas abajo del gen *MUC22* y 11131 pb aguas arriba del gen *HCG22*. Al igual que la inversión *HsInv0031*, la posición respecto al gen *MUC22* hace que sea más difícil pensar en un efecto sobre su regulación. Sin embargo a pesar de que la distancia es mayor, es posible que pueda afectar a la regulación del gen *HCG22* porque se encuentra relativamente más cerca del punto de inicio de la transcripción. El gen *MUC22* codifica

para una proteína mucina que forma parte de la mucosidad y que se ha relacionado con la enfermedad panbronquiolitis difusa [Hijikata et al. 2011]. El gen *HCG22* es no codificante y pertenece a una familia de genes cuya función es codificar proteínas que funcionan como antígenos para los leucocitos en el sistema inmunológico humano. Se ha usado este gen en estudios sobre la enfermedad panbronquiolitis difusa pero no ha sido relacionado con ella, a diferencia de *MUC22* [Hijikata et al. 2011].

El segundo grupo de inversiones no provocan estructuración de la población, pero sí tienen frecuencias diferentes en algunas poblaciones. De la misma forma que en el grupo anterior, no podemos descartar la implicación de la selección natural en su distribución mundial. Forman parte de este grupo las inversiones *HsInv0003, HsInv0004* y *HsInv0041*. Si suponemos que la selección no está implicada, las inversiones se habrían propagado por deriva genética mediante la migración de la población humana, y precisamente en el proceso migratorio fuera de África, se habrían generado diferencias visibles en cuanto a frecuencia en algunas poblaciones. El fenómeno que estaría detrás de estas diferencias es el denominado efecto fundador. Se trata de un proceso estocástico que afecta al establecimiento de una población hija a partir del movimiento migratorio de una parte pequeña de la población madre. Suponiendo que los individuos que forman la nueva población se escogen al azar, sus alelos pueden no representar la frecuencia en la población madre. En la expansión de la población hija, todos los individuos provendrían de los individuos fundadores y por lo tanto, las frecuencias alélicas serían diferentes a las de la población madre.

En el caso de la inversión *HsInv0003*, la frecuencia del alelo invertido en la población Asiática es visiblemente más frecuente, alrededor de un 20%, que en las poblaciones Africana y Europea, al igual que la población Americana que también tiene una frecuencia alrededor de un 10% mayor. Podemos ver la distribución poblacional en la **Figura 4.2**. La distribución se puede explicar por efecto fundador de los individuos que se separaron de la corriente migratoria de África a Eurasia para formar las poblaciones asiáticas, que tendrían en conjunto una frecuencia más alta del alelo invertido, al igual que los individuos que emigraron a América. En la inversión *HsInv0004* ocurre lo mismo. En este caso la población Africana tiene una frecuencia visiblemente menor que el resto de poblaciones, alrededor del 10-15% menor (véase la distribución poblacional en la **Figura 4.3**). Podemos pensar que en este caso la frecuencia del alelo invertido era mayor en el conjunto de los individuos que emigraron de África, de ahí que todas las poblaciones hijas tengan mayor frecuencia de la inversión. En la inversión *HsInv0041* ocurre lo contrario, las poblaciones fuera del continente Africano tienen menor frecuencia del alelo invertido (véase la distribución poblacional en la **Figura 4.4**). Hay que destacar que para estas tres inversiones el alelo derivado es el más frecuente en la población Africana en vez del alelo ancestral. Intuitivamente esto hace que parezca que la selección ha podido actuar en su propagación en algún momento de su historia evolutiva. Las limitaciones de nuestro análisis no nos permiten saber si ha habido participación de la

selección natural o no, pero podemos dar una explicación sin que la selección participe en la distribución de la inversión a través de la deriva genética. Ésta podría haber hecho que la frecuencia del alelo ancestral disminuya en favor del alelo derivado.

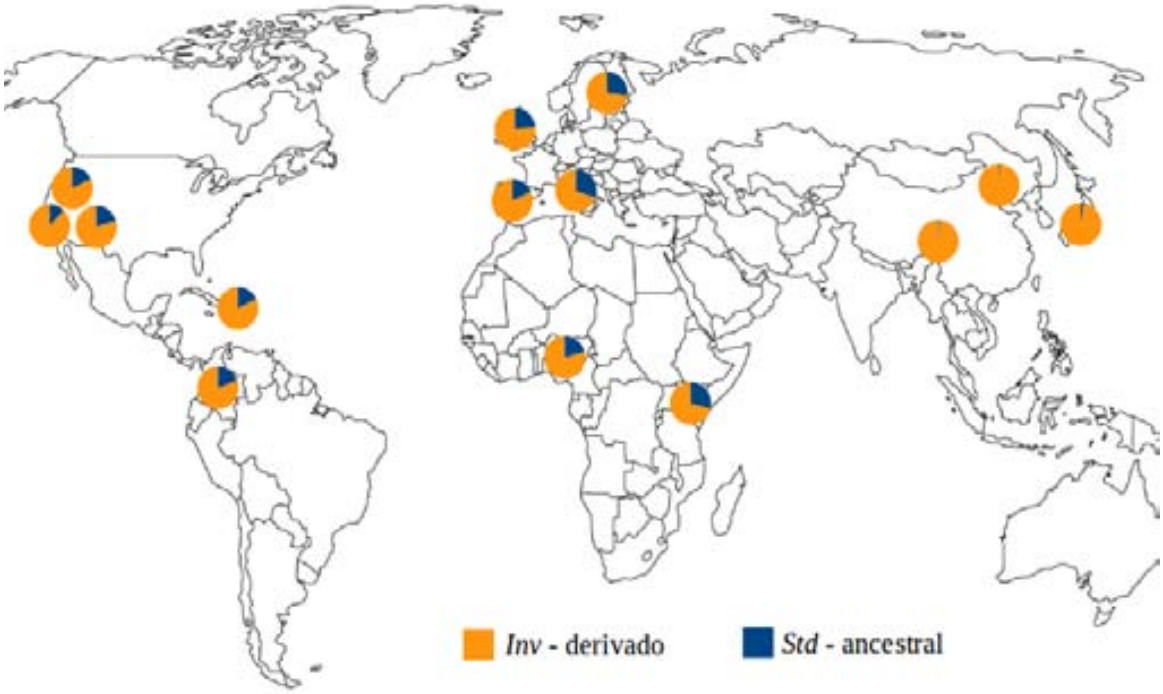

Figura 4.2: Distribución de ambos alelos de la inversión *HsInv0003* en las 14 poblaciones del proyecto de los 1000 Genomas. Se muestra la frecuencia del alelo invertido en color naranja y del alelo estándar en color azul.

En cuanto a la posible implicación en enfermedades y posibles efectos funcionales de este grupo de inversiones, no existen interacciones directas con genes por la posición de sus puntos de rotura ni indirectas por estar localizadas cerca del inicio de transcripción de ningún gen. Esto no quiere decir que no puedan estar afectando a la regulación de los genes más próximos, pero debido a su distancia o posición relativa no destacan como posibles genes afectados.

Finalmente encontramos un grupo de inversiones que tienen valores de *Fst* por encima del umbral de estructura basal de la población humana, y que por lo tanto pueden ser responsables de las diferencias en las frecuencias alélicas entre poblaciones. Se trata de las inversiones *HsInv0006*, *HsInv0052*, *HsInv0059*, *HsInv0063* y *HsInv0068*. Este conjunto de inversiones son candidatas a tener efectos que han estado o están seleccionados y por lo tanto pueden estar implicadas en la adaptación local.

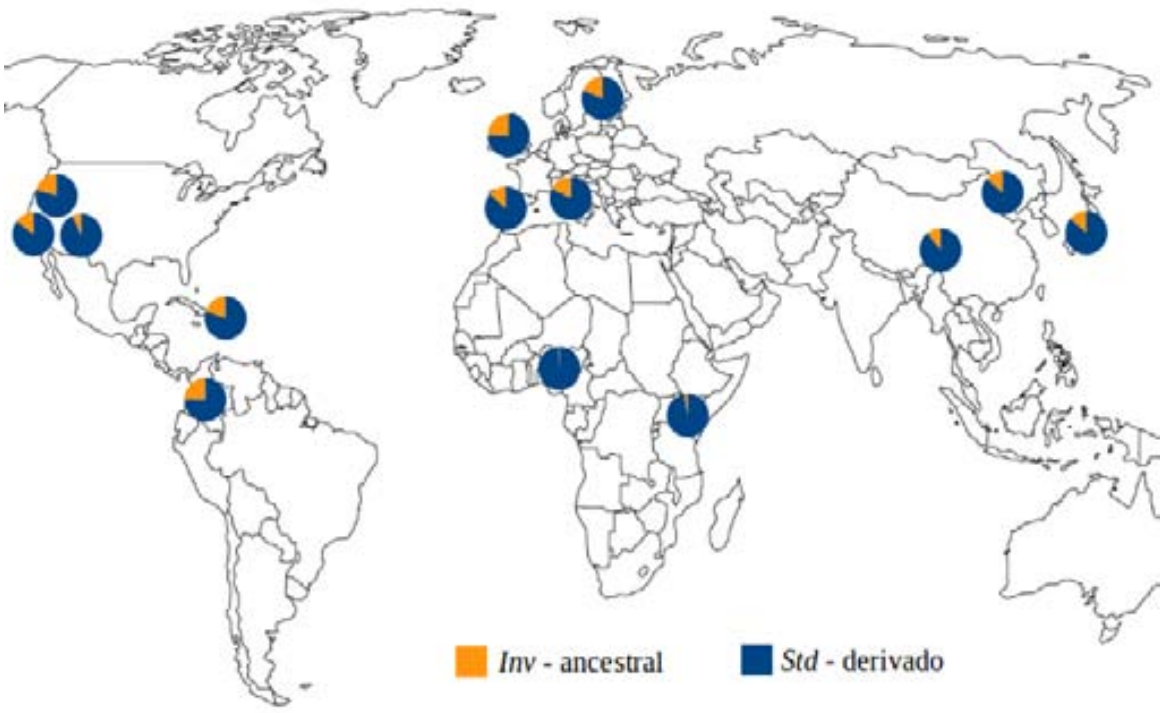

Figura 4.3: Distribución de ambos alelos de la inversión *HsInv0004* en las 14 poblaciones del proyecto de los 1000 Genomas. Se muestra la frecuencia del alelo invertido en color naranja y del alelo estándar en color azul.

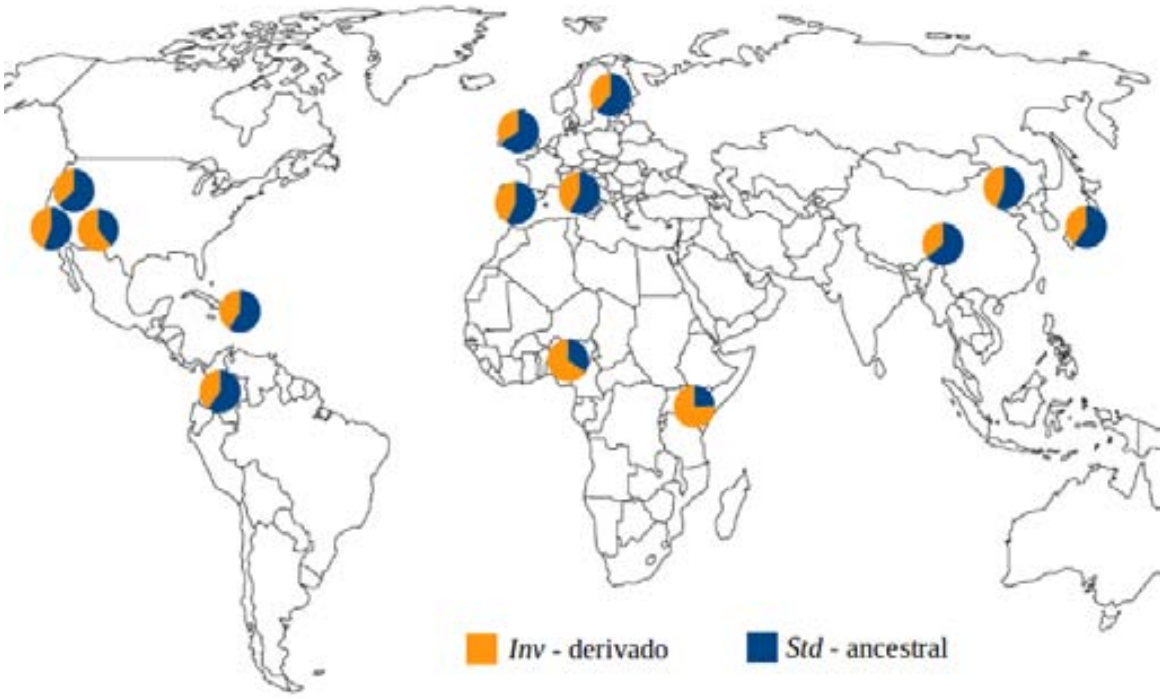

Figura 4.4: Distribución de ambos alelos de la inversión *HsInv0041* en las 14 poblaciones del proyecto de los 1000 Genomas. Se muestra la frecuencia del alelo invertido en color naranja y del alelo estándar en color azul.

La inversión *HsInv0006* tiene un valor de *Fst* de 0.22 para todas las poblaciones y de 0.25 para las poblaciones agrupadas por continentes. Se trata de un valor que supera claramente el umbral de la estructura de población basal. En las diferencias de frecuencia de la inversión entre poblaciones destaca la población Africana, que tiene una frecuencia visiblemente inferior del alelo invertido, a pesar de que se trata del alelo ancestral. La distribución poblacional que se muestra en la **Figura 6** del apartado de resultados se ha podido formar por la acción de distintas fuerzas evolutivas. En este caso no se ha podido comparar su distribución con la obtenida en el estudio de S. Villatoro y colaboradores ya que esta inversión no está incluida. La distribución actual se podría haber generado por deriva genética, en la que podría haber disminuido la frecuencia del alelo invertido en la población africana y luego por efecto fundador establecerse en el resto de poblaciones una frecuencia del alelo invertido mucho mayor. Pero en este caso parece más probable que se haya dado selección además de deriva genética, ya que la distribución actual implicaría dos cambios bruscos de frecuencia por deriva genética, el de la población Africana y el efecto fundador en las poblaciones fuera de África, siendo más improbable la disminución de la frecuencia del alelo invertido ancestral en el continente Africano que su fijación. Además, la frecuencia muy baja del alelo invertido en la población Africana hace más difícil que emigrasen aleatoriamente los individuos portadores de la inversión generando el efecto fundador en las poblaciones hijas. Por lo tanto, es posible que la fuerza evolutiva responsable de la propagación de la inversión sea la selección natural por sí misma o en combinación con la deriva genética. No podemos determinar si realmente se produjo selección natural ni en qué momento, pero las frecuencias alélicas de las diferentes poblaciones y el alelo ancestral sugieren que es posible que la deriva genética actuase en un principio durante la salida de África y que la frecuencia del alelo invertido ancestral inicialmente fuese mayor en África, donde *a posteriori* fue seleccionado negativamente y disminuyó su frecuencia debido a un posible cambio en el ambiente. Otra posibilidad es que la selección mediase la propagación de la inversión desde un principio, fuese seleccionada en el continente Africano y que en su propagación a otros continentes hubiese un cambio en los factores ambientales que provocara selección hacia el otro alelo o que dejara de estar seleccionada y evolucionase de manera neutra, siendo la deriva genética la fuerza evolutiva responsable de su distribución poblacional fuera de África. En cualquier caso, es necesaria la intervención de la selección natural para dar una explicación sencilla a su distribución actual.

En este caso, la localización de la inversión en el primer intrón del gen *DSTYK* puede relacionarse con los efectos posicionales de los puntos de rotura de la inversión, ya que se encuentran muy cerca del punto de inicio de la transcripción y podrían alterar la regulación de la expresión del gen. También podría alterar el *splicing* al estar localizada en un intrón, a través de la alteración de señales de *splicing*. Los efectos de inhibición de la recombinación parecen menos probables por el tamaño de la inversión, que es de 83 pb, un tamaño muy pequeño que hace menos probable que capture combinaciones de alelos coadaptados. El gen *DSTYK* produce una proteína con función quinasa que

modifica otras proteínas mediante la fosforilación, por lo que actúa como un interruptor que puede activar o desactivar otras proteínas. Además se conoce también como *RIP5* y se ha relacionado con la inducción de apoptosis tanto dependiente como independiente de caspasa, por lo que está involucrado en la muerte celular [Zha et al. 2004]. También se ha asociado este gen con malformaciones en el riñón y tracto urinario en humanos, por lo que se relaciona con su desarrollo [Sanna-Cherchi et al. 2013]. Evidentemente son necesarios estudios dirigidos a la detección de la selección natural y estudios de expresión del gen en individuos portadores y no portadores de la inversión para poder sacar conclusiones sobre el posible efecto selectivo de la inversión.

La inversión *HsInv0052* fue validada por Pang y colaboradores en el año 2013 [Pang et al. 2013] junto con la deleción que afecta a la región, pero no se analizó su distribución poblacional. Tiene valores de *Fst* de 0.19 para todas las poblaciones y de 0.22 para las poblaciones agrupadas por continentes, por lo que podemos esperar efectos adaptativos. En este caso las poblaciones Africanas y Asiáticas tienen una frecuencia menor del alelo invertido que las poblaciones Europeas y Americanas. Esto es consistente con el estado ancestral que corresponde al alelo estándar. Esta inversión tampoco se encuentra en el estudio de Villatoro y colaboradores. La distribución poblacional actual de la inversión (véase la **Figura 4.5**) se puede explicar por migración y por tanto por la acción de la deriva genética. En concreto se explicaría por el efecto fundador en las poblaciones Europeas y Americanas, pero no podemos descartar que haya implicación de la selección natural, por lo que ambas fuerzas evolutivas pueden haber tenido un papel importante en la estructura actual de la población. La inversión está localizada en el intrón del gen *BC073807* no codificante de proteína, por lo que es difícil saber si podría estar seleccionada. Su tamaño de 2281 pb hace pensar que es posible que tenga efectos por inhibición de la recombinación, aunque no sea una inversión grande. Lamentablemente, los datos de que disponemos no nos permiten determinar si se dio o está dando una selección a favor o en contra del alelo invertido y en qué poblaciones. Una posibilidad es que la selección natural sea negativa en las poblaciones Africanas y Asiáticas, de manera que se dirigen hacia la fijación del alelo estándar, pero esta hipótesis no es más válida que otras. Cabe destacar que el alelo invertido y el alelo estándar conviven con una deleción mucho más grande que incluye toda la región de la inversión. Mediante la genotipación por *SNPs* marcador se ha visto que esta deleción también presenta estructuración poblacional y los valores de *Fst* son de 0.26 para todas las poblaciones y de 0.30 para las poblaciones agrupadas por continentes, valores más altos que para la inversión. Es comprensible que la deleción de la secuencia pueda tener efectos más negativos y por lo tanto más fuertemente seleccionados que los de una inversión. La deleción tiene un tamaño de 114.2 Kb y no afecta a ningún gen excepto el mismo que la inversión, que tiene un tamaño de unas 507 Kb, y se desconoce si puede tener efectos adaptativos que puedan explicar la propagación de la deleción. A diferencia de la inversión, las frecuencias del alelo delecionado son visiblemente más altas en las poblaciones Africanas y Asiáticas. Podemos ver la distribución poblacional en la **Figura 4.6**. Se trata de un caso

más complejo que el resto y es mucho más difícil tener una idea de qué es lo que ha ocurrido.

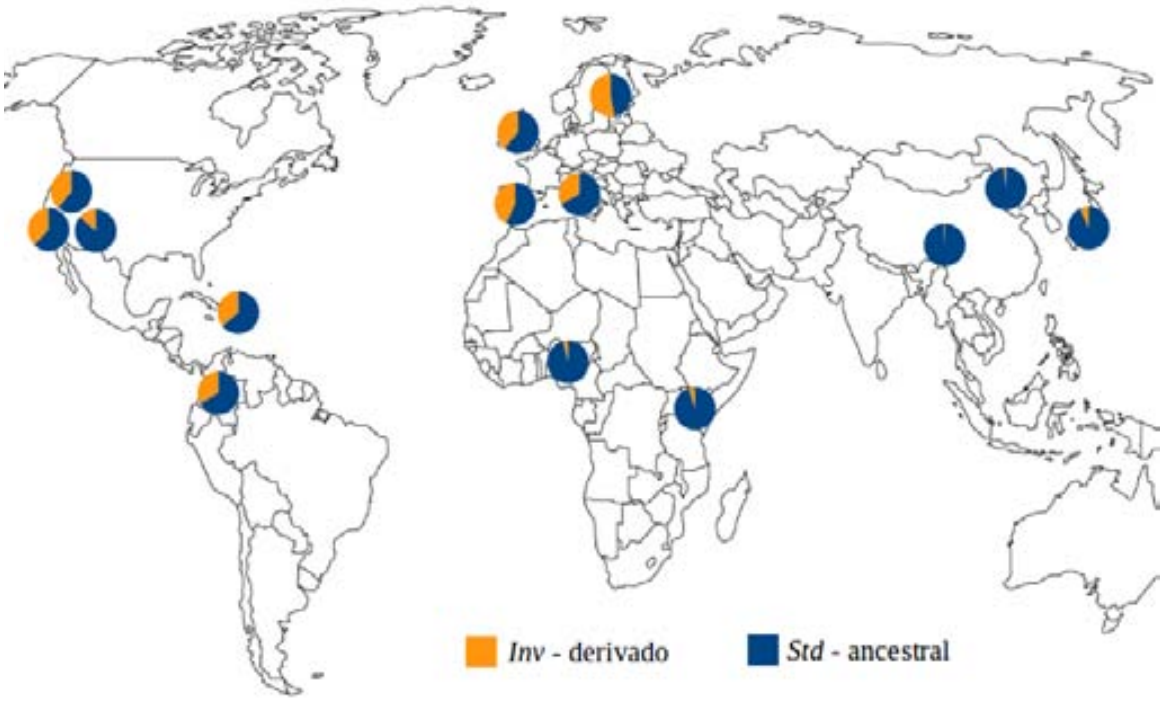

Figura 4.5: Distribución de ambos alelos de la inversión *HsInv0052* en las 14 poblaciones del proyecto de los 1000 Genomas. Se muestra la frecuencia del alelo invertido en color naranja y del alelo estándar en color azul.

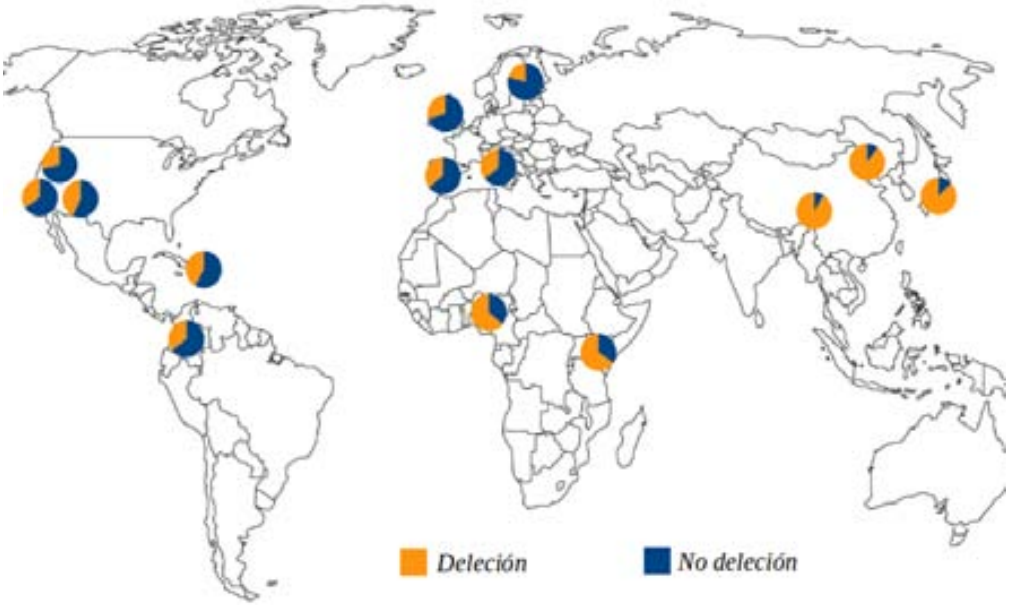

Figura 4.6: Distribución del alelo delecionado y no delecionado en la región de la inversión *HsInv0052* en las 14 poblaciones del proyecto de los 1000 Genomas. Se muestra la frecuencia del alelo delecionado en color naranja y del alelo no delecionado en color azul.

La inversión *HsInv0059* tiene valores de *Fst* de 0.15 y 0.20 para todas las poblaciones según las frecuencias calculadas a partir de genotipación bioinformática y a partir de *SNPs* marcador, respectivamente, y de 0.21 y 0.25 para las poblaciones agrupadas por continentes. En este caso se trata de valores menores que las anteriores inversiones pero superiores a los que indican la estructura basal de la población, por eso podemos esperar efectos adaptativos. El alelo invertido es el más frecuente en todas las poblaciones excepto en la población Asiática (véase la **Figura 4.7**), por lo que en esta población puede haberse dado selección o un efecto fundador. Esta distribución es muy similar a la obtenida por genotipación experimental a gran escala en 7 poblaciones humanas (S. Villatoro y M. Cáceres, resultados no publicados). El estado ancestral es el alelo invertido por lo que la distribución actual se puede explicar por migración fuera de África y la baja frecuencia del alelo invertido en la población Asiática se podría explicar por efecto fundador. No obstante, no podemos descartar un modelo mixto, donde la deriva genética ha actuado en todas las poblaciones y en las poblaciones asiáticas se ha dado además selección en contra del alelo invertido o positiva para el estándar. También es posible que la selección haya actuado sola. De nuevo se trata de una inversión pequeña, de 309 pb, por lo que parece que los efectos seleccionados podrían ser efectos directos de sus puntos de rotura. Al igual que la inversión *HsInv0006*, ésta se encuentra en el primer intrón de un gen, el gen *GABRR1*, y por lo tanto sus puntos de rotura pueden afectar a la regulación del gen por la cercanía al sitio de inicio de transcripción y afectar también a su *splicing*. Este gen codifica para una proteína receptora del neurotransmisor GABA, que tiene funciones inhibidoras en el cerebro de mamíferos. Además ha sido asociado con la susceptibilidad a trastornos mentales como la esquizofrenia, el trastorno obsesivo-compulsivo [Zai et al. 2005] y la epilepsia del lóbulo temporal [Xi et al. 2011]. Al igual que en la inversión *HsInv0006*, son necesarios estudios de expresión y de selección para poder llegar a una conclusión sobre si tiene efectos sobre la eficacia biológica.

La inversión *HsInv0063* ya había sido analizada previamente por Pang y colaboradores en el año 2013. Se trata de una variante estructural compleja que incluye una inversión y una deleción, y los autores del estudio encontraron una asociación entre ambas [Pang et al. 2013]. En nuestro estudio sólo hemos analizado la inversión, ya que la deleción no afecta a la región invertida sino que está localizada en el primer punto de rotura. En el estudio de Pang y colaboradores se calculó una frecuencia del alelo invertido mayor en poblaciones europeas y asiáticas, y un valor de *Fst* de 0.38. Los resultados de nuestro estudio corroboran la menor frecuencia del alelo invertido en las poblaciones Africanas, consistente con el estado ancestral del alelo estándar, pero obtuvimos valores de *Fst* más bajos. Su distribución no muestra diferencias respecto al estudio de Villatoro y colaboradores. Podemos ver la distribución poblacional en la **Figura 4.8**. A partir de las frecuencias obtenidas por genotipación bioinformática obtuvimos valores de *Fst* de 0.11 para todas las poblaciones y 0.15 para las poblaciones agrupadas por continentes. Los valores obtenidos por genotipación a partir de *SNPs* marcador son muy similares, 0.09

para todas las poblaciones y 0.12 para las poblaciones agrupadas por continentes.

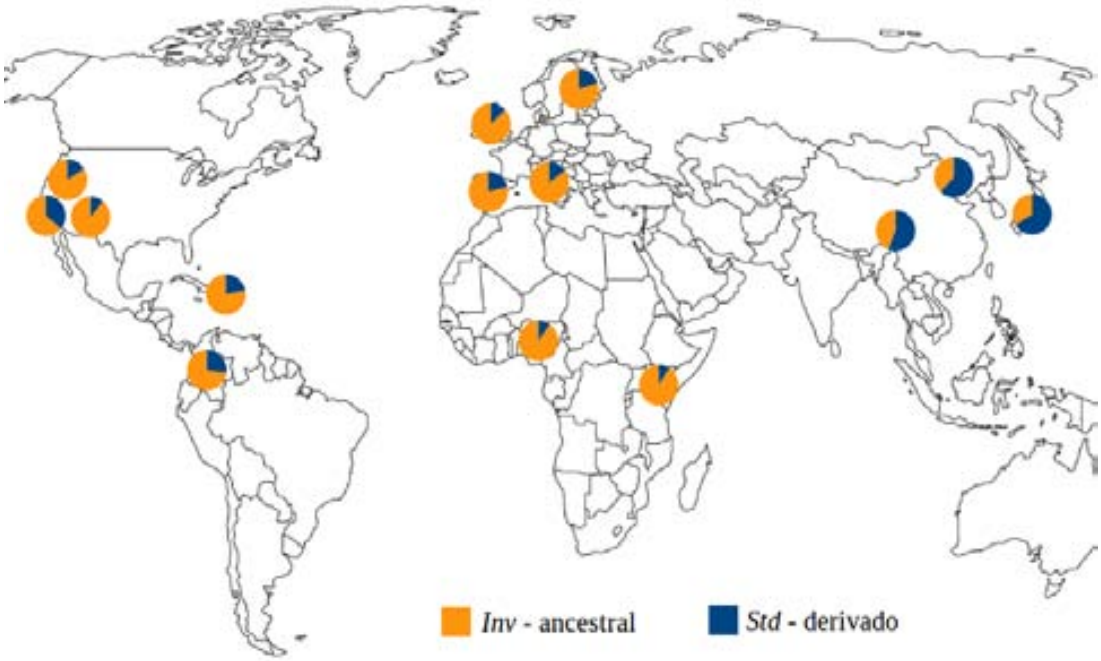

Figura 4.7: Distribución de ambos alelos de la inversión *HsInv0059* en las 14 poblaciones del proyecto de los 1000 Genomas. Se muestra la frecuencia del alelo invertido en color naranja y del alelo estándar en color azul.

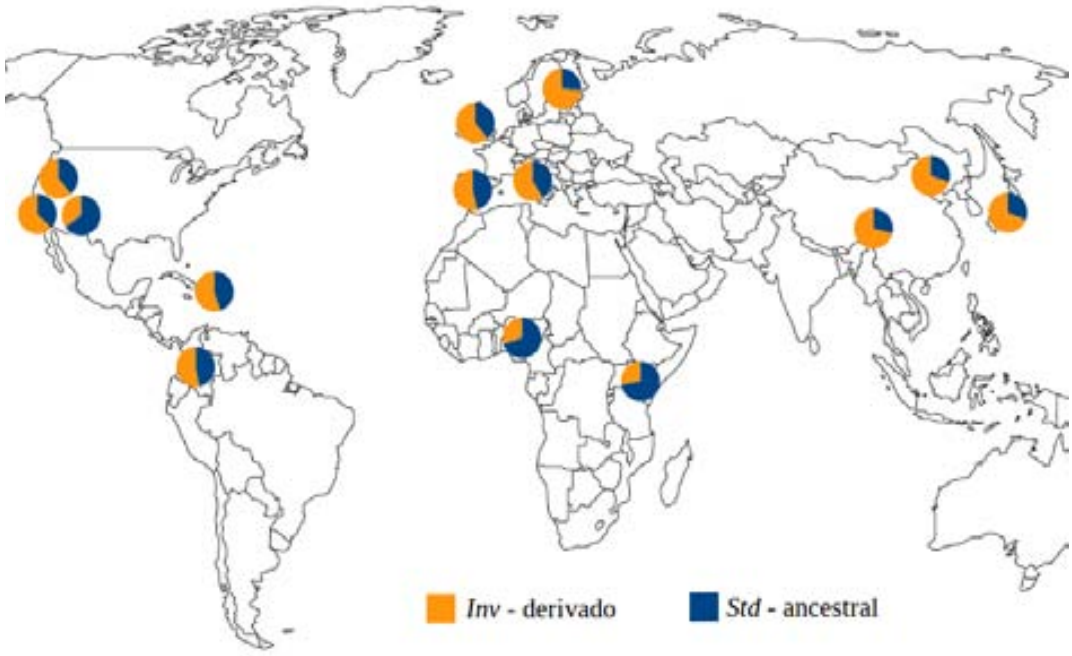

Figura 4.8: Distribución de ambos alelos de la inversión *HsInv0063* en las 14 poblaciones del proyecto de los 1000 Genomas. Se muestra la frecuencia del alelo invertido en color naranja y del alelo estándar en color azul.

En global, superan muy tímidamente el umbral de la variación basal y por lo tanto no esperamos efectos adaptativos evidentes. La distribución actual se explicaría por deriva genética y un efecto fundador en las poblaciones fuera del continente Africano, aunque no podemos descartar la acción de la selección. Este efecto fundador sería menos acusado que en otras inversiones ya que la frecuencia del alelo invertido en las poblaciones Africanas alcanza el 30%, y eso explica los valores bajos de *Fst*. En cuanto a las diferencias con los resultados del estudio de Pang y colaboradores [Pang et al. 2013], las atribuimos a las diferencias de criterio a la hora de seleccionar los *SNPs* marcador. En nuestro estudio se seleccionaron *SNPs* marcador con un valor de r^2 mínimo de 0.99 en todas las poblaciones, y se comprobaron los genotipos de los individuos para el *SNP* marcador con los genotipos de la genotipación bioinformática. De esta manera nos aseguramos de que realmente se trata de un *SNP* que segrega con la inversión, y las frecuencias estimadas a partir de los genotipos para este *SNP* son fiables. En cambio, en el estudio de Pang y colaboradores consideraron *SNPs* marcador aquellos con valores r^2 superiores a 0.8, por lo que no se trata de *SNPs* que segreguen siempre con la inversión, sino que hasta un 20% de los individuos pueden tener genotipos para el *SNP* que no están ligados al alelo invertido, con la consiguiente variación en las frecuencias estimadas. Además las frecuencias poco ajustadas conllevan un cálculo poco fiable del índice *Fst*. Por otra parte, no se especifica el valor de r^2 para cada *SNP* marcador y cada inversión por lo que seguramente han seleccionado el *SNP* más ligado a la inversión y en aquellos casos en que estén fijados o tengan valores de r^2 superiores a 0.95, las frecuencias alélicas serán fiables. En resumen, la inversión no parece tener efectos sobre genes que puedan explicar su distribución actual a partir de la adaptación.

Finalmente, la inversión *HsInv0068* tiene valores de *Fst* de 0.25 para todas las poblaciones y de 0.32 para las poblaciones agrupadas por continentes. A pesar de que se trata de los valores más altos entre las inversiones que superan el umbral de la variación basal entre poblaciones, en este caso no pudimos calcular frecuencias para todas las poblaciones a partir de la genotipación bioinformática debido a la falta de individuos con cobertura de *reads* suficiente. Entre ellas, para una de las poblaciones Europeas, dos poblaciones Africanas y todas las poblaciones Americanas, no se pudieron genotipar al menos 10 individuos, el umbral que usamos para calcular frecuencias fiables. Por lo tanto, no tenemos frecuencias para estas poblaciones y se han eliminado del cálculo de *Fst*. En los resultados del estudio de Villatoro y colaboradores se incluye la frecuencia en la población de Kenya, ausente en nuestro estudio por falta de individuos con suficiente cobertura de *reads*. En el caso de esta inversión, la inclusión de más poblaciones hace que el índice *Fst* que se obtenga sea del 0.06 para todas las poblaciones por lo que no supera el umbral basal (S. Villatoro y M. Cáceres, resultados no publicados). Esto nos indica que el valor del índice *Fst* que hemos obtenido es mayor debido a la inclusión de pocas poblaciones. Por eso no podemos sacar conclusiones sin más datos de frecuencias alélicas. Aun así podemos analizar las que sí tenemos. Podemos ver la distribución poblacional en la **Figura 4.9**.

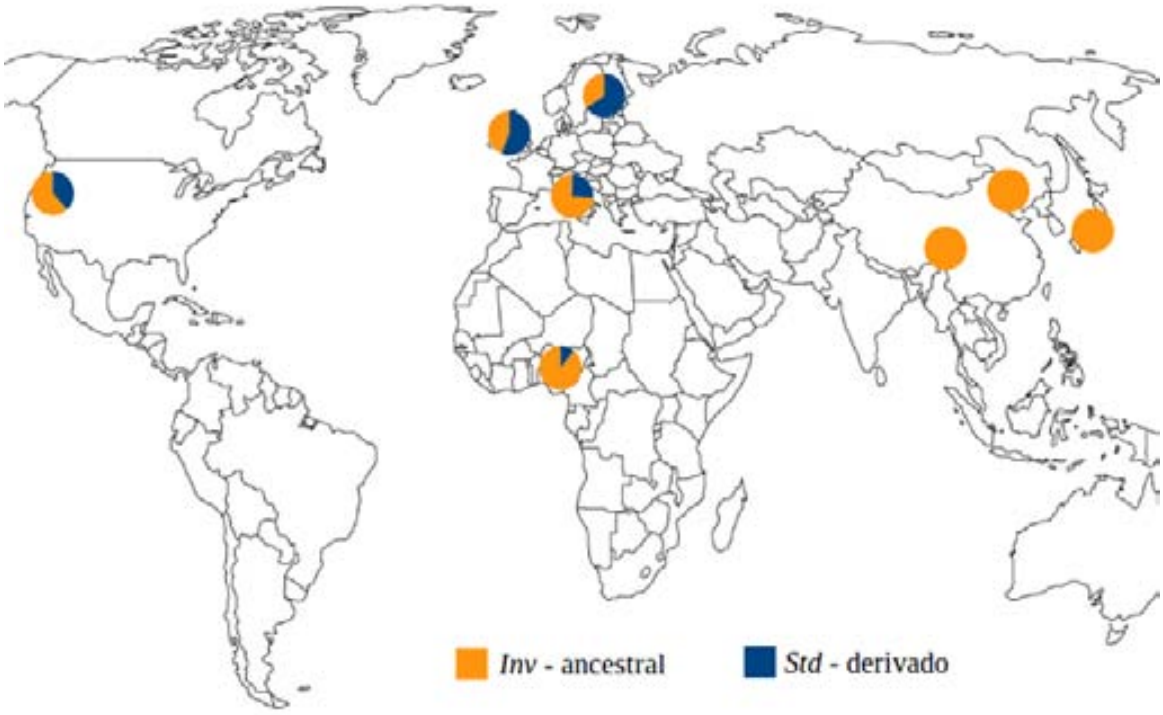

Figura 4.9: Distribución de ambos alelos de la inversión *HsInv0068* en 8 poblaciones del proyecto de los 1000 Genomas. Se muestra la frecuencia del alelo invertido en color naranja y del alelo estándar en color azul.

En la única población Africana para la que tenemos una frecuencia alélica válida, el alelo invertido tiene una frecuencia del 90%, consistente con el estado ancestral del alelo invertido. En Asia la inversión está fijada para todas las poblaciones y en Europa el alelo estándar ha aumentado su frecuencia especialmente en las poblaciones del Norte hasta un 54%, aunque no tenemos datos de todas las poblaciones del Sur. En el continente Americano no hay datos de frecuencia de ninguna población. En conjunto, la distribución actual se puede explicar por deriva genética, pero implica un efecto fundador en las poblaciones Asiáticas diferente del efecto fundador en las poblaciones Europeas, e incluso diferentes efectos fundador entre las poblaciones Europeas del norte y del sur. También podría haber actuado la selección natural en Europa, pero los valores bajos de *Fst* obtenidos para esta inversión en el estudio de Villatoro y colaboradores descartarían esta hipótesis. En este caso es necesaria la genotipación de individuos de todas las poblaciones para poder analizar correctamente su distribución y poder estimar que es lo que pudo ocurrir. En cuanto a los posibles efectos sobre genes serían con mayor probabilidad efectos de los puntos de rotura debido al tamaño pequeño de la inversión, de 249 pb. Aunque no está localizada cerca de ningún gen, se han detectado señales de expresión en la zona. En concreto hay una *EST* o marcador de secuencia expresada, con código *BG613972*, que hace que la inversión pueda está localizada dentro de un transcrito detectado en carcinoma embrionario. Al igual que para el resto de inversiones candidatas a estar seleccionadas, se necesitan estudios dirigidos de detección de la selección natural

y estudios de expresión para poder sacar conclusiones válidas sobre las fuerzas evolutivas que han mediado su expansión. Por eso hay que tomar la discusión actual como un análisis preliminar con el que hemos intentado obtener una primera impresión de lo que ha ocurrido.

Además, existen otras inversiones que no hemos analizado a nivel poblacional que pueden tener efectos funcionales. Se trata de las inversiones *HsInv0030, HsInv0036 y HsInv0069*. La inversión *HsInv0036* está localizada 1712 pb aguas abajo del gen *ANKRD62*, pero a pesar de la cercanía, la posición relativa de ambos hace difícil pensar en un efecto sobre la regulación ya que el inicio de la transcripción queda justo al otro lado del gen. El gen codifica para una proteína que contiene un dominio denominado Anquirina. Las dos inversiones restantes tienen características similares. Sus puntos de rotura están localizados en repeticiones invertidas y los genes a los que afectan están duplicados y forman parte de esas repeticiones invertidas. La diferencia entre sus posibles efectos reside en que la inversión *HsInv0030* intercambia parte de los genes *CTRB2* y *CTRB1* mientras que la inversión *HsInv0069* invierte completamente los genes *FAM225B* y *FAM225A*, de ahí que esperemos mayores efectos funcionales en la inversión *HsInv0030*. La inversión *HsInv0069* al invertir completamente los genes afectados reduce los posibles efectos a un cambio en la regulación de la expresión génica, ya que se invierte la secuencia aguas arriba del gen. Por lo tanto esta inversión es candidata a tener efectos sobre la regulación de ambos genes, que codifican para RNAs largos intergénicos no codificantes de proteínas y cuyas funciones no son conocidas.

Por el contrario la inversión *HsInv0030* intercambia el primer exón de los genes *CTRB2* y *CTRB1* que codifican quimotripsinógenos, proteínas proteasas de serina que son secretadas en el tracto intestinal como precursores inactivos que se activan tras su lisis proteolítica con tripsina. Ambos genes están relacionados con la diabetes [Hart et al. 2013]. La identidad entre ellos es del 97% aunque la identidad de su primer exón que es codificante es del 82%, por lo que la inversión realmente implica cambios de hasta un 18% de la secuencia del primer exón [Pang et al. 2013]. Además se han encontrado 5 transcritos que indican el intercambio del primer exón debido a la inversión [Pang et al. 2013]. Todos estos datos muestran que esta inversión es una seria candidata a tener efectos funcionales. Además se ha detectado una deleción que afecta al exón 5 del gen *CTRB2* que segrega siempre con el alelo invertido [Pang et al. 2013]. En conjunto, la inversión y la deleción modifican la estructura de ambos genes, en especial de *CTRB2* y sus efectos pueden tener carácter adaptativo.

Hemos visto que tanto la inversión *HsInv0069* como la inversión *HsInv0030* tienen sus puntos de rotura localizados en RIs, pero además comparten otra característica, tienen origen recurrente. En primer lugar, esto implica que no se pueden genotipar en individuos de distintas poblaciones por *SNPs* marcador para analizar su distribución y analizar sus implicaciones adaptativas, porque la recurrencia rompe la inhibición de la recombinación

y los alelos de los *SNPs* no segregan siempre con los alelos de la inversión. En el caso de la inversión *HsInv0030*, Pang y colaboradores usaron *SNPs* con valores de r^2 con el alelo invertido iguales o mayores a 0.8, que como hemos visto antes pueden conllevar un error en el cálculo de las frecuencias alélicas. Es un ejemplo de como las inversiones recurrentes pueden llevar a conclusiones erróneas si se genotipan mediante el desequilibrio de ligamiento. Por otro lado, su origen recurrente las hace muy interesantes en cuanto a sus implicaciones adaptativas. En este tipo de inversiones, parece más probable que las consecuencias adaptativas estén mediadas por los efectos posicionales de sus puntos de rotura, que por los efectos de inhibición de la recombinación; ya que estos se romperían en cada evento de recurrencia. Aun así, no podemos descartar que se den efectos adaptativos por mantención de alelos favorables y en cualquier caso estaría relacionado con el tiempo entre eventos de recurrencia. Además estos posibles efectos funcionales no se identificarían en la gran cantidad de estudios de *GWAS* que se han realizado basándose en *SNPs*.

En conjunto, nuestro estudio proporciona un grupo de inversiones candidatas a ser analizadas por sus posibles efectos adaptativos sutiles, que escaparían de una selección negativa drástica, y que pueden estar detrás de la determinación de caracteres favorables en condiciones ambientales determinadas. Por el momento son candidatas a tener efectos adaptativos y para varias de ellas un primer paso es estudiar la expresión de los genes afectados por sus puntos de rotura. Por otro lado, son necesarios estudios dirigidos a la detección de las huellas de la selección natural para poder llegar a conclusiones sobre el papel de esta fuerza evolutiva en la propagación de este conjunto de inversiones.

En ese sentido, el trabajo realizado en esta tesis abre las puertas a la futura determinación del impacto funcional y evolutivo de las inversiones en el genoma humano.

5. CONCLUSIONES

5. CONCLUSIONES

1. El análisis manual y experimental de los genomas de Referencia y de J. Craig Venter mediante el diseño de ensayos de *PCR* y *PCR* inversa ha permitido detectar y descartar como falsos positivos el 65% de las inversiones inicialmente predichas, 29 errores en la comparación genómica y 30 errores de ensamblaje en los genomas comparados; mejorar la definición de los puntos de rotura de 30 inversiones y validar 18 de ellas. Este estudio pone de manifiesto la importancia del análisis específico de estas variantes estructurales y contribuye a la generación de un catálogo de inversiones fiables y no redundantes en el genoma humano.

2. Se ha realizado un gran esfuerzo para mejorar la fiabilidad y calidad del genoma de Referencia resolviendo 25 regiones cuya orientación era errónea y que están relacionadas con la presencia de regiones duplicadas, mediante la determinación de su orientación real a partir de ensayos de *PCR* sobre los *BACs* originales usados en la secuenciación.

3. Aunque el conjunto analizado no representa el total de las inversiones en *HuRef*, la naturaleza no sesgada del método de detección por comparación genómica ha permitido identificar inversiones con diferentes características que forman dos grupos: las inversiones de mayor tamaño tienen sus puntos de rotura localizados en repeticiones invertidas y están formadas por mecanismos homólogos, mientras que las inversiones con puntos de rotura no localizados en repeticiones invertidas tienden a tener un tamaño menor y están formadas por mecanismos no homólogos.

4. Se ha detectado una gran proporción de inversiones que forman parte de variantes complejas compartiendo puntos de rotura no localizados en repeticiones invertidas con inserciones y deleciones, y que en muchos casos incluyen secuencias de micro-homología, por lo que el mecanismo replicativo de *FoSTeS/MMBIR* responsable de su formación tendría un papel muy importante en la determinación del paisaje genómico humano.

5. Mediante el análisis de la variación nudeotídica y haplotípica, se ha demostrado el origen recurrente de 3 inversiones en la población Europea que se caracterizan por tener un número elevado de *SNPs* compartidos entre ordenaciones y tener sus puntos de rotura localizados en repeticiones invertidas que han generado la inversión por el mecanismo homólogo de *NAHR*. Estas características contrastan con las inversiones de origen único con *SNPs* marcador que tienen sus puntos de rotura fuera de repeticiones invertidas y han sido formadas por mecanismos no homólogos.

6. La genotipación experimental de 90 individuos de población Europea para 17 de las inversiones polimórficas ha permitido establecer una frecuencia del alelo diferente al genoma de Referencia de entre el 20% y el 99%, demostrar que todas las inversiones están en equilibrio de Hardy-Weinberg y se trasmiten correctamente en 30 tríos, analizar su origen mediante la asociación con la variación nucleotídica y establecer *SNPs* marcador para su genotipación ($r^2 \geq 0.99$) en 6 inversiones.

7. El genoma humano reutiliza los sitios donde se forman las inversiones cromosómicas mediante la formación de variantes estructurales complejas en que varias variantes comparten los puntos de rotura y mediante la existencia de inversiones de origen recurrente que pueden aparecer y desaparecer en cada evento de recurrencia usando los mismos puntos de rotura.

8. La localización de las 18 inversiones polimórficas analizadas en detalle descarta efectos drásticos sobre genes, ya que no rompen exones ni separan partes de genes. Su localización en regiones intergénicas, intrones o la inclusión de sus puntos de rotura en repeticiones invertidas donde invierten genes duplicados, hace pensar en que sus posibles efectos son sutiles y por lo tanto no están seleccionados fuertemente.

9. Se ha analizado la distribución poblacional de 11 inversiones genotipadas con puntos de rotura simples o *SNP* marcador en 14 poblaciones y se han obtenido valores de diferenciación poblacional muy variables. Se han clasificado las inversiones según las diferencias de frecuencia del alelo invertido entre poblaciones, su estado ancestral, el índice de estructuración de la población *Fst* y sus posibles efectos sobre genes, determinados por la posición de sus puntos de rotura, obteniéndose un grupo de inversiones candidatas a tener efectos adaptativos.

10. Entre las inversiones candidatas a tener efectos adaptativos a través de la alteración de la expresión de genes destaca la inversión *HsInv0006*, de 83 pb, que está localizada en el primer intrón del gen *DSTYK* por lo que puede tener efectos sobre su regulación y sobre el *splicing*. Las frecuencias del alelo diferente al genoma de Referencia muestran diferencias entre poblaciones, y destaca la población Africana, que tiene una frecuencia muy inferior del alelo invertido, a pesar de que se trata del alelo ancestral; lo que sugiere una posible implicación de la selección natural. Además muestran diferenciación en la población con valores de *Fst* de 0.22 para todas las poblaciones y de 0.25 para las poblaciones agrupadas por continentes, que superan claramente el umbral de la estructura de población basal (*Fst* = 0.1).

BIBLIOGRAFÍA

1000 Genomes Project Consortium. (2010). *A map of human genome variation from population-scale sequencing. Nature, 467(7319), 1061-1073.*

1000 Genomes Project Consortium. (2012). *An integrated map of genetic variation from 1,092 human genomes. Nature, 491(7422), 56-65.*

Aguado, C., Gayà-Vidal, M., Villatoro, S., Oliva, M., Izquierdo, D., Giner-Delgado, C., ... & Cáceres, M. (2014). *Validation and Genotyping of Multiple Human Polymorphic Inversions Mediated by Inverted Repeats Reveals a High Degree of Recurrence. PLoS genetics, 10(3), e1004208.*

Ahn, S. M., Kim, T. H., Lee, S., Kim, D., Ghang, H., Kim, D. S., … & Kim, S. J. (2009). *The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome research, 19(9), 1622-1629.*

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). *Genome structural variation discovery and genotyping. Nature Reviews Genetics, 12(5), 363-376.*

Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., ... & Eichler, E. E. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. Nature genetics, 41(10), 1061-1067.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.

Alves, J. M., Chikhi, L., Amorim, A., & Lopes, A. M. (2014). The 8p23 inversion polymorphism determines local recombination heterogeneity across human populations. Genome biology and evolution, 6(4), 921-930.

Alves, J. M., Lopes, A. M., Chikhi, L., & Amorim, A. (2012). On the structural plasticity of the human genome: chromosomal inversions revisited. Current genomics, 13(8), 623.

Anger, G. J., Crocker, S., McKenzie, K., Brown, K. K., Morton, C. C., Harrison, K., & MacKenzie, J. J. (2014). X-Linked Deafness-2 (DFNX2) Phenotype Associated With a Paracentric Inversion Upstream of POU3F4. American journal of audiology, 23(1), 1-6.

Anton, E., Blanco, J., Egozcue, J., & Vidal, F. (2005). Sperm studies in heterozygote inversion carriers: a review. Cytogenetic and genome research, 111(3-4), 297-304.

Anton, E., Vidal, F., Egozcue, J., & Blanco, J. (2006). Genetic reproductive risk in inversion carriers. Fertility and sterility, 85(3), 661-666.

Antonacci, F., Kidd, J. M., Marques-Bonet, T., Ventura, M., Siswara, P., Jiang, Z., & Eichler, E. E. (2009). Characterization of six human disease-associated inversion polymorphisms. Human molecular genetics, 18(14), 2555-2566.

Armengol, G., Knuutila, S., Lozano, J. J., Madrigal, I., & Caballín, M. R. (2010). Identification of human specific gene duplications relative to other primates by array CGH and quantitative PCR. Genomics, 95(4), 203-209.

Armour, J. A., Sismani, C., Patsalis, P. C., & Cross, G. (2000). Measurement of locus copy number by hybridisation with amplifiable probes. Nucleic acids research, 28(2), 605-609.

Ayala, D., Guerrero, R. F., & Kirkpatrick, M. (2013). Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. Evolution, 67(4), 946-958.

Azim, M. K., Yang, C., Yan, Z., Choudhary, M. I., Khan, A., Sun, X., ... & Zhang, Y. (2013). Complete genome sequencing and variant analysis of a Pakistani individual. Journal of human genetics, 58(9), 622-626.

Bagnall, R. D., Waseem, N., Green, P. M., & Giannelli, F. (2002). Recurrent inversion breaking intron 1 of the factor VIII gene is a frequent cause of severe hemophilia A. Blood, 99(1), 168-174.

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., & Eichler, E. E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. Genome research, 11(6), 1005-1017.

Bandelt, H. J., Forster, P., & Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. Molecular biology and evolution, 16(1), 37-48.

Bansal, V., Bashir, A., & Bafna, V. (2007). Evidence for large inversion polymorphisms in the human genome from HapMap data. Genome research, 17(2), 219-230.

Barrett, J. C., Fry, B., Maller, J. D. M. J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics, 21(2), 263-265.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., ... & Anastasi, C. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456(7218), 53-59.

Bondeson, M. L., Dahl, N., Malmgren, H., Kleijer, W. J., Tönnesen, T., Carlberg, B. M., & Pettersson, U. (1995). Inversion of the IDS gene resulting from recombination with IDS-related sequences in a common cause of the Hunter syndrome. Human molecular genetics, 4(4), 615-621.

Bosch, N., Morell, M., Ponsa, I., Mercader, J. M., Armengol, L., & Estivill, X. (2009). Nucleotide, cytogenetic and expression impact of the human chromosome 8p23. 1 inversion polymorphism. PloS one, 4(12), e8269.

Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., ... & Futreal, P. A. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nature genetics, 40(6), 722-729.

CASPERSSON, T., LOMAKKA, G., & ZECH, L. (1971). The 24 fluorescence patterns of the human metaphase chromosomes—distinguishing characters and variability. Hereditas, 67(1), 89-102.

Chaisson, M. J., Raphael, B. J., & Pevzner, P. A. (2006). Microinversions in mammalian evolution. Proceedings of the National Academy of Sciences, 103(52), 19824-19829.

Chen, J. M. *Genomic Rearrangements: Mutational Mechanisms. eLS.*

Chen, J. M., Cooper, D. N., Chuzhanova, N., Férec, C., & Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. Nature Reviews Genetics, 8(10), 762-775.

Cheung, J., Estivill, X., Khaja, R., MacDonald, J. R., Lau, K., Tsui, L. C., & Scherer, S. W. (2003). Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol, 4(4), R25.

Colombo, P. C. (2013). Micro-evolution in grasshoppers mediated by polymorphic Robertsonian translocations. Journal of Insect Science, 13.

Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E., & Pritchard, J. K. (2005). A high-resolution survey of deletion polymorphism in the human genome. Nature genetics, 38(1), 75-81.

Conrad, D. F., Bird, C., Blackburne, B., Lindsay, S., Mamanova, L., Lee, C., ... & Hurles, M. E. (2010). Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nature genetics, 42(5), 385-391.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., ... & Hurles, M. E. (2009). Origins and functional impact of copy number variation in the human genome. Nature, 464(7289), 704-712.

Corbett-Detig, R. B., & Hartl, D. L. (2012). Population genomics of inversion polymorphisms in Drosophila melanogaster. PLoS genetics, 8(12), e1003056.

Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., ... & Dudakia, D. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature, 464(7289), 713-720.

Cáceres, A., Sindi, S. S., Raphael, B. J., Cáceres, M., & González, J. R. (2012). Identification of polymorphic inversions from genotypes. BMC bioinformatics, 13(1), 28.

Cáceres, M., Sullivan, R. T., & Thomas, J. W. (2007). A recurrent inversion on the eutherian X chromosome. Proceedings of the National Academy of Sciences, 104(47), 18571-18576.

Deeb, S. S. (2006). Genetics of variation in human color vision and the retinal cone mosaic. Current opinion in genetics & development, 16(3), 301-307.

del Gaudio, D., Fang, P., Scaglia, F., Ward, P. A., Craigen, W. J., Glaze, D. G., ... & Roa, B. B. (2006). Increased MECP2 gene copy number as the result of genomic duplication in neurodevelopmentally delayed males. Genetics in Medicine, 8(12), 784-792.

Deng, L., Zhang, Y., Kang, J., Liu, T., Zhao, H., Gao, Y., ... & Zeng, C. (2008). An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism. Human mutation, 29(10), 1209-1216.

Dobzhansky, T., & Levene, H. (1948). Genetics of natural populations. XVII. Proof of operation of natural selection in wild populations of Drosophila pseudoobscura. Genetics, 33(6), 537.

Dobzhansky, T., & Sturtevant, A. H. (1938). Inversions in the chromosomes of Drosophila pseudoobscura. Genetics, 23(1), 28.

Dumas, L., Kim, Y. H., Karimpour-Fard, A., Cox, M., Hopkins, J., Pollack, J. R., & Sikela, J. M. (2007). Gene copy number variation spanning 60 million years of human and primate evolution. Genome research, 17(9), 1266-1277.

Dutrillaux, B. (1979). Chromosomal evolution of the great apes and man. Journal of reproduction and fertility. Supplement, 105-111.
Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32(5), 1792-1797.
Eichler, E. E., Nickerson, D. A., Altshuler, D., Bowcock, A. M., Brooks, L. D., Carter, N. P., ... & Waterston, R. H. (2007). Completing the map of human genetic variation. Nature, 447(7141), 161-165.
Entesarian, M., Carlsson, B., Mansouri, M. R., Stattin, E. L., Holmberg, E., Golovleva, I., ... & Dahl, N. (2009). A chromosome 10 variant with a 12 Mb inversion [inv (10)(q11. 22q21. 1)] identical by descent and frequent in the Swedish population. American Journal of Medical Genetics Part A, 149(3), 380-386.
Excoffier, L., Laval, G., & Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary bioinformatics online, 1, 47.
Feuk, L. (2010). Inversion variants in the human genome: role in disease and genome architecture. Genome Med, 2(11).
Feuk, L., Carson, A. R., & Scherer, S. W. (2006) Structural variation in the human genome. Nature Reviews Genetics, 7(2), 85-97.
Feuk, L., MacDonald, J. R., Tang, T., Carson, A. R., Li, M., Rao, G., ... & Scherer, S. W. (2005). Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. PLoS genetics, 1(4), e56.
Flores, M., Morales, L., Gonzaga-Jauregui, C., Domínguez-Vidaña, R., Zepeda, C., Yañez, O., ... & Palacios, R. (2007). Recurrent DNA inversion rearrangements in the human genome. Proceedings of the National Academy of Sciences, 104(15), 6099-6106.
Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., ... & Sikela, J. M. (2004). Lineage-specific gene duplication and loss in human and great ape evolution. PLoS biology, 2(7), e207.
Fujimoto, A., Nakagawa, H., Hosono, N., Nakano, K., Abe, T., Boroevich, K. A., ... & Tsunoda, T. (2010). Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nature genetics, 42(11), 931-936.
Gazave, E., Darré, F., Morcillo-Suarez, C., Petit-Marty, N., Carreño, A., Marigorta, U. M., ... & Navarro, A. (2011). Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. Genome research, 21(10), 1626-1639.
Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., … & Zhang, H. (2003). The international HapMap project. Nature, 426(6968), 789-796.
Giglio, S., Broman, K. W., Matsumoto, N., Calvari, V., Gimelli, G., Neumann, T., ... & Zuffardi, O. (2001). Olfactory receptor–gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. The American Journal of Human Genetics, 68(4), 874-883.
Gilling, M., Dullinger, J. S., Gesk, S., Metzke-Heidemann, S., Siebert, R., Meyer, T., ... & Thomas, N. S. (2006). Breakpoint cloning and haplotype analysis indicate a single origin of the common Inv (10)(p11. 2q21. 2) mutation among northern Europeans. The American Journal of Human Genetics, 78(5), 878-883.

Gimelli, G., Pujana, M. A., Patricelli, M. G., Russo, S., Giardino, D., Larizza, L., ... & Zuffardi, O. (2003). Genomic inversions of human chromosome 15q11–q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. Human molecular genetics, 12(8), 849-858.

Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C., & Gelbart, W. M. (2000). Chromosome mutation II: changes in chromosome number.

Groth, M., Szafranski, K., Taudien, S., Huse, K., Mueller, O., Rosenstiel, P., ... & Platzer, M. (2008). High-resolution mapping of the 8p23. 1 beta-defensin cluster reveals strictly concordant copy number variation of all genes. Human mutation, 29(10), 1247-1254.

Gu, W., Zhang, F., & Lupski, J. R. (2008) Mechanisms for human genomic rearrangements. Pathogenetics, 1(1), 4.

Guerrero, R. F., Rousset, F., & Kirkpatrick, M. (2012). Coalescent patterns for chromosomal inversions in divergent populations. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1587), 430-438.

Gupta, R., Ratan, A., Rajesh, C., Chen, R., Kim, H. L., Burhans, R., ... & Thomas, G. (2012). Sequencing and analysis of a South Asian-Indian personal genome. BMC genomics, 13(1), 440.

Hall, I. M., & Quinlan, A. R. (2012). Detection and interpretation of genomic structural variation in mammals. Genomic Structural Variants, 225-248.

Hall, T. A. (1999, January). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In Nucleic acids symposium series (Vol. 41, pp. 95-98).

Haraksingh, R. R., & Snyder, M. P. (2013). Impacts of variation in the human genome on gene regulation. Journal of molecular biology, 425(21), 3970-3977.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., ... & Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome research, 22(9), 1760-1774.

Hastings, P. J., Ira, G., & Lupski, J. R. (2009). A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS genetics, 5(1), e1000327.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009) Mechanisms of change in gene copy number. Nature Reviews Genetics, 10(8), 551-564.

Heyer, W. D., Ehmsen, K. T., & Liu, J. (2010). Regulation of homologous recombination in eukaryotes. Annual review of genetics, 44, 113-139.

Hijikata, M., Matsushita, I., Tanaka, G., Tsuchiya, T., Ito, H., Tokunaga, K., ... & Keicho, N. (2011). Molecular cloning of two novel mucin-like genes in the disease-susceptibility locus for diffuse panbronchiolitis. Human genetics, 129(2), 117-128.

Hoffmann, A. A., & Rieseberg, L. H. (2008). Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation?. Annual review of ecology, evolution, and systematics, 39, 21.

Huret, J. L., Leonard, C., & Savage, J. R. K. (2000). Chromosomes, chromosome anomalies.

Hurles, M. E., Dermitzakis, E. T., & Tyler-Smith, C. (2008). The functional impact of structural variation in humans. Trends in Genetics, 24(5), 238-245.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., ... & Lee, C. (2004). Detection of large-scale variation in the human genome. Nature genetics, 36(9), 949-951.

International HapMap Consortium. (2005). A haplotype map of the human genome. Nature, 437(7063), 1299-1320.

Istrail, S., Sutton, G. G., Florea, L., Halpern, A. L., Mobarry, C. M., Lippert, R., ... & Venter, J. C. (2004). Whole-genome shotgun assembly and comparison of human genome assemblies. Proceedings of the National Academy of Sciences of the United States of America, 101(7), 1916-1921.

Johansson, I., Lundqvist, E., Bertilsson, L., Dahl, M. L., Sjöqvist, F., & Ingelman-Sundberg, M. (1993). Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. Proceedings of the National Academy of Sciences, 90(24), 11825-11829.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., ... & Broad Institute Genome Sequencing Platform & Whole Genome Assembly Team. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. Nature, 484(7392), 55-61.

Kehrer-Sawatzki, H., & Cooper, D. N. (2007). Structural divergence between the human and chimpanzee genomes. Human genetics, 120(6), 759-778.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. Genome research, 12(6), 996-1006.

Khaja, R., Zhang, J., MacDonald, J. R., He, Y., Joseph-George, A. M., Wei, J., ... & Feuk, L. (2006). Genome assembly comparison identifies structural variants in the human genome. Nature genetics, 38(12), 1413-1418.

Khorana, H. G., Büuchi, H., Ghosh, H., Gupta, N., Jacob, T. M., Kössel, H., ... & Wells, R. D. (1966, January). Polynucleotide synthesis and the genetic code. In Cold Spring Harbor symposia on quantitative biology (Vol. 31, pp. 39-49). Cold Spring Harbor Laboratory Press.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., ... & Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature, 453(7191), 56-64.

Kidd, J. M., Graves, T., Newman, T. L., Fulton, R., Hayden, H. S., Malig, M., ... & Eichler, E. E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell, 143(5), 837-847.

Kidd, J. M., Newman, T. L., Tuzun, E., Kaul, R., & Eichler, E. E. (2007). Population stratification of a common APOBEC gene deletion polymorphism. PLoS genetics, 3(4), e63.

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. PLoS biology, 8(9), e1000501.

Koolen, D. A., Vissers, L. E., Pfundt, R., de Leeuw, N., Knight, S. J., Regan, R., ... & de Vries, B. B. (2006). A new chromosome 17q21. 31 microdeletion syndrome associated with a common inversion polymorphism. Nature genetics, 38(9), 999-1001.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., ... & Snyder, M. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science, 318(5849), 420-426.

Krimbas, C. B., & Powell, J. R. (1992). Drosophila inversion polymorphism.

Lakich, D., Kazazian, H. H., Antonarakis, S. E., & Gitschier, J. (1993). Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. Nature genetics, 5(3), 236-241.

Lam, H. Y., Mu, X. J., Stütz, A. M., Tanzer, A., Cayting, P. D., Snyder, M., ... & Gerstein, M. B. (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nature biotechnology, 28(1), 47-55.

Langer-Safer, P. R., Levine, M., & Ward, D. C. (1982). Immunological method for mapping genes on Drosophila polytene chromosomes. Proceedings of the National Academy of Sciences, 79(14), 4381-4385.

Lee, J. A., Carvalho, C., & Lupski, J. R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. cell, 131(7), 1235-1247.

Lejeune, J. T. R. G. M., Turpin, R., & Gautier, M. (1959). Le mongolisme, premier exemple d'aberration autosomique humaine. Ann Genet, 1(4), 1-49.

Lejeune, J., Dutrillaux, B., Rethoré, M. O., & Prieur, M. (1973). Comparaison de la structure fine des chromatides d'Homo sapiens et de Pan troglodytes. Chromosoma, 43(4), 423-444.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., ... & Venter, J. C. (2007). The diploid genome sequence of an individual human. PLoS biology, 5(10), e254.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., ... & Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome research, 20(2), 265-272.

Li, Y., Zheng, H., Luo, R., Wu, H., Zhu, H., Li, R., ... & Wang, J. (2011). Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. Nature biotechnology, 29(8), 723-730.

Lilleoja, R., Sarapik, A., Reimann, E., Reemann, P., Jaakma, Ü., Vasar, E., & Kõks, S. (2012). Sequencing and annotated analysis of an Estonian human genome. Gene, 493(1), 69-76.

Liu, P., Erez, A., Nagamani, S. C. S., Dhar, S. U., Kołodziejska, K. E., Dharmadhikari, A. V., ... & Bi, W. (2011). Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell, 146(6), 889-903.

Lledó, J. I. L., & Cáceres, M. (2013). On the power and the systematic biases of the detection of chromosomal inversions by paired-end genome sequencing. PloS one, 8(4), e61292.

Lowry, D. B., & Willis, J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. PLoS Biology, 8(9), e1000500.

Lucas-Lledó, J. I., Vicente-Salvador, D., Aguado, C., & Cáceres, M. (2014). Population genetic analysis of bi-allelic structural variants from low-coverage sequence data with an expectation-maximization algorithm. BMC bioinformatics, 15(1), 163.

MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L., & Scherer, S. W. (2014). The database of genomic variants: a curated collection of structural variation in the human genome. Nucleic acids research, 42(D1), D986-D992.

Marques-Bonet, T., Kidd, J. M., Ventura, M., Graves, T. A., Cheng, Z., Hillier, L. W., ... & Eichler, E. E. (2009). A burst of segmental duplications in the genome of the African great ape ancestor. Nature, 457(7231), 877-881.

Mars, W. M., Patmasiriwat, P., Maity, T., Huff, V., Weil, M. M., & Saunders, G. F. (1995). Inheritance of unequal numbers of the genes encoding the human neutrophil defensins HP-1 and HP-3. Journal of Biological Chemistry, 270(51), 30371-30376.

Martínez-Fundichely, A., Casillas, S., Egea, R., Ràmia, M., Barbadilla, A., Pantano, L., ... & Cáceres, M. (2013). InvFEST, a database integrating information of polymorphic inversions in the human genome. Nucleic acids research, gkt1122.

Mathews, J., Duncavage, E. J., & Pfeifer, J. D. (2013). Characterization of translocations in mesenchymal hamartoma and undifferentiated embryonal sarcoma of the liver. Experimental and molecular pathology, 95(3), 319-324.

McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., ... & International HapMap Consortium. (2005). Common deletion polymorphisms in the human genome. Nature genetics, 38(1), 86-92.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., ... & Blanchard, A. P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome research, 19(9), 1527-1541.

Medvedev, P., Stanciu, M., & Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. Nature methods, 6, S13-S20.

Merla, G., Brunetti-Pierri, N., Micale, L., & Fusco, C. (2010). Copy number variants at Williams–Beuren syndrome 7q11. 23 region. Human genetics, 128(1), 3-26.

Miller, R. J., & Reis, D. J. (1982). The origin of man: a chromosomal pictorial legacy. Science, 215, 1526.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., & Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. Genome research, 16(9), 1182-1190.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., … & Wang, J. (2011). Mapping copy number variation by population-scale genome sequencing. Nature, 470(7332), 59-65.

Morel, F., Laudier, B., Guerif, F., Couet, M. L., Royere, D., Roux, C., ... & Douet-Guilbert, N. (2007). Meiotic segregation analysis in spermatozoa of pericentric inversion carriers using fluorescence in-situ hybridization. Human Reproduction, 22(1), 136-141.

Mullis, K. F. F. S. S. S. R. H. G., Faloona, F. A., Scharf, S. J., Saiki, R. K., Horn, G. T., & Erlich, H. (1992). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. BIOTECHNOLOGY SERIES, 17-17.

M't Hart, L., Fritsche, A., Nijpels, G., van Leeuwen, N., Donnelly, L. A., Dekker, J. M., ... & Diamant, M. (2013). The CTRB1/2 locus affects diabetes susceptibility and treatment via the incretin pathway. Diabetes, DB_130227.

Navarro, A., & Barton, N. H. (2003). Chromosomal speciation and molecular divergence--accelerated evolution in rearranged chromosomes. Science, 300(5617), 321-324.

Newman, T. L., Tuzun, E., Morrison, V. A., Hayden, K. E., Ventura, M., McGrath, S. D., ... & Eichler, E. E. (2005) A genome-wide survey of structural variation between human and chimpanzee. Genome research, 15(10), 1344-1356.

Nguyen, D. Q., Webber, C. P., Hehir-Kwa, J., Pfundt, R., Veltman, J., & Ponting, C. P. (2008). Reduced purifying selection prevails over positive selection in human copy number variant evolution. Genome research, gr-077289.

Ochman, H., Gerber, A. S., & Hartl, D. L. (1988). Genetic applications of an inverse polymerase chain reaction. Genetics, 120(3), 621-623.

Onishi-Seebacher, M., & Korbel, J. O. (2011). Challenges in studying genomic structural variant formation mechanisms: The short-read dilemma and beyond. Bioessays, 33(11), 840-850.

Osborne, L. R., Li, M., Pober, B., Chitayat, D., Bodurtha, J., Mandel, A., ... & Scherer, S. W. (2001). A 1.5 million–base pair inversion polymorphism in families with Williams-Beuren syndrome. Nature genetics, 29(3), 321-325.

Pang, A. W. C., Migita, O., MacDonald, J. R., Feuk, L., & Scherer, S. W. (2013). Mechanisms of formation of structural variation in a fully sequenced human genome. Human mutation, 34(2), 345-354.

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., ... & Scherer, S. W. (2010). Research Towards a comprehensive structural variation map of an individual human genome.

Pearson, H. (2006). Genetics: what is a gene?. Nature, 441(7092), 398-401.

Pegueroles, C., Ordonez, V., Mestres, F., & Pascual, M. (2010). Recombination and selection in the maintenance of the adaptive value of inversions. Journal of evolutionary biology, 23(12), 2709-2717.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., ... & Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. Nature genetics, 39(10), 1256-1260.

Perry, G. H., Tchinda, J., McGrath, S. D., Zhang, J., Picker, S. R., Cáceres, A. M., ... & Lee, C. (2006). Hotspots for copy number variation in chimpanzees and humans. Proceedings of the National Academy of Sciences, 103(21), 8006-8011.

Perry, G. H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A. S., ... & Redon, R. (2008). Copy number variation and evolution in humans and chimpanzees. Genome research, 18(11), 1698-1710.

Quinlan, A. R., & Hall, I. M. (2012). Characterizing complex structural variation in germline and somatic genomes. Trends in Genetics, 28(1), 43-53.

Ray, F. A., Zimmerman, E., Robinson, B., Cornforth, M. N., Bedford, J. S., Goodwin, E. H., & Bailey, S. M. (2013). Directional genomic hybridization for chromosomal inversion discovery and detection. Chromosome Research, 21(2), 165-174.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … & Hurles, M. E. (2006). Global variation in copy number in the human genome. nature, 444(7118), 444-454.

Rodríguez, L., Bhatt, S. S., García-Castro, M., Plasencia, A., Fernández-Toral, J., Abarca, E., ... & Liehr, T. (2014). A unique case of a discontinuous duplication 3q26. 1-3q28 resulting from a segregation error of a maternal complex chromosomal rearrangement involving an insertion and an inversion. Gene, 535(2), 165-169.

Salm, M. P., Horswell, S. D., Hutchison, C. E., Speedy, H. E., Yang, X., Liang, L., ... & Shoulders, C. C. (2012). The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome research, 22(6), 1144-1153.

Sanna-Cherchi, S., Sampogna, R. V., Papeta, N., Burgess, K. E., Nees, S. N., Perry, B. J., ... & Gharavi, A. G. (2013). Mutations in DSTYK and dominant urinary tract malformations. New England Journal of Medicine, 369(7), 621-629.

Santoro, M., Melillo, R. M., & Fusco, A. (2006). RET/PTC activation in papillary thyroid carcinoma: European Journal of Endocrinology Prize Lecture. European Journal of Endocrinology, 155(5), 645-653.

Scambler, P. J., Kelly, D., Lindsay, E., Williamson, R., Goldberg, R., Shprintzen, R., ... & Burn, J. (1992). Velo-cardio-facial syndrome associated with chromosome 22 deletions encompassing the DiGeorge locus. The Lancet, 339(8802), 1138-1139.

Schouten, J. P., McElgunn, C. J., Waaijer, R., Zwijnenburg, D., Diepvens, F., & Pals, G. (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic acids research, 30(12), e57-e57.

Schwartz, D. C., & Cantor, C. R. (1984). Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell, 37(1), 67-75.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., ... & Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. Science, 305(5683), 525-528.

Sellner, L. N., & Taylor, G. R. (2004). MLPA and MAPH: new techniques for detection of gene deletions. Human mutation, 23(5), 413-419.

Sgardioli, I. C., Simioni, M., Viguetti-Campos, N. L., Prota, J. R., & Gil-da-Silva-Lopes, V. L. (2013). A new case of partial 14q31. 3-qter trisomy due to maternal pericentric inversion. Gene, 523(2), 192-194.

Sharakhova, M. V., ANTONIO-NKONDJIO, C., Xia, A., Ndo, C., AWONO-AMBENE, P., Simard, F., & Sharakhov, I. V. (2013). Polymorphic chromosomal inversions in Anopheles moucheti, a major malaria vector in Central Africa. Medical and veterinary entomology.

Sharp, A. J., Cheng, Z., & Eichler, E. E. (2006). Structural variation of the human genome. Annu. Rev. Genomics Hum. Genet., 7, 407-442.

Sharp, A. J., Hansen, S., Selzer, R. R., Cheng, Z., Regan, R., Hurst, J. A., ... & Eichler, E. E. (2006). Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nature genetics, 38(9), 1038-1042.

She, X., Jiang, Z., Clark, R. A., Liu, G., Cheng, Z., Tuzun, E., ... & Eichler, E. E. (2004). Shotgun sequence assembly and recent segmental duplications within the human genome. Nature, 431(7011), 927-930.

Shen, H., Li, J., Zhang, J., Xu, C., Jiang, Y., Wu, Z., ... & Deng, H. W. (2013). Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. PloS one, 8(4), e59494.

Sindi, S. S., & Raphael, B. J. (2010). Identification and frequency estimation of inversion polymorphisms from haplotype data. Journal of computational biology, 17(3), 517-531.

Small, K., Iber, J., & Warren, S. T. (1997). Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. Nature genetics, 16(1), 96-99.

Soda, M., Choi, Y. L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., ... & Mano, H. (2007). Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. Nature, 448(7153), 561-566.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. Journal of molecular biology, 98(3), 503-517.

Spielmann, M., & Mundlos, S. (2013). Structural variations, the regulatory landscape of the genome and their alteration in human disease. Bioessays, 35(6), 533-543.

Spitz, F., Herkenne, C., Morris, M. A., & Duboule, D. (2005). Inversion-induced disruption of the Hoxd cluster leads to the partition of regulatory landscapes. Nature genetics, 37(8), 889-893.

Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. TRENDS in Genetics, 18(2), 74-82.

Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. Annual review of medicine, 61, 437-455.

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., ... & Stefansson, K. (2005). A common inversion under selection in Europeans. Nature genetics, 37(2), 129-137.

Stephens, M., Smith, N. J., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. The American Journal of Human Genetics, 68(4), 978-989.

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., ... & Campbell, P. J. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell, 144(1), 27-40.

Stevison, L. S., Hoehn, K. B., & Noor, M. A. (2011). Effects of inversions on within-and between-species recombination and divergence. Genome biology and evolution, 3, 830-841.

Sturtevant, A. H. (1921). A case of rearrangement of genes in Drosophila. Proceedings of the National Academy of Sciences of the United States of America, 7(8), 235.

Tan, J. C., Tan, A., Checkley, L., Honsa, C. M., & Ferdig, M. T. (2010). Variable numbers of tandem repeats in Plasmodium falciparum genes. Journal of molecular evolution, 71(4), 268-278.

Teague, B., Waterman, M. S., Goldstein, S., Potamousis, K., Zhou, S., Reslewic, S., ... & Schwartz, D. C. (2010). High-resolution human genome structure by single-molecule analysis. Proceedings of the National Academy of Sciences, 107(24), 10848-10853.

Thabet, M. M., Huizinga, T. W. J., Marques, R. B., Stoeken-Rijsbergen, G., Bakker, A. M., Kurreeman, F. A., ... & Van Der Helm-Van Mil, A. H. M. (2009). Contribution of Fcγ receptor IIIA gene 158V/F polymorphism and copy number variation to the risk of ACPA-positive rheumatoid arthritis. Annals of the rheumatic diseases, 68(11), 1775-1780.

Thomas, N. S., Bryant, V., Maloney, V., Cockwell, A. E., & Jacobs, P. A. (2008). Investigation of the origins of human autosomal inversions. Human genetics, 123(6), 607-616.

Tishkoff, S. A., & Verrelli, B. C. (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. Annual review of genomics and human genetics, 4(1), 293-340.

Turner, D. J., Shendure, J., Porreca, G., Church, G., Green, P., Tyler-Smith, C., & Hurles, M. E. (2006). Assaying chromosomal inversions by single-molecule haplotyping. Nature methods, 3(6), 439-445.

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., ... & Eichler, E. E. (2005). Fine-scale structural variation of the human genome. Nature genetics, 37(7), 727-732.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., & Leunissen, J. A. (2007). Primer3Plus, an enhanced web interface to Primer3. Nucleic acids research, 35(suppl 2), W71-W74.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., ... & Ye, J. (2008). The diploid genome sequence of an Asian individual. Nature, 456(7218), 60-65.

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids. Nature, 171(4356), 737-738.

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nature Reviews Genetics, 14(2), 125-138.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., ... & Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. nature, 452(7189), 872-876.

Wilson, G. M., Flibotte, S., Missirlis, P. I., Marra, M. A., Jones, S., Thornton, K., ... & Holt, R. A. (2006). Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla. Genome research, 16(2), 173-181.

Wu, Y. L., Savelli, S. L., Yang, Y., Zhou, B., Rovin, B. H., Birmingham, D. J., ... & Yu, C. Y. (2007). Sensitive and specific real-time polymerase chain reaction assays to accurately determine copy number variations (CNVs) of human complement C4A, C4B, C4-long, C4-short, and RCCX modules: elucidation of C4 CNVs in 50 consanguineous subjects with defined HLA genotypes. The Journal of Immunology, 179(5), 3012-3025.

Xi, B., Chen, J., Yang, L., Wang, W., Fu, M., & Wang, C. (2011). GABBR1 gene polymorphism (G1465A) is associated with temporal lobe epilepsy. Epilepsy research, 96(1), 58-63.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., & Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC bioinformatics, 13(1), 134.

Yon, J., & Fried, M. (1989). Precise gene fusion by PCR. Nucleic Acids Research, 17(12), 4895.

Yunis, J. J., & Dunham, K. (1980). The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee. Science, 208(4448), 1145-1148.

Zai, G., Arnold, P., Burroughs, E., Barr, C. L., Richter, M. A., & Kennedy, J. L. (2005). Evidence for the gamma-amino-butyric acid type B receptor 1 (GABBR1) gene as a susceptibility factor in obsessive-compulsive disorder. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 134(1), 25-29.

Zai, G., King, N., Wong, G. W., Barr, C. L., & Kennedy, J. L. (2005). Possible association between the gamma-aminobutyric acid type B receptor 1 (GABBR1) gene and schizophrenia. European neuropsychopharmacology, 15(3), 347-352.

Zha, J., Zhou, Q., Xu, L. G., Chen, D., Li, L., Zhai, Z., & Shu, H. B. (2004). RIP5 is a RIP-homologous inducer of cell death. Biochemical and biophysical research communications, 319(2), 298-303.

Zody, M. C., Jiang, Z., Fung, H. C., Antonacci, F., Hillier, L. W., Cardone, M. F., ... & Eichler, E. E. (2008). Evolutionary toggling of the MAPT 17q21. 31 inversion region. Nature genetics, 40(9), 1076-1083.

AGRADECIMIENTOS

Me gustaría darles las gracias a muchas personas que me han brindado su ayuda en distintos aspectos relacionados con esta tesis doctoral.

En primer lugar a mi director de tesis, el doctor Mario Cáceres, por darme la oportunidad de realizar la tesis doctoral en su laboratorio, por su buena labor en la dirección, por mostrarme por dónde continuar siempre que hubo confusión o no supe que decisión tomar, por su cercanía y por su entrega y dedicación a la ciencia que son para mi un ejemplo.

En segundo lugar, quiero acordarme de mi estancia en la plataforma Bioinformàtica de la UAB, dirigida por el doctor Antonio Barbadilla. Él me dio la oportunidad de aprender bajo la dirección de la doctora Sónia Casillas. También me habló de Mario y gracias a ambos me pude poner en contacto con él.

Durante mis 4 años en el laboratorio de Mario he contado con la ayuda y dirección específica en distintos apartados de la tesis de muchos de los doctores del grupo. Quiero agradecer a la doctora Marta Puig su ayuda en cuanto a las técnicas de laboratorio, su análisis crítico de los resultados que me ha permitido aprender de los errores y su consejo general. A la doctora Cristina Aguado por su ayuda en el diseño y aplicación del protocolo de *PCR* inversa, al doctor Ignasi Lucas por su ayuda y dedicación en la modificación del software *BreakSeq* que se usó para la genotipación bioinformática de las inversiones, además de su gran aportación estadística en forma del software *svgem*, que nos permitió mejorar la fiabilidad de la genotipación; a la doctora Magda Gayà por su ayuda tanto en el análisis del desequilibrio de ligamiento entre los alelos de *SNPs* e inversiones como en el análisis de los haplotipos. Quiero destacar que no sólo estoy agradecido por su ayuda e implicación sino también por el trato que he recibido por parte de todos ellos y de todos los integrantes del grupo, incluidas las personas que ya no se encuentran en él.

Esta experiencia ha sido muy fructífera y quiero agradecer también el apoyo y la comprensión de quien realiza un camino paralelo, a mis compañeros de laboratorio y a mis excompañeros en la plataforma bioinformática y compañeros en el IBB, Maite Barrón y Miquel Ramia. Finalmente, quiero agradecer su infinita ayuda en temas de administración de servidores y su consejo en temas informáticos a Òscar Conchillo.

Evidentemente la ciencia no se queda encerrada en el laboratorio, sino que forma parte de nuestras vidas, por eso quiero agradecer a mi pareja, Gemma Gou, su apoyo incondicional, su paciencia y comprensión en los momentos en que la tesis ha estado por encima de todo y la energía que me ha transmitido de manera constante que me ha ayudado mucho en los momentos en que las cosas no van como uno espera. También

quiero agradecer a mi familia que hayan creído en mí y que me hayan apoyado en mis decisiones, en especial a mis padres que han hecho muchas veces un esfuerzo extra para que yo pudiera seguir viviendo fuera de casa. Finalmente quiero agradecer a un buen amigo, Alex Chamorro, su visión de la ciencia y de la vida, que distinta a la mía, me ha hecho reflexionar muchas veces y que sin duda ha contribuido de una manera u otra al crecimiento personal que he adquirido con este proyecto.