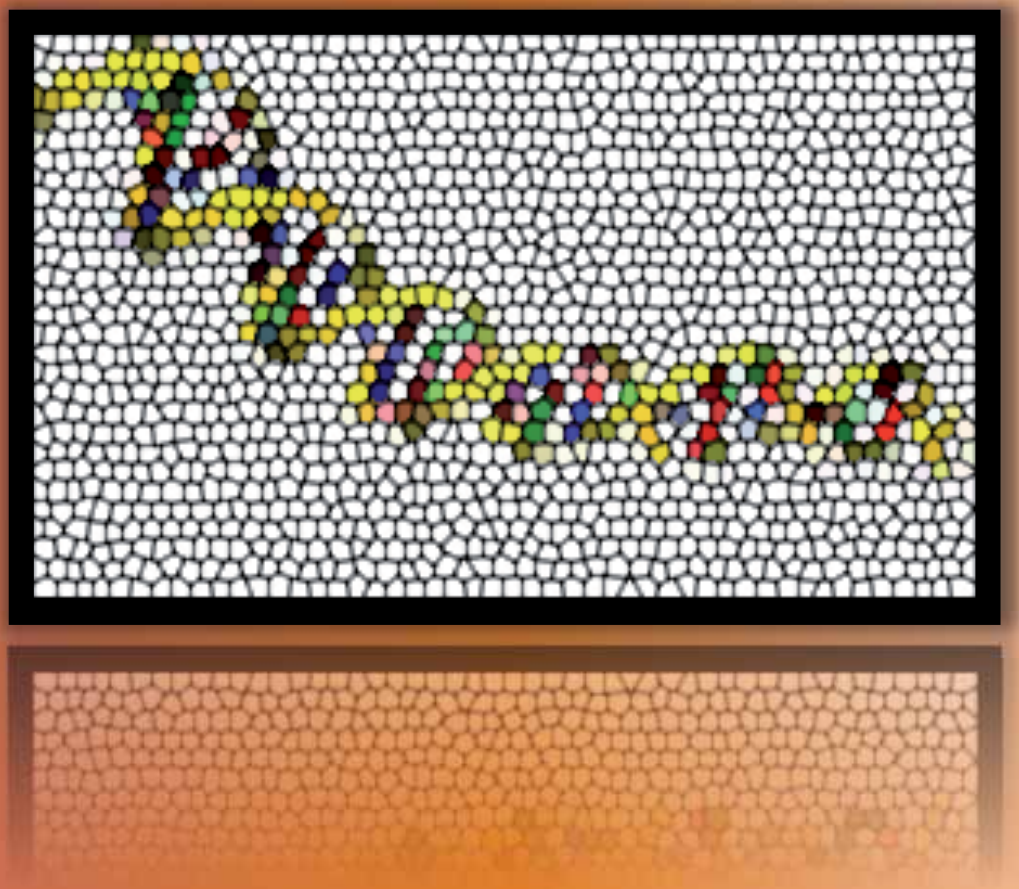
Functional impact of polymorphic inversions in the human genome

Effect on gene expression patterns



Meritxell Oliva Pavia

Doctor of Philosophy – Universitat Autònoma de Barcelona – 2014

“I've paid my dues
Time after time.
I've done my sentence
But committed no crime.
And bad mistakes –
I've made a few.
I've had my share of sand kicked in my face
But I've come through.

[…]

But it's been no bed of roses,
No pleasure cruise.

[…]

We are the champions.”

Queen

“The most practical solution is a good theory.”

Albert Einstein

“The good thing about science is that it's true whether or not you believe in it.”

Neil deGrasse Tyson

Functional impact of polymorphic inversions in the human genome

Effect on gene expression patterns

Memòria presentada per

Meritxell Oliva Pavia

per optar al grau de

Doctora en Genètica per la Universitat Autònoma de Barcelona

Aquesta Tesi Doctoral ha estat realitzada sota la direcció del

Dr. **Mario Cáceres Aguilar**

a l'Institut de Biotecnologia i de Biomedicina de la

Universitat Autònoma de Barcelona

Mario Cáceres Aguilar — El Director de la Tesi

Meritxell Oliva Pavia — L'Autora

Doctor of Philosophy – Universitat Autònoma de Barcelona – 2014

CONTENTS

Acknowledgements ii

List of Abbreviations and Acronyms iii

Index of figures iv

Index of tables v

Preface vi

I INTRODUCTION 3

Inversions overview 5

1.1 Inversion classification and mechanisms of origin 5

1.2 Inversion effects on recombination, evolution and speciation 8

1.3 Inversion positional and mutational effects on genes 13

1.3.1 Direct mutational effect 14

1.3.2 Positional effect 15

1.3.3 Predisposition to further rearrangements 17

1.4 Methods for inversion discovery and genotyping 18

1.4.1 Traditional methods 18

1.4.2 State-of-the-art genomic methods 19

The human genome 29

2.1 Genomic variation in the human genome 29

2.1.1 HapMap project 31

2.1.2 1000 Genomes Project 31

2.1.3 HGSV project 32

2.1.4 Databases and repositories of structural variation 33

2.2 Function of the human genome 35

2.2.1 Genetic determinants of gene expression: expression Quantitative Trait Loci (eQTLs) 36

2.2.2 eQTLs in humans: studies, databases and repositories 37

2.2.3 eQTLs and disease 41

2.2.4 eQTL methods 44

Inversions in the human genome **45**

3.1 Overview 45

3.2 Inversion origins in humans: mechanisms of formation and recurrence of inversion mutational events 46

3.3 Human polymorphic inversions: well-studied cases 49

3.3.1 8p23.1 Inversion 50

3.3.2 17q21.31 Inversion 51

3.4 Inversions and human diseases 53

3.4.1 Genome-wide association studies 54

3.5 Expanding knowledge of inversions in the human genome: the InvFEST project 55

II OBJECTIVES **59**

III MATERIALS AND METHODS **63**

IV RESULTS **79**

Chapter 1 **81**

Refining a catalogue of human polymorphic inversions **81**

1.1 Benchmarking of GRIAL against alternative PEM-based methods for inversion prediction 83

1.2 Refinement of the inversion catalogue: filtering false positive predictions and validation of inverted region candidates 89

1.2.1 PCR amplification of inversion predictions supported by duplicated fosmids. 90

1.2.2 Artefactual fosmid paired-end detection by misspriming analysis 91

1.2.3 Detection of false positive inversion predictions by remapping of fosmid paired ends 95

1.3 GRIAL inversion candidates validation 99

1.3.1 HsInv0409 100

1.3.2 HsInv0410 102

Chapter 2 **105**

Functional impact of polymorphic inversions on gene expressions **195**

2.1 Selection of candidate polymorphic inversions and predicted functional effects 107
2.2 LCL differential expression analysis 112
2.3 LCL DE analysis on well-studied inversions 120
2.4 Inversion-eQTL analyses 127

Chapter 3 **135**

Characterization of candidate inversions with functional effects **135**

3.1 HsInv0058 inversion 137
3.1.1 Recurrence, population distribution and evolutionary history 139
3.1.2 Functional characterization 140
3.2 HsInv0124 inversion 155
3.2.1 Recurrence, population distribution and evolutionary history 155
3.2.2 Functional characterization 157
3.3 Other inversion candidates 165

Chapter 4 **175**

Inversions and disease **175**

4.1 Meta-analyses of disease and complex phenotype GWAS variants associated to inversions 177
4.2 Inversion candidates 180

V DISCUSSION **185**

Refining the catalogue of human polymorphic inversions 187
Methodology 189
Functional impact of polymorphic inversions on gene expression 192
Inversions and disease 208

VI FUTURE DIRECTIONS **211**

VII CONCLUSIONS **215**

VIII APPENDIX **219**

IX BIBLIOGRAPHY **229**

Acknowledgements

Finalment. Semblava que no havia d'arribar mai. Després d'una ingesta quantitat de temps escrivint, corregint, re-escrivint, re-corregint i re-re-escrivint aquesta tesis, és un plaer indescriptible col·locar el plomí (virtual, és clar) en aquesta secció.

Per on començo? Crec que l'inici del recorregut científic personal que dóna lloc a aquest treball es remunta a la decisió de cursar el màster de Bioinformàtica de la UPF (BIOINFO, 2007-2009). En aquests dos anys, a part d'adquirir els coneixements bàsics per poder fer aquesta tesis, vaig tenir el plaer de conèixer una colla d'entranyables "bioinfos" que em van acompanyar en els meus periples per comprendre l'apassionant (o no) món dels "scripts", "pipelines" i "workflows". Ignasi i Pau: gràcies per ensenyar-me la diferència entre "hash" i "array", i a tu, Bàrbara, per compartir amb mi hores d'estudi i presentacions, a vosaltres Laia i Mª Dolors per no rebentar-me els timpans (tot i intentar-ho a força de crits aguts), també gràcies a tu Cristian per aportar la nota de color a l'hora de dinar amb els teus macarrons radioactius, i finalment a tu Eva, per ser única. Tu ets la culpable que relacioni bicicletes mal aparcades amb biologia estructural… No em vull oblidar de certs professors/coordinadors de BIOINFO als qui estic molt agraïda en varis aspectes: gràcies Arcadi per oferir-me el meu primer treball remunerat en recerca, que em va donar la oportunitat d'aprendre sobre genètica de poblacions i compartir, amb la gent del teu laboratori i de tot BIOEVO, interessants discussions científiques a part de xocolatines i dolços després de dinar cada dos per tres... Gràcies Olga, Rui, Ángel, Txema, Valeria, Belén, Hafid (mil gracies per tot H), Óscar, Marc, etc. Torno amb els professors de BIOINFO: gràcies Cedric per presentar l'alineament múltiple de seqüències d'una forma tan apassionant, gràcies Eduardo per fer-nos entendre les cadenes de Markov amb divertits exemples de casinos deshonestos, i finalment, merci Jordi Villà per ser una persona optimista, motivant i motivada.

Després del laboratori de l'Arcadi, vaig aterrar al grup de Genòmica Comparativa del CRG liderat pel Cedric Notredame, on vaig fer les pràctiques del màster i vaig posar i treure el peu en un 1r doctorat. Dono gràcies a en Cedric per l'oportunitat donada, i també a tots els membres del seu grup i estimats "geeks" amb qui vaig coincidir: Ema B., Ema R., Jia-Ming, J.F., Paolo, Ionas, Giovanni, Mathias, Isabel, Thomas i sobretot a tu, Carsten, per ser tan bon amic. Merci per haver tingut

la paciència de respondre les meves preguntes, donar-me un cop de mà quan em feia falta, per ser uns excel·lents companys, per participar en el "lunch club" (Jia-Ming: mai no oblidaré la teva sopa de fideus amb vedella taiwanesa. Mmmm...) i en general per ser com sou! Gràcies també als PIs i membres dels laboratoris del departament de Bioinformàtica i Genòmica del CRG: mencions especials per a en Toni Gabaldón (per ser una persona tan propera, a més a més del PI més "fiestero"), Fyodor Kondrashov (encarnació de la definició "científic" en el seu estat pur: digueu-me, qui més té articles amb el seu pare i la seva àvia???) i també per a en Roderic Guigó per dirigir amb entusiasme el departament i per donar-me consells científics (i personals) quan en necessitava. També no em vull oblidar d'en Juan Valcárcel i la Isabel Vernós, per la seva tasca de donar suport els estudiants de doctorat (jo inclosa) quan en necessiten, als "sysadmin" Óscar i la Judit, per donar-me suport (bio)informàtic, i en general, a tothom qui va contribuir a fer que la meva estada al CRG fos memorable.

Al CRG vaig aprendre quelcom molt important: la recerca surt molt millor si es combina amb una "beer party" cada últim dijous de mes. Merci a tots els "CRGians" amb qui he compartit tans bons moments i converses científicoetíliques. M'agradaria anomenar-los a tots, però tinc certes llacunes... Si un encara vol veure augmentada la seva productivitat de manera espectacular, està clar que ha de participar en al torneig de vòlei platja del PRBB!!! Merci a totes les persones amb qui he compartit equip ("Geeks", "Not only Geeks", "Pringats") i/o m'he enfrontat! Mai no m'hauria imaginat que guanyaria una lliga anomenada "Masters of Disaster" o "Quasi-que-no Cracks"; o que m'emocionaria al sentir aficionats (gràcies Colin, Sara, Sònia, Anne, Eric etc.) cridar fins a esgargamellar-se "Allez les geeks!!". Qui em diria, també, que al CRG coneixeria la meva segona família? Mil gràcies a les meves "nenes", l'Elena i l'Eli, amb qui he compartit de tot: rialles, llàgrimes, experiències nipones i en general, una amistat que espero que duri per sempre, fem el que fem i estem on estem. Gràcies a en Marc, per ser un molt bon amic amb qui he compartit xerrades de ciència i política tot assaborint experiències enogastronòmiques. Gràcies a en Peter per haver-me fet gaudir de moments inoblidables. Gràcies a l'Almer, home-enciclopèdia i inspirador d'un entrepà amb nom forani i a en Raik, el científic més aventurer i poc ortodox que he conegut. Gràcies a la Johanna per ensenyar-me que ser científica i "sex-symbol" no és incompatible. Gràcies a l'Ester pels bons "tannat-moments". Gràcies a la Kiana, en Toni, en Tobias i la Camila. Gràcies a tots, nois!

Vet aquí que arribo al vaixell on he navegat (a vegades amb rumb fix i a vegades a la deriva) al llarg d'aquesta tesis: el grup de Genòmica Comparativa i Funcional de l'IBB, a la UAB. Dono gràcies a tots els "grumets" amb qui he coincidit en aquesta travessia: a l'Ester, la "crack" il·luminada de les llibreries genòmiques, a la Marta per l'ajuda prestada al "wet lab", a en David I. per la seva capacitat organitzativa tant al laboratori com fora (merci per organitzar les barbacoes i altres events de grup/departament!), a en Sergi per ser compartir alegries i penes amb la arxifamosa HsInv0102 i altres inversions, errors i encerts genotípics i per ser un alumne avantatjat en aprendre bioinformàtica a marxes forçades! Gràcies a l'Àlex, per la seva gran capacitat d'oratòria i per ser el meu company de "benchmarking" de `GRIAL` (entre els dos hem venut la moto que va molt bé, eh? Shhh...). Gràcies a la Sònia per l'ajuda bioinformàtica prestada i per la paciència i serenor infinites que té. Gàcies a eixe xic, l'Ignasi, per ensenyar-me un poc de valencià i per donar-me consell estadístic i suport bioinformàtic sempre que l'he necessitat. Gràcies a na Magda, una al·lota ben trempada que no se va cansar d'explicar-me una i altra vegada que rediantres és el "linkage desequilibrium". Merci a la Carla per la seva curiositat i simpatia innata i a en David C. per ser un "crack" de la genètica de poblacions. Merci a en David V. per ser encarregar-se tan bé del "crick" i facilitar-nos la vida a tots els bioinformàtics del grup (i també al "xino" pel suport donat). Merci a la Lorena per ser tan competent i per donar-me un cop de mà sempre que l'he necessitat. Gràcies al capità del vaixell, en Mario, per fer possible INVFEST i per haver supervisat aquesta tesis. Gràcies Mario per tenir la porta del despatx sempre oberta per donar-me consells i resoldre dubtes, destaco la teva tenacitat, gran capacitat de treball i amor per la ciència. GRÀCIES a tots els membres del grup. Sou bons científics i millor persones.

Dins la UAB però fora del grup, vull donar gràcies a l'Alfredo i l'Antonio del departament de Genètica per les seves xerrades filosòfiques i motivants sobre genètica de poblacions. També a la gent que des de consergeria m'han facilitat la vida a l'hora de fer paperassa (Maite) o acudir al meu rescat per entrar/sortir de l'IBB sense targeta (Miguel). Gràcies al la família CBATEG per compartir espectaculars dinars de "tupper" a Medicina; especial menció a en Luca, per venir amb "delicatessen" cuinades per compartir amb mi i a en Carles, per oferir-se sempre a anar a buscar el tallat de rigor després de l'àpat. Merci a la científica més "chic" de tot l'IBB i potser de la UAB, l'Àngels (akas Angelina), moltíssimes gràcies per ser al meu costat quan t'he necessitat i per inspirar-me amb el teu calçat extravagant amb estampats exòtics. Merci a tota la gent de l'IBB i organitzadors de la "Fondue". Merci també a tots els jugadors de bàsquet de dilluns, per haver-me proporcionat

molt bones estones de joc “canyero” (els jovenets) o més relaxat (sèniors). Dono també gràcies a la gent que des de fora la UAB ha contribuït a la realització d’aquesta tesis amb el seu input: mil gràcies Robert i Tomàs per donar-me consell i “feedback” quan em feia més falta, i també a tots els membres del grup d’en Tomàs amb qui he coincidit: Javi, Marc, Tiago, Marcos, Jessica, Irene, Belén, etc. Merci al meus companys de “zulo” Guillem i Maria. Merci a la gent de Cambrige: al professor J. Trowsdale, a la Jyothi a en Clemens i a en John. També agraeixo al “Cambridge Mixed Social Basketball Club” haver-me acceptat com a membre, he disfrutat com una “enana” cada partit jugat.

També voldria agrair a tots el qui que no tenen relació amb el meu doctorat i/o amb les institucions en les que he estat, però que m’han donat el seu suport en tot moment i han estat amb mi aquest temps. Gràcies als amics biotecnòlegs de la meva promoció: Jordi, Gina, Irma, Esther, Helena, Ponsi, Bernat, Rodri, Laura C., Laura M., Laura S., Marta, Marc, Borja, Mercè, Pou, etc. Moltíssimes gràcies als Catalins&Catalines per haver brindat els èxits amb mi (i també ofegat els fracassos): Jordi L., Toni, Ona, Jordi V., Georgina, Johan, Laurence, Tessa, Helena, Enric, Mireia, etc. Gràcies també a la colla de Calaf, que em coneixen i suporten des que era un marrec: Lídia (akas. alcaldessa), Planella, Patri, Stany, Rusita, Rous, Graells, Fiti, Vero, Judit i Anna T. Gràcies a les meves companyes de pis: Maggie, Payal i Anna, per la convivència d’aquests últims anys, sobre tot a la “rubia”. Gràcies també a en Vishal (akas Vishi) , una persona molt especial i font d’inspiració permanent, benvinguda la casualitat que ens va fer creuar.

Finalment, voldria donar les gràcies a la meva família, que heu estat, sou i sereu indispensables en la meva vida. Papa, mama, Eva i iaia Ción, no tinc pas prou paraules d’agraïment, sobretot per vosaltres, pares. Merci per donar-me la moral per arribar fins al final, gràcies per les vostres paraules amables en els moments durs, per ser al meu costat en tot moment (malgrat no compartir sovint el meu criteri), per ser la força invisible que m’empeny endavant, i en general, per acceptar-me i apreciar-me tal com sóc. Us ho dec tot i us estimo. A vosaltres dedico aquest treball.

GRÀCIES

List of Abbreviations and Acronyms

BP	Breakpoint
CNP	Copy number polymorphism
CNV	Copy number variant
DE	Differential Expression
DECIPHER	Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources
DGV	Database of Genomic Variants
DGVa	Database of Genomic Variants archive
DNA	Deoxyribonucleic acid
DSB	Double-strand break
DSBR	Double strand break repair
EBV	Epstein-Barr virus-transformed
ENCODE	The Encyclopaedia of DNA Elements Project
FISH	Fluorescence in situ hybridization
FoSTeS	Fork stalling and template switching
Fst	Fixation index: a measure of population differentiation due to genetic structure.
GRC	Genome Reference Consortium
GTEx	Genotype-Tissue Expression Project
GWAS	Genome-wide association studies
HET	Heterozigote for standard and inverted allele
HGP	The Human Genome Project
HTS	High-throughput sequencing
INV	Homozigote for inverted allele
Inv	Inverted allele
ISV	Intermediate-sized structural variation
kb	Kilobase
LCL	Epstein-Barr virus-transformed B-lymphoblastoid cell line
LCR	Low copy repeat
LD	Linkage disequilibrium
LSV	Large-scale structural variation
LogFC	Logarithm base 2 of fold change
LTR	Long Terminal Repeat
Mb	Megabase
MEPS	Minimal efficient processing segments
MMBIR	Microhomology mediated break induced replication
MMEJ	Microhomology-mediated end-joining
NAHR	Non allelic homologous recombination
NGS	Next generation sequencing
NH	Non-homology
NHEJ	Non-homologous end-joining
PEM	Paired-end mapping

PFGE	Pulsed-field gel electrophoresis
PR	Paired Read
RD	Read depth
RNA	Ribonucleic acid
RNA-Seq	RNA Sequencing
SD	Segmental duplication
SINE	Short interspersed nuclear elements
SNP	Single nucleotide polymorphism
SP	Split Read
STD	Homozygote for standard allele
Std	Standard allele
STR	Short tandem repeat
SV	Structural variant
TSS	Transcription start site
aCGH	Microarray comparative genomic hybridization
alt-EJ	Alternative end-joining
bp	Base pair
eQTL	Expression quantitative trait loci
indels	Short insertions and deletions
lincRNA	Long intergenic non-coding RNA
lncRNA	Long non-coding RNA
1000GP	1000 Genomes Project

Index of figures

Figure 1 - Classification of inversions _____5
Figure 2 – Genomic rearrangements mediated by NAHR _____7
Figure 3 - Schematic diagram showing the suppression of recombination in an inversion heterozygote _____9
Figure 4 – Suppression of recombination _____10
Figure 5 - Inversion effects on gene coding sequences _____14
Figure 6 - Possible mechanisms causing a positional effec _____16
Figure 7 - Inversion detection techniques _____20
Figure 8 - PEM technique _____22
Figure 9 - FISH genotyping of inversion polymorphisms _____28
Figure 10 - Classes of structural variation _____30
Figure 11 - Summary of the InvFEST database content _____35
Figure 12 - DGV content _____46
Figure 13 - Mechanisms of formation of inversions _____48
Figure 14 – Frequency of inversions in HapMap population _____58
Figure 15 – VH and PEMer inversion BP inference _____68
Figure 16 - Comparison of GRIAL with PEM methods inversion predictions _____85
Figure 17 - BP detection accuracy by SV prediction methods _____89
Figure 18 - Artefactual paired-end fosmid mapping _____92
Figure 19 - Problematic fosmid paired-end read sequence _____92
Figure 20 – Sequence quality of problematic reads _____93
Figure 21 - Fosmid library paired-end sequencing _____94
Figure 22 - Remapped discordant in signal fosmids _____96
Figure 23 - HsInv0306 and HsInv0710 region _____99
Figure 24 - HsInv0409 inversion regio _____100
Figure 25 - HsInv0409 validation _____101
Figure 26 - HsInv0410 inversion region _____103
Figure 27– HsInv0410 validation _____103
Figure 28 – Inversion and BP size distributio) _____107
Figure 29 - LCL expression datasets _____113
Figure 30 – LCL DE Pipeline _____116
Figure 31 – Replicability of LCL DE results across datasets _____116

Figure 32 - Overview of LCL DE analysis results 119
Figure 33 - 17q21.31 differentially expressed genes 124
Figure 34 - 17q21.31 structural haplotypes 125
Figure 35 – Scheme of the pipeline to find inversion-eQTLs in non-LCL derived tissue 128
Figure 36 – Inversion eQTLs distribution 131
Figure 37 – Overview of MHC region 137
Figure 38 – Mapping of clone AC207175.3 against HG19 and HsInv0058 complex rearrangement – 139
Figure 39 - Top LCL DE analysis candidates of expression association with HsInv0058 genotype 142
Figure 40 - Expression of HsInv0058 gene candidates in multiple tissue 145
Figure 41 - HCG22 expression association with HsInv0058 genotype in multiple tissues 149
Figure 42– Linkage disequilibrium of HCG22 eQTLs with putative causal variants 152
Figure 43 - Linkage disequilibrium of HCG27, HLA-B, HLA-C eQTLs with putative causal variants - 153
Figure 44 - HsInv0124 genomic region 156
Figure 45 – LCL expression of top DE analysis candidates for HsInv0124 162
Figure 46 - IFITM2 and IFITM3 expression profile in multiple tissues 165
Figure 48 – LCL expression of RHOH inverted exon association with HsInv0102 genotype 168
Figure 49 – HsInv0059 complex rearrangement 169
Figure 50 – Expression of GABRR1 in multiple tissues 171
Figure 51 - Expression of SPINK14 in multiple tissues 172
Figure 52 – SPINK14 ~ HsInv0201 and GABRR1 ~ HsInv0059 associations in multiple tissues 173
Figure 53 - INVFEST predictions overlapping genes 191
Figure 54 – CTRB1 and CTRB2 eQTL data 199
Figure 55 – LD in HsInv0124 region 202
Figure 56 – HsInv0058 DE candidate-genes protein-protein interactions 205
Figure 57 - SPINK14 coding sequence 206
Figure 58 – SPINK6 eQTL association 207
Figure 59 - NLGN4X expression profile in multiple tissues 211
Figure 60 – Expression of genes contained or overlapping inversions 221

Index of tables

Table 1 - Inversions predispose to disease _____18
Table 2 - seeQTL summary table of eQTLs in each dataset _____41
Table 3 - Scheme of fosmid PEMs classification after remapping _____73
Table 4 - GRIAL unique inversion predictions_____86
Table 5 - Benchmarking of different inversion prediction methods against gold-standard inversion dataset_____87
Table 6 - Duplicated fosmids features _____91
Table 7 - Summary of characteristics of the inversion set _____108
Table 8 – Genes contained in inversions_____110
Table 9 – Genes overlapping inversion BPs _____115
Table 11 – Filtered LCL DE candidates _____118
Table 12 – 17q21.31-inv differentially expressed genes _____123
Table 13 - 17q21.31 rearrangements differentially expressed genes _____123
Table 14 - 8p23.1-inv differentially expressed genes_____127
Table 15 - Characteristics of eQTL datasets - _____130
Table 16 – Inversion-eQTLs gene associations _____133
Table 17 - HsInv0058 distribution in HapMap populations _____140
Table 18 - HsInv0058 top LCL DE analysis candidates _____141
Table 19 – HsInv0058 top inversion-eQTL candidates _____143
Table 20 – HsInv0058 candidate genes eQTL associations and eQTLs linked with HsInv0058 _____151
Table 21 – Correlations of candidate genes eQTL scores with putative causal variants _____151
Table 22 – HsInv0124 distribution in HapMap populations _____157
Table 23 – HsInv0124 top LCL DE analysis candidates _____160
Table 24 – HsInv0124 top inversion-eQTL candidates _____161
Table 25 - Top LCL DE analysis candidates _____167
Table 26 - Inversions with tag-SNPs associated to diseases and complex phenotypes_____178
Table 27 - Inversion tag-SNPs associated to diseases and complex phenotypes and effects on gene expression _____179
Table 28 – HsInv0058 linked variants associated to disease and complex traits_____181
Table 30 – Transcripts overlapping inversion BPs _____227

Preface

An inversion is a genomic rearrangement that alters the orientation of a specific genomic sequence. Inversions are balanced rearrangements and do not involve a gain or loss of DNA (at least no significant change in theory, although in practice losses at breakpoints are commonly found), but it still can alter the original genetic background which may have consequences.

Inversions constitute the first type of structural variants (SVs) identified by the inventor of genetic mapping, Sturtevant (1921). Over the next half century, these rearrangements were extensively studied (particularly in *Drosophila* species) due to their particular features and implications. Inversions can exert important functional effects directly, by mutational and positional effects on genes, or indirectly, by imposing new regimens of molecular evolution both on the DNA sequences encompassed by them and in their vicinity. Therefore, inversions may associate with certain phenotypes or diseases, shape the evolutionary fate of the carriers by adaptive processes or even play a role in the origin of new species. For all that, increasing our knowledge of inversions constitutes a key issue for in-depth understanding of genome variation and its consequences.

Throughout the last century history, technological breakthroughs have heavily impacted the study of inversions: interest in inversions decreased together with the popularity of empirical population genetics starting in the 1970s, with the appearance of biochemical and then molecular genetics. Conversely, for the last 10 years, attention on inversions is ascending again due to the advent of novel genomic technologies that have enabled the study of SVs, including inversions, in a high-throughput fashion within and across species.

Similarly to the genomic variation field, transcriptomics has undergone major advances in the last 20 years. Thanks to the appearance of different technologies, from the initial sequencing of expression sequence tag (EST) strategy to gene chips, and most recently RNA-Seq, it is nowadays possible to accurately measure gene expression genome-wide in a cost-effective manner. Consequently, of the recent large efforts for the analysis of genome variation and genome expression empower the study of genome function. One of the most prominent outcomes in this direction has been the publication of a large set of studies on genetic variants that explain

variation in gene-expression levels, named expression quantitative trait loci (eQTLs). Such studies have enhanced the comprehension of basic mechanisms of gene regulation and interpretation of genome-wide association studies. Although most of eQTL studies have focused on single nucleotide polymorphisms (SNPs), the study of other genetic variants, such as indels or inversions, is of particular interest especially if they are not tagged by SNPs and have potential functional implications.

For all that, we think that state of the art of genomic research provides the means and tools to explore the human genome in detail and expand our knowledge of inversions. The aim of this thesis has been: i) help build the most accurate and exhaustive catalogue of human polymorphic inversions to date; ii) gain further insight on the functional impact of polymorphic inversions in the human genome, particularly on the inverted rearrangements that modulate gene expression (inversion-eQTLs); iii) elucidate the mechanisms by which the effect on gene expression occurs; and iv) find possible associations of inversions with disease.

I INTRODUCTION

Inversions overview

1.1 Inversion classification and mechanisms of origin

Inversions are classified according to several criteria (**Figure 1**). For instance, inversions can be pericentric or paracentric depending on the inclusion of the centromere. In pericentric inversions, the centromere is included in the inverted region, and in meiosis, a single crossover event that takes place between the breakpoints of a heterozygote inversion carrier produces unbalanced gametes that carry deletions or insertions. This event can cause a reduction of fertility, making the inversions underdominant (lower heterozygote fitness) and purged by purifying selection. However, not all inversions produce a detrimental effect, as in *D. melanogaster*, some paracentric inversions like the cosmopolitan inversion In(3R)Payne[1], apparently escape fitness costs when heterozygous, possibly because they suppress recombination and confer selective advantage to the carriers by holding together favorable combinations of alleles that act together to facilitate adaptive shifts (Kennington, Partridge, and Hoffmann 2006).

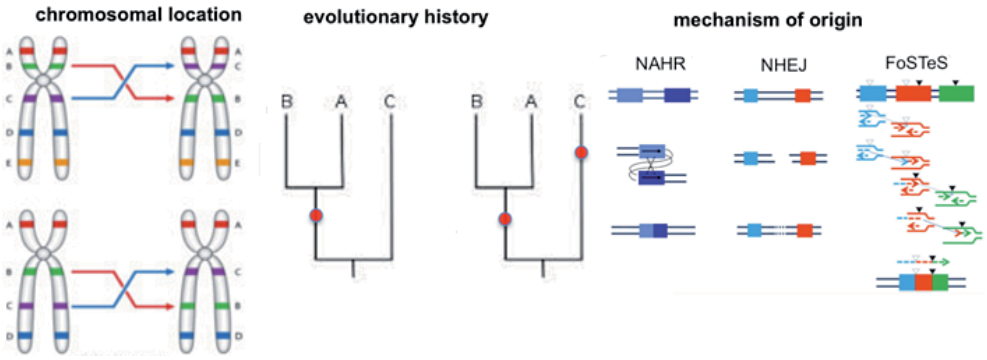


Figure 1 - Classification of inversions – Left panel contains schematic representations of a pericentric (top) and a paracentric (bottom) inversions. Middle panel, contains a schematic representation of a unique (left) and recurrent (right) inversion event in the evolutionary history of species A,B,C. Right panel contains a scheme of different mechanisms that may originate inversions, such as NAHR (left), NHEJ (middle) and FoSTeS (right).

[1] Well-studied inversion of *Drosophila melanogaster*. The rearrangement spans several megabases in size, includes several percent of the entire genome and contains hundreds or thousands of genes.

Paracentric inversions do not include the centromere in the inverted region. In this case, a crossover between the breakpoints produces unbalanced gametes that carry deletions or insertions, acentric and dicentric gametes. Unlike pericentric inversions, many of the paracentric inversions segregating in nature do not suffer from underdominance in *Drosophila* and other insects. Hence, paracentric inversions are far more common than pericentric ones, both as polymorphisms within and fixed differences between species (Hoffmann and Rieseberg 2008).

Another criteria to classify inversions is based on their chromosomal location; whether they are in autosomal or sex chromosomes (chromosome X or Y in case of humans and other mammalian species). A widely accepted theory postulates that inversions in sex chromosomes play an important role in the evolution of sex chromosomes (see section I1.2).

We can classify inversions on the basis of the amount of sequence similarity at the breakpoints of the SV and this is related to the different mechanisms involved in the generation of the inversions. Homology based mechanisms, such as non-allelic homologous recombination (NAHR), are mediated by high sequence similarity at the breakpoints. Contrarily, non-homology (NH) or microhomology based mechanisms, such as no homologous end-joining (NHEJ), microhomology-mediated end-joining (MMEJ), fork stalling and template switching (FoSTeS) (Lee, Carvalho, and Lupski 2007) or microhomology-mediated break-induced repair (MMBIR) (Shaikh et al. 2000) require little or no sequence similarity. It is pertinent to note that besides the mentioned ones, there may be other currently unknown molecular mechanisms that can generate inverted rearrangements apart from NAHR and NH processes.

NAHR is similar to the normal biological process of homologous recombination that occurs during meiosis and exchanges DNA between two homologous chromosomes (**Figure 2**). However, as the name states, NAHR is a rearrangement that occurs between homologous sequences that are not the same allele on homologous chromosomes and that can generate several kinds of SVs (Darai-Ramqvist et al. 2008; Kidd et al. 2008). In the case of inversions though, the homologous sequences need to be inverted repeats (IRs). Different types of homologous sequences are natural substrates for NAHR, spanning different lengths and having different origins.

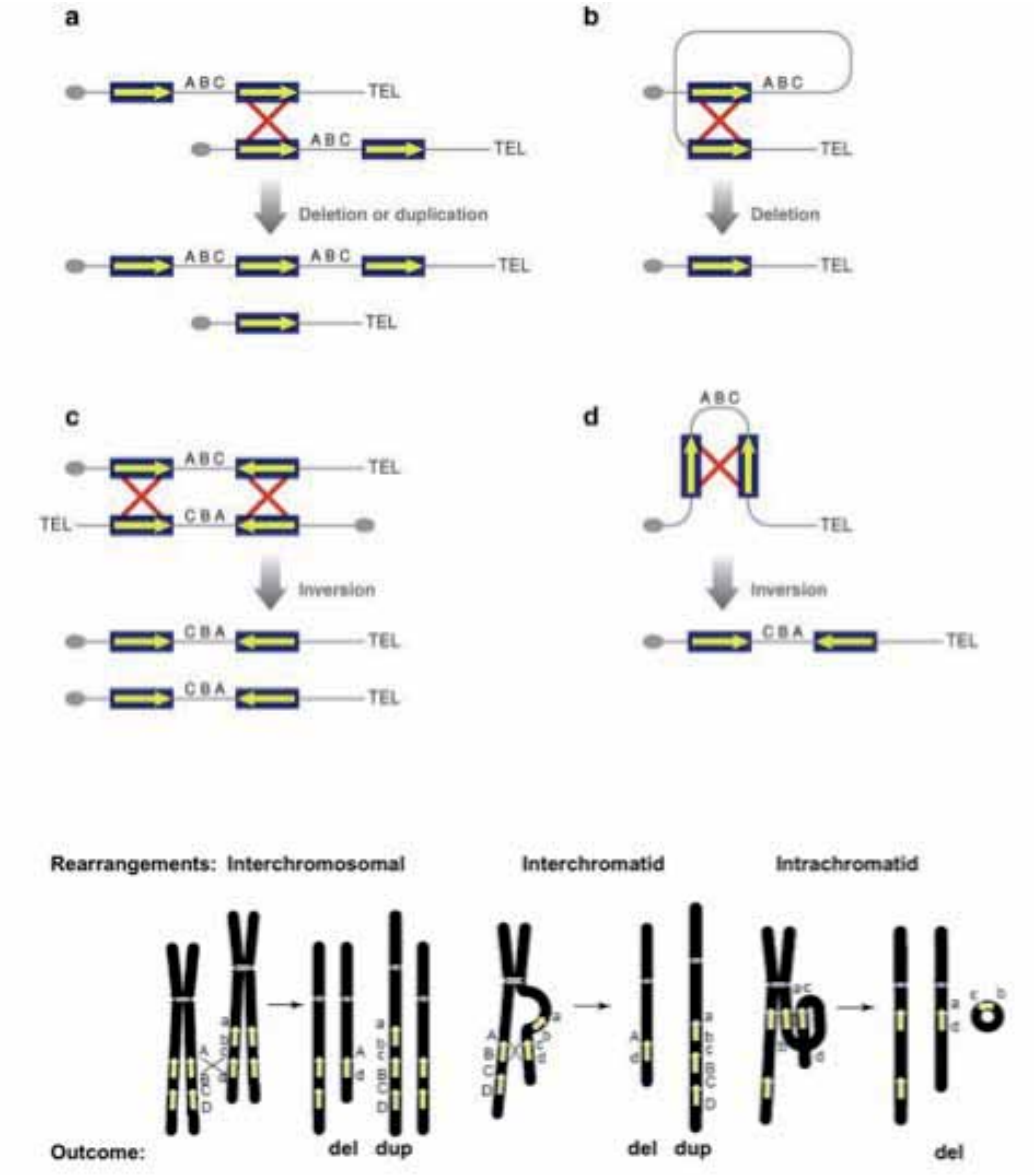


Figure 2 – Genomic rearrangements mediated by NAHR – Top panel: (***a*** and ***b***) Interchromosomal, intrachromosomal, or intrachromatid NAHR between directly orientated repeats causes deletion and/or duplication of the intervening sequence. (***c*** and ***d***) Interchromosomal, intrachromosomal, or intrachromatid NAHR between inverted repeats causes inversion of the intervening sequence. Repeat sequences are depicted as black boxes, with their orientation indicated by yellow arrows, and recombination is shown by red crosses. Bottom panel: Scheme of interchromosomal, interchromatid and intrachromatid NAHR-mediated rearrangements that cause deletions (del) and duplications (dup). Adapted from Gu at al. (Gu, Zhang, and Lupski 2008; Sharp, Cheng, and Eichler 2006).

As mentioned, there are other mechanisms for the formation of SVs (see Stankiewicz and Lupski (2010)), and the distinction between homology mediated and non-homologous mechanisms may not be so straightforward to identify. Inversions arising putatively from NHEJ events, resulting from random (or near random) double-strand breaks, display some degree of microhomology (e.g. 2–25 bp of sequence similarity) at their breakpoints. Other mechanisms such as FoSTeS have also been proposed. FoSTeS occurs when the active replication fork stalls and switches templates using complementary template microhomology to anneal and prime DNA replication. As a result, there could be interrupted duplications in which stretches of DNA of normal copy number were punctuated by stretches of DNA that were amplified two or three times (Lee, Carvalho, and Lupski 2007). The FoSTeS events occur preferentially in regions of complex genomic architecture that contain abundant low-copy repeats with high sequence identity and in various orientations. These regions are prompt to bring into proximity highly similar DNA segments or repetitive sequences that normally lie far apart. Therefore, this context can favor replication of long-distance template-switching models between different replication forks stalling and slippage and, consequently, enables the joining or template-driven juxtaposition of different sequences from discrete genomic positions, generating complex genome structural rearrangements, including inversions (Lee, Carvalho, and Lupski 2007).

1.2 Inversion effects on recombination, evolution and speciation

Sturtevant and others (Krimbas and Jeffrey, 1992) stated and proved that inversions have a dramatic effect on genetic transmission. When heterozygous, inversions form loops during the pairing of chromosomes in meiosis and suppress recombination (**Figure 3**, **Figure 4**) with the exception of double crossovers within large inverted regions and allelic gene conversion events (homologous DNA sequence replacement between alleles of the same gene).

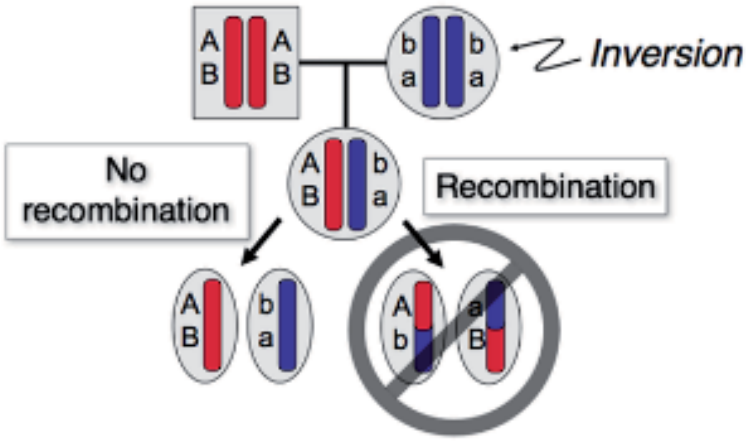


Figure 3 - Schematic diagram showing the suppression of recombination in an inversion heterozygote - Two loci segregate for the alleles (A, a) and (B, b). An individual that is heterozygous at both loci and for the inversion does not produce the recombinant gametes A/b and a/B. Adapted from Kirkpatrick (2010).

Although it is still not totally clear how suppression of recombination occurs, there are two mechanisms that could contribute to this molecular event. The first is a real suppression of recombination due to the difficulty of complete synapsis between the two homologous in the regions at the ends of the inversion loop during meiosis. The second mechanism considers that recombination occurs at a fairly normal frequency within the inversion region. However, the gametes produced from recombination within the inverted region in heterokaryotypes are usually unable to produce viable offspring (Stevison, Hoehn, and Noor 2011).

As discussed, suppression of recombination can be explained by the loss of unbalanced gametes that result from recombination, the lack of synapsis in inverted regions in heterozygotes and probably other mechanisms not yet understood. Recombination is responsible for the genetic shuffling and generation of new allelic combinations upon which selection can act, and therefore is considered a major evolutionary process. Consequently, the effect of inversion variants in suppressing recombination constitutes a key evolutionary effect. Moreover, inversions can affect recombination rates outside the inverted region. For example, in *Drosophila* species, the recombination rate throughout the rest of the genome is significantly increased by inversions (Stevison, Hoehn and Noor 2011). This pattern has also been found in humans for a particular inversion (Chowdhury et al. 2009; Stefansson et al. 2005).

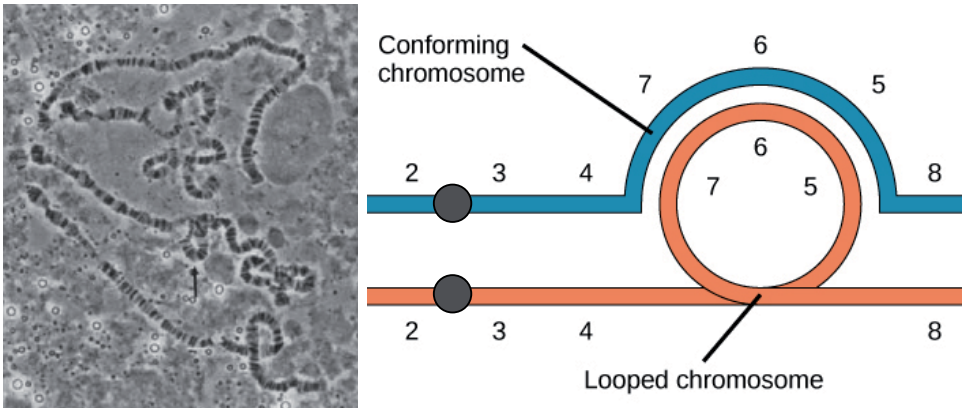


Figure 4 – Suppression of recombination - Left, inversion causing chromosomal loop during meiosis in *Drosophila*. Adapted from Mestres et al. (1998). Right, schematic representation of the synapsis of a paracentric inversion heterozygote. During the process, the inverted region (5,6,7) is incorporated into a loop to maximize synapsis along the length of both chromosomes. A real suppression of recombination occurs near the breakpoints (at the base of the loop) due to the difficulty of synapsing in this region. An apparent suppression of recombination occurs in the rest of the region due to the formation of inviable recombinant chromatids. Centrosomes depicted by grey circles. Adapted from ("Chromosomal Basis of Inherited Disorders" n.d.).

Inversions are thought to shape the evolution of species and participate in speciation processes. Some evolutionary models state that inversions may foster reproductive isolation and play a role in speciation processes in several ways (Kirkpatrick and Barton 2006; Navarro and Barton 2003; Ortíz-Barrientos et al. 2002). The commonality of these models is that they suggest that differences between the two inverted rearrangements accumulate and can evolve to cause reproductive isolation. Observations supporting these new models have been described in several species, including birds, mammals, insects, and plants. In primates however they have not been definitively proven (see Alves et al. (2012) and Hoffman et al. (2008) for a review).

Inversions have also played a key role in the evolution of sex chromosomes. In groups like mammals, the X and Y chromosome do not recombine along almost its entire length. This lack of recombination is thought to be originally caused by inversions: at an early stage of the evolution of mammalian Y, the non-recombining Y genomic region was gradually enlarged by a run of overlapping inversions (Rice 1987). Inversions play a role in creation of non-recombining blocks and may spread if they capture the male-determining factor and a male-beneficial allele. This occurs because of a decrease in recombination between the locus that determines sex (Charlesworth 1991; Rice 1987) and genes under sex-antagonistic selection. Then,

when the Y is isolated from the X, the former evolves as an asexual genetic unit and can degenerate (Bachtrog 2006). Inversions may also be crucial to the very origin of sex chromosomes by contributing to the creation of a neo-sex chromosome that can hijack sex determination from the ancestral sex chromosomes (van Doorn and Kirkpatrick 2007).

The spread of inverted rearrangements can result from a combination of factors largely influenced by a population's demography, ecology and evolutionary history. As in any other mutation, gene flow (e.g. migration), random genetic drift and selection and can play major roles in modulating inversion frequencies and distribution across populations. Most of the inversions, principally the small intergenic ones, are likely to evolve neutrally (by drift alone). Selection can work in three ways. Purifying selection purges inversions with a clear detrimental effect, such as pericentric inversions that can generate structural problems with meiosis or inversions disrupting a gene or altering its expression and causing disease. However, a mutation caused by an inversion can sometimes be adaptive and therefore be selected (positive selection caused by a positional effect). Finally, selection can act on an inversion when it does not generate but rather carries one or more selected alleles (see Kirkpatrick et al. (2010) for a review).

One of the adaptive models that may explain the spread of an inversion is the "coadaptation hypothesis", which states that inversions can contain well-adapted genes that have a more advantageous effect in combination than individually (epistasis). The inverted rearrangement provides conditions favorable for the alleles to become "protected" from introgression, due to lack of recombination, which would cause an increase in frequency of the inverted allele (Dobzhansky 1970). In this case, the selective advantage is not directly related to the new chromosomal structure but to its favorable genetic composition, the haplotype "trapped" by the inversion (Dobzhansky 1970; Krimbas and Powell 1992; Spirito F. 1998). It can also happen that an inversion occurs in a genomic region containing two or more alleles that are well adapted to certain environmental conditions without necessarily interacting in an epistatic manner.

In any case, the selective advantage conferred by coadapted alleles can cause the inversion to spread in the particular environment and/or geographic area where the genes are favored. This scenario is known as "local adaptation" and it is characterized by the inversion showing a clinal distribution (Kirkpatrick and Barton 2006). A good example of the contribution of the inversions to the local adaptation is

the inversion In(3R)Payne in *D. melanogaster*, which shows parallel latitudinal clines on three continents (Hoffmann and Rieseberg 2008). Several examples of local adaptation can be found in other insects such as species of *Anopheles* mosquitos (Ayala, Guerrero, and Kirkpatrick 2013). Local adaptation is found in species of different kingdoms as well, for instance in monkeyflowers (see Kirkpatrick et al. (2006) for a compilation of illustrative examples).

Another example of an phenomenon that can affect the frequency and distribution of the inversions is meiotic drive (White 1978), by exherting a bias in the transmission of inversion alleles. Some inversions show signatures of meiotic drive: the gametes of heterozygotes carry the inversion more than 50% of the time. Meiotic drive systems often involve a pair of interacting loci that must be coinherited for the system to invade a population. This system would be usually disrupted by recombination, but inversion suppresses recombination so meitotic drive may occur and driving alleles are spread.

Inversions can be underdominant (inferior in fitness as heterozygotes). If single crossovers are frequent within the inversion, underdominance might occur. This phenomenon leads to the generation of unbalanced and inviable gametes, which seems frequent for multiple inversions that differentiate plant and insect species, but not seemingly in mammals (Coyne 2004; Rieseberg 2001). The most convincing evidence for chromosomal underdominance comes from recovery of chromosome pairing and fertility following genome doubling of structural heterozygotes, reported in many plant species (Stebbins 1958). Other evidence consistent with the underdominance of inversions includes mapping of underdominant quantitative trait loci (QTLs) to inversion breakpoints also reported in several plant species (Lai et al. 2005).

Conversely, inversions can also be overdominant (superior as heterozygotes) (Dobzhansky 1970). Overdominance can arise from deleterious alleles in the region covered by an inversion. However, the genetic basis for overdominance seems not to have been clearly determined for any inversion. In principle, it could result from mutational effects of inversion breakpoints. On the other hand, overdominance could also occur by one allele of a polymorphic locus contained inside the inverted region becoming fixed on the ancestral chromosome and the other allele on the inverted chromosome, conferring heterokaryotype advantage. A third hypothesis is “associative overdominance", which happens when an inversion captures by chance one or more deleterious recessive alleles, more likely on the case of a large inversion

(Sturtevant and Mather 1938). The inversion can be otherwise selectively favored when rare. Then, it can spread to the point where recessive homozygotes become frequent enough to counteract the initial advantage. As a result, a balanced polymorphism that has the same evolutionary properties as conventional overdominance is obtained. Both conventional and associative overdominance are two different forms of balancing selection, as the inverted rearrangement is actively maintained in the population gene pool with high frequency (see Charlesworth (2006) for a review). Interestingly, inversions show large systematic differences between taxa in the frequency and severity of fitness effects. For example, heterozygotes for inversions seem to show decreased fertility in plants much more commonly than in animals (Hoffmann and Rieseberg 2008). This observation may relate to the different meiotic cost in terms of infertility in inversion heterozygotes across species. For instance, in *Drosophila* females, three of the four gametes become polar cells and only one forms the egg; the recombinant gamete is less likely to become the egg because it migrates more slowly, reducing any cost associated with unbalanced recombinant gametes (Hoffmann and Rieseberg 2008). This factor coupled with the absence of recombination in males might explain the persistence of inversion polymorphisms in *Drosophila melanogaster* and other species.

1.3 Inversion positional and mutational effects on genes

Inversions, similarly to other type of SVs such as deletions or insertions, can exert a functional effect on genes by disrupting gene sequences (mutational effect, **Figure 5**) or by creating fusion genes (**Figure 5**). However, inversions have additional mechanisms that affect genes, either indirectly by inhibition of recombination (see section I1.2) or by positional effects, changing the positional distribution of genes with respect to their regulatory elements and therefore affecting the dynamics of gene expression. In addition, inversions may constitute the substrate for further genomic rearrangements that can affect genes. Therefore, if taken together, inversions effects are more complex and differ from those of other SVs.

1.3.1 Direct mutational effect

One of the possible consequences of inversions is the mutational effect at the breakpoints. The consequences of this break will depend on its location with respect to gene sequences. The rupture of an exon by the breakpoint can impair the functionality of the gene, especially if the exon is protein-coding, and tends to be a deleterious change (**Figure 5**).

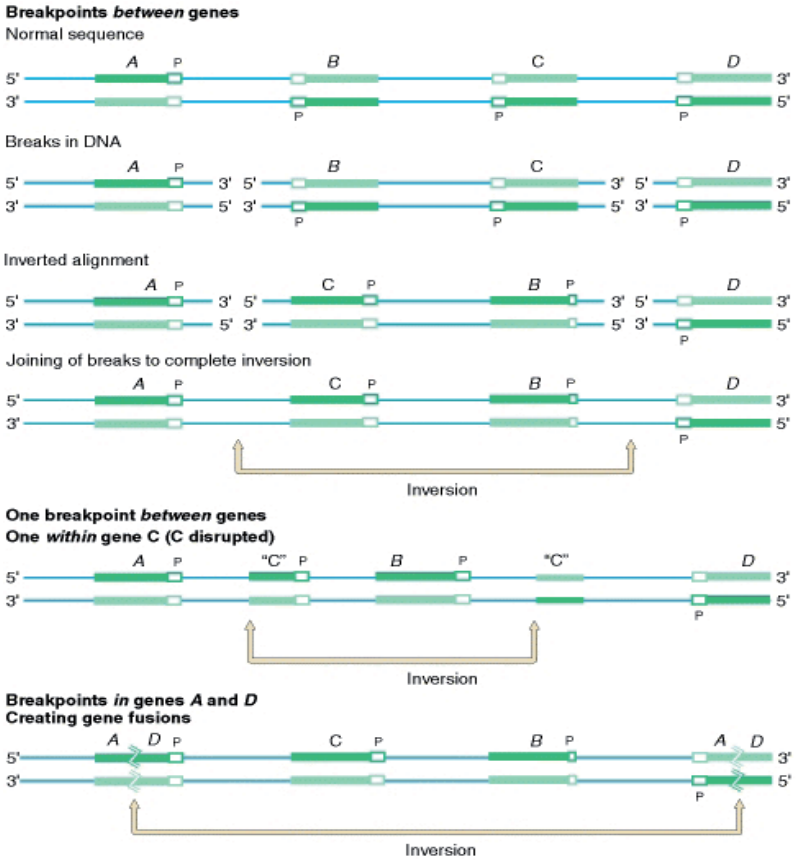


Figure 5 - Inversion effects on gene coding sequences - Effects of inversions at the DNA level. Genes are represented by *A*, *B*, *C*, and *D*. Template strand is dark green; non-template strand is light green; jagged lines indicate break in DNA. The letter P stands for promoter; thick arrow indicates the position of the breakpoint. Taken from Griffiths AJF, Miller JH, Suzuki DT, et al. (2010).

Furthermore, inversion mutational events that alter gene-coding structures by breaking within introns and subsequently reordering the distribution of exons within

the gene might lead into genomic disorder as well. In humans, there is evidence of inversions that disrupt genes and are related to diseases, ranging from single-gene inherited disorders to complex, polygenic pathologies (Feuk 2010). For instance, the most prevalent case is the X-linked disorder caused by an inversion that breaks the factor VIII gene, which gives rise to hemophilia A (Lakich et al. 1993). Another example is an inversion that breaks the IDS iduronate 2-sulfatase gene and causes Hunter syndrome in 13% of diagnosed patients (Bondeson et al. 1995). Finally, inversions have also been associated with certain forms of cancer. For instance, in non-small cell lung carcinoma (NSCLC), an inversion located in the chromosome band 2p23 causes the fusion of *EML4* (echinoderm microtubule-associated protein-like 4) gene with the *ALK* (anaplastic large cell lymphoma kinase) gene in 3% to 5% of the cases diagnosed with NSCLC (Soda et al. 2007). However, 2p21-p23 inversion and other alterations in the region seem to be acquired but not driving mutational events in lung cancer development and progression (Perner et al. 2008).

1.3.2 Positional effect

Even when the location of the inversion within the genome does not break a functional element, it is important to appreciate that inversions can have a significant effect at a distance (Sharp, Cheng, and Eichler 2006). Positional effect is a direct consequence of the inversion due to the relocation of genomic segments from one region to another (**Figure 6**). However, it is not clear from literature whether a direct mutational effect can be included in this category. For the purpose of this dissertation, a position effect is defined as a change in the level of gene expression caused by a change in the position of the gene relative to its normal chromosomal environment, but not associated with an intragenic mutation caused by the inversion.

Position effects can be produced by translocation of a gene into a heterochromatic region, resulting in the methylation of promoter regions and consequent down-regulation of expression (Kleinjan 1998). The phenomenon by which genes become silenced by heterochromatinisation has been called position-effect variegation (PEV), and has been well described in Drosophila as being responsible for changes in eye color (Weiler and Wakimoto 1995). Another described positional effect is caused by intergenic genomic rearrangements that detach a gene from its transcriptional regulatory elements or that bring a gene into proximity to another regulatory element, altering gene expression. A good example

is the study that proved the loss of expression of Hoxd genes during limb development in mice, as well as subsequent phenotypic alterations (Spitz et al. 2005). In this study, the authors induced an inversion that split the mammalian Hoxd gene cluster into two independent pieces (Spitz et al. 2005).

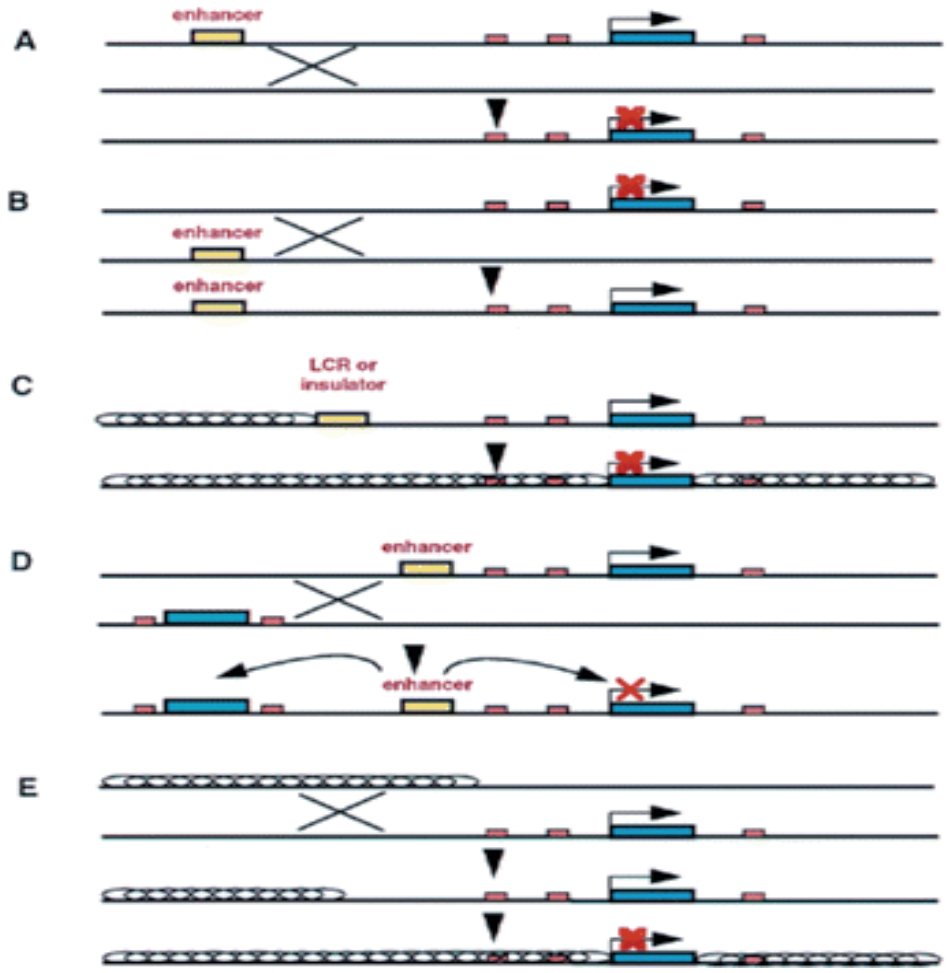


Figure 6 - Possible mechanisms causing a positional effect - A) The chromosomal rearrangement (cross) caused by an inversion (BP indicated by a downwards black triangle) separates the promoter/transcription unit (blue box) from a distant *cis*-acting regulatory element. The removal of an enhancer element will result in complete or partial silencing from the affected allele (red cross). Alternatively, if a silencer element is removed (zigzagged block), allele activation may occur. **B**) Juxtaposition of the gene with an enhancer element from another gene may also lead to activation of gene expression. **C**) Removal of a long-range insulator or boundary element may lead to inappropriate shutting down of the locus. **D**) Enhancer competition. A gene residing at the translocation site competes for interaction with the regulatory element(s) of another gene, thereby reducing its level of expression. **E**) PEV. PEV can occur when a chromosomal rearrangement causes the juxtaposition of a euchromatic gene with a region of heterochromatin. The heterochromatin DNA organization is thought to spread into the juxtaposed euchromatic region, thereby silencing the nearby gene in a stochastic manner. Taken from Kleinjan et al. (1998).

Although the current estimated fraction of the genome that is evolutionarily conserved through purifying selection represents an small portion, around 10%, the recent results of the ENCODE project (see section I2.2) suggest that many other hidden elements are potentially functional, most of which have a role in gene regulation (ENCODE Project Consortium et al. 2007). Thus, inversions cannot be presumed to be functionally harmless or neutral because they encompass only non-coding segments. Rather, a careful assessment of nearby genes that may be affected via a positional effect mechanism needs to be considered.

1.3.3 Predisposition to further rearrangements

The potential effect of inversions might not be directly associated with the alteration of gene expression, either by disrupting coding regions that span the breakpoints or by position effects acting on genes adjacent to the breakpoints. Instead, the real effect of an inversion could be that it can act as a risk factor for other genomic changes (Sharp, Cheng, and Eichler 2006). That is the case of several polymorphic inversions generated between flanking duplications which do not have any direct consequence. It is thought that these result in abnormal meiotic pairing, leading to an increased susceptibility to unequal NAHR. In this situation, the inversions presumably act as catalyst and cause a predisposition to secondary rearrangement by switching the orientation of large, highly identical stretches of sequence on homologous chromosomes. This switching of orientation allows their subsequent misalignment during synapsis and hence facilitates illegitimate recombination (Sharp, Cheng, and Eichler 2006)

Therefore, these inversions have been associated with an increased susceptibility to rearrangements at these loci (Sharp, Cheng, and Eichler 2006). So far, there is growing evidence for several polymorphic inversions flanked by highly homologous segmental duplications. Parents that carry the inversions in heterozygosis confer a predisposition to further deletion of the inverted segment in subsequent generations. Most of these cases have been described as microdeletion syndromes in the offspring of inversion heterozygotes, such as Sotos syndrome, Angelman syndrome, W-Beuren syndrome and Koolen-de Vries syndrome (**Table 1**) (see Sharp, Cheng, and Eichler (2006) for a review).

Locus	Cytogenetic location	Population frequency	Size of inversion region	Associated predisposition
OR genes	4p16	12%	~6 Mb	t(4;8)(p16;p23) translocation
Sotos syndrome critical region	5q35	Unknown	2.2 Mb	Deletion of SoS critical region
Koolen-De Vries syndrome	17q21.31	20%*	900 Kb	Deletion of 424-kb encompassing 6 genes including *MAPT*
Williams-Beuren syndrome critical region	7q11.23	Unknown	1.6 Mb	Deletion of WBS critical region (and atypical WBS phenotype?)
OR genes	8p23	26%	4.7 Mb	inv dup(8p), +der(8)(pter-p23.1::p23.2-pter) and del(8)(p23.1;p23.2)
Angelman syndrome critical region	15q11-q13	9%	~4.5 Mb	Deletion of AS critical region
Proximal Yp	Yp11.2	33%	~4 Mb	*PRKX*/*PRKY* translocation (sex reversal)

Table 1 - Inversions predispose to disease - Summary of polymorphic inversions that predispose to further rearrangements that cause disease. Adapted in part from Sharp, Cheng, and Eichler (2006). *Frequency in Europeans.

1.4 Methods for inversion discovery and genotyping

1.4.1 Traditional methods

Inversions were first seen in the giant salivary gland chromosomes of larval flies and *Diptera* remains the group in which large inversions can be most easily detected. Chromosome staining techniques are able to visualize inversions in some other groups. However, the detection of inversions was traditionally limited to large-scale microscopically visible rearrangements via karyotype analysis using classical

G-banding techniques (de la Chapelle et al. 1974; O'Neill, Eldridge, and Metcalfe 2004; Wilson et al. 1970). Nowadays most experimental techniques such as fluorescence in situ hybridization (FISH), pulsed-field gel electrophoresis (PFGE), haplotype fusion-PCR remain laborious and target-based (Turner et al. 2006). Using these approaches, one can only test the presence of a predicted inversion in a specific genomic location. Nevertheless, they are still used for validating macroscopical SVs in some studies (Antonacci et al. 2009). Novel approaches based on SNP array data and next-generation sequencing (NGS) data have been recently introduced to identify or predict the location of SVs in general, including inversions at a genome-wide level. Some of these approaches are described in this section.

1.4.2 State-of-the-art genomic methods

The appearance of novel and powerful molecular biology techniques (e.g. high-throughput sequencing platforms, array-Comparative Genomic Hybridization (CGH) and SNP microarrays) has provided the means for studying genome structural variation in a high-throughput manner (see Alkan, Coe, and Eichler (2011) for a review). Nevertheless, not all types of SVs have benefited equally from the advent of these approaches. Inversion discovery and genotyping has lagged behind other kinds of SVs because of features particular to inversion: techniques particularly suited to CNVs discovery, such as array-CGH and microarrays cannot be used to detect inversions because they do not imply gain or loss of genetic material. Suitable techniques for inversion detection (**Figure 7**) are explained in the following subsections.

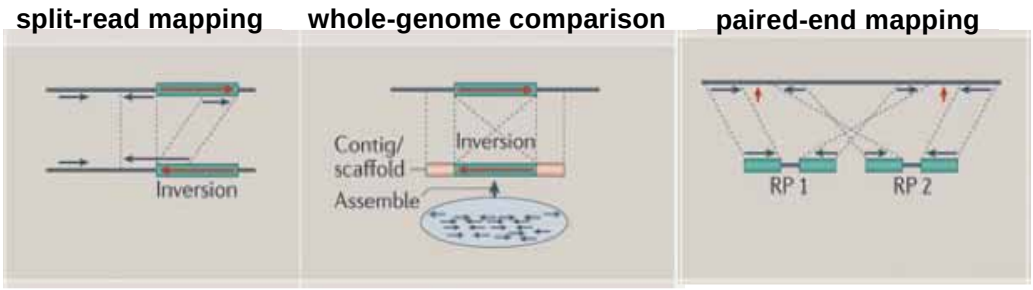


Figure 7 - Inversion detection techniques - Reads are represented as arrowed lines and the inversion region as a green box with a red arrow inside. Leftmost panel: identification of inversion BPs by split-read mapping. Reads from an inversion-carrier sample genome (bottom) are aligned to a reference genome (top). A single read targeting the boundaries of an inversion region produces 2 split reads that map to different genomic locations with opposite orientation. Middle panel: identification of inversion regions by whole-genome comparison. Reads from an inversion-carrier sample genome (bottom) are assembled and the resulting contig is aligned to a reference genome (top). Leftmost panel: identification of inversion BPs by paired-read mapping. Paired reads from an inversion-carrier sample genome (bottom) are aligned to a reference genome (top). Paired reads targeting the boundaries of an inversion region map to unexpected genomic locations with the same orientation. Adapted from Alkan, Coe and Eichler (2011).

1.4.2.1 Whole-genome sequence comparison

One possible approach for inversion detection is whole-genome sequence comparison, used recently by a few studies that compared the reference human genome to a *de novo* assembled sample genome (Levy et al. 2007; Li et al. 2010) and to the chimp genome (Feuk et al. 2005). The strategy has two phases. First, *de novo* assembly is carried out and the reads from the sample genome are concatenated to obtain large contigs/scaffolds. The second step is alignment, where the different SVs can be detected through the alignment of the contigs produced in the first step against the assembled sequences of the same genome region in other individual or the reference genome (**Figure 7**).

This strategy is theoretically the most comprehensive method for the study of genomic structural variation, because it allows the detection of all types of variation (SVs, small variants and SNPs), as well as the discovery of novel sequences. In addition, it gives the exact location of the variants, allowing to resolve the breakpoints at nucleotide resolution. However, despite all these advantages, this approach is expensive and complex, as it usually involves whole-genome genome sequencing coupled with *de novo* assembly. Therefore, it is not currently used for detecting SVs in a large number of samples.

1.4.2.2 Split-read mapping

This approach uses the profile of incompletely aligned reads to pinpoint the exact breakpoints of SV events. It is based on the mapping pattern of reads from a sample genome that span SV breakpoints, which will be mapped partially between both sides of the breakpoint in the reference genome (Ye et al. 2009). That means the read will be broken into two segments (**Figure 7**). Since a split-read signature indicates a breakpoint, a deletion in the sample genome will be associated to split reads mapping with an inner gap that represents the extra sequence in the reference genome. Insertions in sample genome will be associated with a set of reads that map partially to the reference genome, with just the left or right extreme aligned, depending on which breakpoint (left or right) in the sample genome is bridged by the split reads. In the case of inversions, the reads spanning the breakpoints in the sample genome will be associated with read mapping divided into two fragments in relative inverted orientation one to the other and spanning an inner gap, which in this case represents the inverted sequence in the sample genome. Split read strategy is able to detect breakpoints of SV with very high resolution, theoretically at nucleotide resolution, especially in unique regions. In addition, the identification of such a 'split-read' alignment signature complements the paired-end mapping (PEM) approach in widening the size space search towards the low end, as significantly smaller insertions and deletions can be discovered. However, in the case of inversions, split-read mapping cannot solve mapping ambiguity at inversion breakpoints when there is presence of IRs.

One of the first algorithms using split reads approaches to identify SVs is `Pindel` (Ye et al. 2009). Several other algorithms using the split reads approach exist nowadays, such as `Splitread` (Karakoc et al. 2011), which clusters split-read mappings based on the maximum parsimony method. In addition, there are algorithms that use evidence from both paired-ends and split reads to improve the definition of breakpoints, such as `DELLY` (Rausch et al. 2012) or `PRISM` (Jiang, Wang, and Brudno 2012). Finally, there are several other algorithms that use the split read approach applied to specific features, such as `TopHat` (Trapnell, Pachter, and Salzberg 2009), `Dissect` (Yorukoglu et al. 2012), or more recently `segemehl` (Hoffmann et al. 2014). These methods are specialized in deciphering transcriptome structure using RNA-Seq data by detecting splicing, *trans*-splicing and gene fusion events from single-end read data.

1.4.2.3 Paired-end mapping

PEM is one of the most commonly used methods as it has shown promising results in genome-wide detection of inversion rearrangements. This technique is able to assess the orientation of paired-end reads, therefore allowing the identification of discordant mapping to a reference genome (Korbel et al. 2007; Medvedev, Stanciu, and Brudno 2009; Tuzun et al. 2005). PEM consists of shearing a genome in fragments of a certain size, sequencing fragment ends, mapping them to a reference genome and interpreting the mapping signal to look for patterns indicative of SVs. SV-free fragments will map concordantly, with expected orientation (+−) and size, whereas SV regions will display discordant patterns. Indel fragments will map concordantly in orientation but discordantly in size. A translocation will be associated to a discordant mapping in location, where one read maps in another chromosome. Finally, fragments that bridge inversion breakpoints will map discordantly in orientation (++ or −−) and size (**Figure 8**).

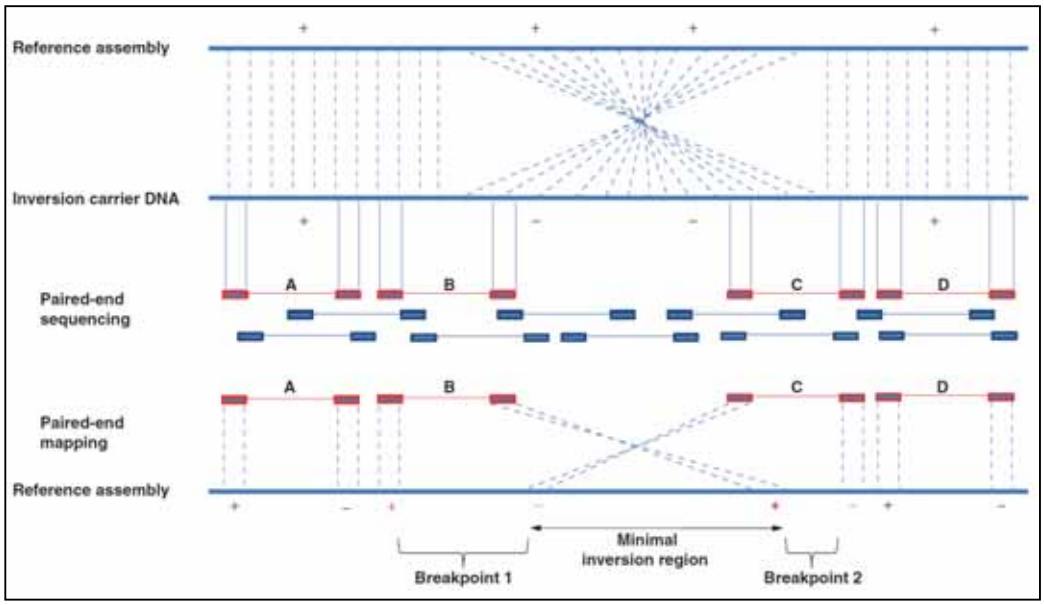


Figure 8 - PEM technique – The genome of an inversion carrier sample is aligned to a reference genome (top part of the figure). The sample genome is fragmented in pieces of a fixed length; the extremes of these genomic fragments are sequenced and the resultant paired ends (blue and red boxes linked) are mapped to the reference genome. If pairs map with expected distance (equivalent to the fragment size) and orientation signal (plus(+)/minus(−)), they are indicative of a region equivalent to the reference genome (see A,D end-pairs). Contrarily, if the end-pairs map ++ or −− (see B,C end-pairs), they are indicative of an inversion occurring in the target region as they span the inversion breakpoints) Taken from Feuk et al. (2010).

Following the advent of protocols based on PEM libraries generation and fragment end sequencing, many different algorithms based on different strategies have been developed for the prediction of SVs from the PEM discordant patterns.

For example, there are some algorithms that employ a "hard clustering" approach, using only the best location for each mapped paired-end read to find SVs, such as `PEMer` (Korbel et al. 2009) or `BreakDancer` (Chen et al. 2009). Alternatively, other SV detection algorithms use multiple mappings for each paired-end read and employ a soft clustering method through a combinatorial optimization framework and maximum parsimony or a heuristic approach that considers several possible positions for each read. An example of this type of algorithms is `VariationHunter` (Hormozdiari et al. 2009), but several others exist.

The approaches discussed are significantly different for each strategy of detection of SVs. Nevertheless, recently some integrative methods have been developed in which multiple signals are used in order to achieve improvements on the SVs discovery. The implementation of multi-approach algorithms that integrate the analysis of varied signals from read mapping stems from the need to improve the accuracy of SV discovery methods, because none of the single approaches provide comprehensive results. Most of these integrative algorithms combine paired-end reads and read-depth patterns. One of them is `GASVPro` (Sindi et al. 2012), which uses paired-end read patterns to find candidate SVs and then uses the read depth information as a posterior filtering. `DELLY` (Rausch et al. 2012) combines paired-end reads and split read approaches, using read pair signatures to detect candidate SVs and then refine as much as possible the breakpoints using split-read information.

A series of recent publications (Ahn et al. 2009; Kidd et al. 2008; Korbel et al. 2007; Pang et al. 2010; Tuzun et al. 2005; Wang et al. 2008) have successfully applied PEM-based methods to identify SVs in the human genome in a genome-wide manner. However, PEM-based techniques present limitations for inversion discovery and breakpoint refinement. For instance, inversion breakpoints are generally enriched in copies of duplicated segments of DNA (e.g. segmental duplications (SDs)), which greatly limits the ability to unambiguously map breakpoint regions (Feuk et al. 2005). Moreover, the repetitive nature of the genome causes high rates of false positives for inversion predictions (Lucas-Lledó and Cáceres 2013; Onishi-Seebacher and Korbel 2011). PEM-based methods try to overcome this issue by implementing parameters that favor concordant mappings. As a consequence, many reads sequenced across inversion breakpoints are missmapped concordantly, and the

power of PEM methods to detect them is significantly reduced. Thus, this sometimes results in an increase of false negatives (Lucas-Lledó and Cáceres 2013). In addition, another source of false negatives is the fact that some methods make compulsory the detection of both breakpoints to predict an inversion. Nonetheless, for inversions smaller than the insert size, chances are that only one (or none) of the breakpoints is bridged by the fragment sequenced and therefore the inversion will not be predicted. Another problem is that a good definition of an inversion should provide refined breakpoint regions, and yet this is not the case for inversions predicted out of a set of discordant mappings targeting only one of the two breakpoints.

As reported by Lucas-Lledó and Cáceres (2013), building libraries with longer reads and longer templates is more important for increasing inversion detectability than increasing the coverage. Ideally, for a paired-end to target an inversion breakpoint with no ambiguity, it should bridge the repetitive sequence at the breakpoint. However, the insert size of most used next generation sequencing technologies is yet significantly under 3 kb. This figure is much less that the average size of a segmental duplication in the human genome or some common repetitive elements such as retrotransposons, that range from ~100 bp to over 5 kb in size, or most frequently LINEs, that are about 6,500 bp in length. Thus, a big fraction of the polymorphic inversions in the human genome remain difficult to discover by PEM using this data. Also, there exist problems with the scalability of the method: PEM may not efficiently apply to the large number of samples that are needed to characterize inversions in a population and detect their association with diseases.

1.4.2.4 Inversion detection from genotype data

A cost-efficient way of detecting and characterizing inversions may be based on the widely available high-density genotype data of SNPs. In the simplest case, an inversion is tagged by a SNP in perfect linkage disequilibrium (LD), so inversion genotype can be directly inferred from SNP genotype. In other cases, although not in perfect disequilibrium, there still are SNPs tightly linked to an inverted region that can serve as a surrogate marker for the inversion. For instance, Bosch et al. (2009) have identified 16 SNPs in strong LD with an inversion located in 8p23.1 chromosomal region and have used these SNPs to indirectly infer the inversion genotypes in a small number of samples.

Even in absence of strong SNP-inversion linkage, the effects of inversion variations may still manifest in the statistical properties of the SNP genotypes. In the inverted orientation, the physical ordering of the SNPs is different, and one often observes an unusually higher level of long-range linkage disequilibrium (LD) in addition to an unusually lower level of short-range LD across the inversion breakpoints (Pritchard and Przeworski 2001). Taking advantage of this feature, several inversion-detection methods have been developed that use the LD signature in different manners. For example, Bansal, Bashir, and Bafna (2007) developed a statistical method to detect large polymorphic chromosomal segments (> 200 Kb) that are inverted in most of the chromosomes in a population with respect to the human reference sequence and applied it to HapMap data. This method was able to predict 176 inversion regions with minimum frequence of 0.25 (for inversions with lower frequences, the method lacks enough statistical power hence fails to identify them). This list of candidate inversions overlapped with several previously known inversion polymorphisms. However, the method suffers from a high false positive rate: some false positive predicted inversions correspond to regions of high LD due to low recombination or recent selective sweeps. More recently, Sindi and Raphael (2010) applied a probabilistic model to identify inversion polymorphisms, using differences in haplotype block structure. In opposition to Bansal et al. (2007) method, their method was able to detect inversions and predict inversion frequencies that are the minor allele in the population. The authors generated a set of 355 putative inversion polymorphisms using SNP data from 4 HapMap populations: Americans with Caucasian ancestry (CEU), Yorubans (YRI) and Chinese and Japanese pooled together (CHB+JPT). The list overlapped with several already validated inversion polymorphisms. A generalization of Sindi et al. (2010) method has been recently implemented by Cáceres et al. (2012) in an R package called `inveRsion`. Finally, methods based on multidimensional scaling (MDS) (Salm et al. 2012) and principal components analysis (PCA) (Ma and Amos 2012) have been implemented for inversion detection. The former method was developed for genotyping the 8p23.1 inversion using unphased SNP data (for detailed information of this variant, see section I3.3.1). The latter method was used to infer the inversion genotypes of inversion polymorphisms at 8p23.1 and 17q21.31 on HapMap III data, and outcomes are highly in agreement with literature results. In addition, the method was applied using HapMap data to carry out a preliminary genome-wide inversion scan and resulted in 2040 candidate inversions, 169 of which overlapped with previously reported inversions.

However, there are limitations that need to be taken into account when predicting inversion rearrangements from SNP-haplotype data. The proposed approaches rely on the assumption that SNP haplotypes can be used as a proxy for the inversion status and that strong LD is expected in regions harboring inversions. As a consequence, only ancient, unique inversions that have accumulated divergent mutations are expected to be captured, whereas novel, recurrent inversions with no remarkable correlation with SNP-based haplotypes will be neglected. In summary, making use of high density SNP data to identify inversion regions is a promising approach but has some limitations. For instance, in those cases where an inversion has arisen independently on at least 2 distinct haplotype backgrounds (i.e. recurrent inversions), genotyping methods based on SNP data suffer from a false positive and negative rate.

Ultimately, validation studies are needed to measure the error rates of SNP- or NGS- inferred inversion rearrangements. A recent review (Alkan, Sajjadian, and Eichler 2011) explored the main limitations of the current approaches to discovering SVs, highlighting the importance of designing algorithms that incorporate multiple methodologies in an integrative fashion to improve power, robustness, sensitivity and specificity.

1.4.2.5 Experimental genotyping methods

In order to understand the effect of inversion on human disease, complex traits and evolution, one of the preliminary steps required is to infer the inverted allelic composition of a sample of the study. The distinction between discovery and genotyping is important. Once a variant has been detected, validated and characterized at the sequence level (discovery), a different suite of methods may be applied to infer genotypes with relaxed thresholds. However, genotyping SVs and inversions in particular, in a high-throughput manner is still a challenge. In practice, this goal is not completely achieved because inversions tend to reside within repetitive DNA and that makes their characterization more difficult.

FISH-based techniques are an option for inversion genotyping and some studies use this approach both for validating and genotyping at low or medium-throughput level (Antonacci et al. 2009) (**Figure 9**). Still, this approach is very

laborious and only useful for relatively long inversions that can be identified at an optical-microscopy level.

PCR amplification is a better option for high-throughput analysis and different PCR-based techniques have been used to validate inversions. However, all of them have some disadvantages. For instance, regular or long-range PCR are restricted by the size of the fragments to amplify and do not work efficiently for long fragments (~5 kb for PCR, ~10 kb for long-range PCR). Therefore, inversions generated by simple breaks or mediated by small IRs at their breakpoints cannot be genotyped by these techniques. Haplotype-fusion PCR is a very promising technique to study inversions caused by duplicated sequences of almost any kind (Turner, Tyler-Smith, and Hurles 2008; Turner et al. 2006), although it has not yet been optimized for inversion genotyping. Another alternative is inverse PCR (iPCR) (Ochman, Gerber, and Hartl 1988). This protocol is initiated by restriction-enzyme digesting of DNA followed by a ligation step to produce circular molecules of DNA; then inversion genotype is inferred by amplification across a self-ligated site. iPCR has been applied to genotype inversions mediated by 9.5 kb SDs causing hemophilia A in multiple patients and in prenatal diagnosis. Recently, a high-throughput iPCR protocol has been developed (Aguado et al. 2014) that allows to genotype a wide-variety of inversions mediated by IRs in a large number of individuals in just one day. Aguado et al. (2014) used the protocol to analyze 17 human inversions ranging from ~5 kb to 226 kb and mediated by IR sequences of 1.6–24 kb.

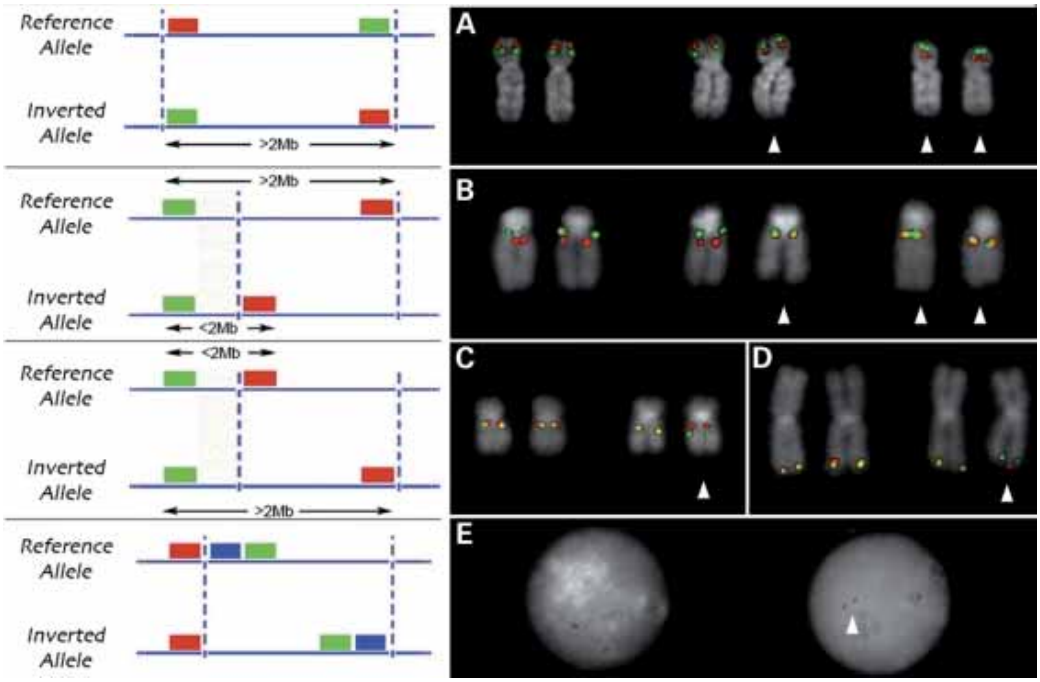


Figure 9 - FISH genotyping of inversion polymorphisms - FISH genotyping of inversion polymorphisms. **A**) Metaphase FISH validation of the 8p23.1 inversion using two probes located inside of the inversion. Metaphase FISH-based assay to resolve inversions <2 Mb using one probe located inside and one outside the inverted region is shown for inversions 15q13.3 (**B**), 17q12 (**C**) and 3q29 (**D**). **E**) Interphase triple colour FISH validation was used for the inversion 15q24. Arrows indicate inverted chromosomes. Taken from Antonacci et al. (2009).

Some studies claim that is not necessary to directly genotype SVs, inversions included, if the genotyping process is conceived only as an inexorable step to gain further knowledge about population statistics of the variant. In addition, SV genotype imputation is usually not an option to replace direct genotyping, as there exists circularity and biases associated with the approach. Instead, Lucas-Lledó et al. (2014) claim that an alternative to genotype imputation is to study the population-level structure of the variation with new methods that take genotype uncertainty into account in the analysis. For that, they present `svgem` (Lucas-Lledó et al. 2014) an expectation-maximization implementation to estimate allele and genotype frequencies, calculate genotype posterior probabilities and test for Hardy-Weinberg equilibrium and for population differences from the number of times the alleles are observed in each individual.

The human genome

One decade since the completion of "The Human Genome Project" (HGP), biological sciences are facing with great impetus one of their main challenges, the deciphering of genome functions and understanding the complex way in which the genome sequences are translated into a big variety of phenotypic characteristics of individuals. Furthermore, biomedical interest has intensified the thorough investigation of individual genome variation. A wide variety of large-scale projects have been already launched to investigate the human genome from diverse perspectives. In this section we will provide a general overview of the objectives of a few of these projects and their results.

2.1 Genomic variation in the human genome

An important scientific aim after the completion of the human genome is the understanding of the nature and patterns of variation within the human species, including both common and rare variants and its use as markers in linkage and association analysis. Initially, the focus of variation discovery was targeted SNPs, which are changes in one base between sequences, but later the focus shifted to SVs. SVs can be defined as a wide variety of genomic rearrangements of different sizes but involving more than 50-100 bp nucleotide and constitute bigger and more complex alterations in the architecture of eukaryotic genomes than SNPs. SVs can be either balanced or unbalanced depending on gain or loss of genetic material: unbalanced SVs are called copy number variants (CNVs) and include insertions, deletions, and duplications whereas balanced SVs include chromosomal inversions and translocations (**Figure 10**).

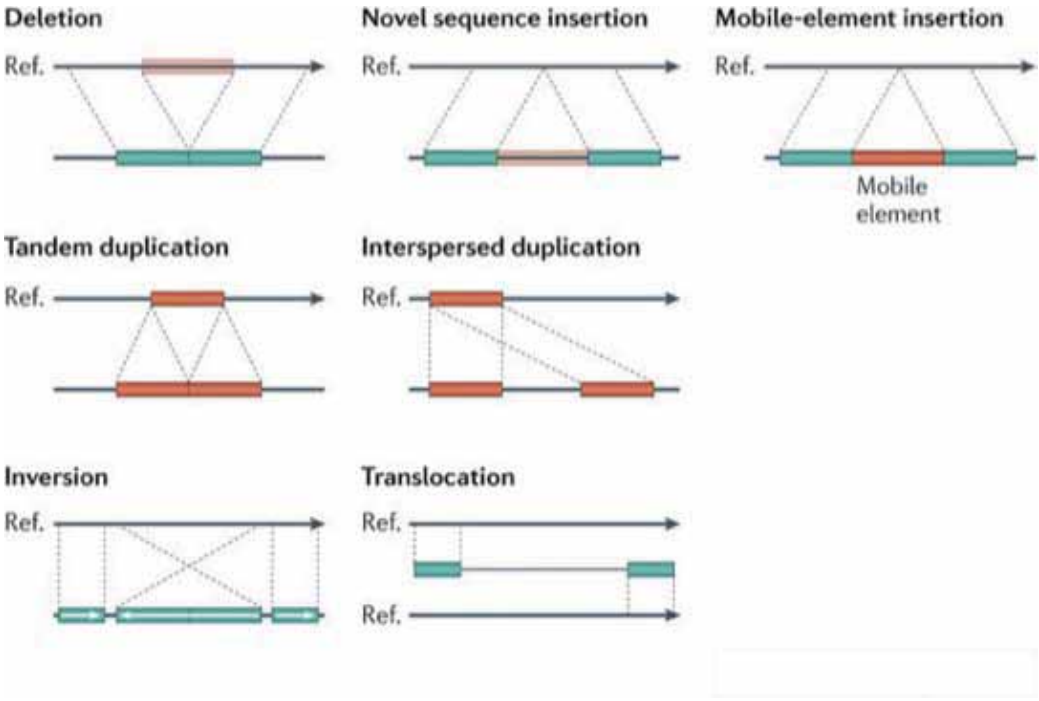


Figure 10 - Classes of structural variation – Schematic representation of deletions, novel sequence insertions, mobile-element insertions, tandem and interspersed segmental Duplications, inversions and translocations in a sample genome compared to a reference genome (rightwards arrow line). Taken from Alkan, Coe and Eichler (2011).

Over the last years, in the human genomic variation field, attention has shifted from SNPs towards SVs, because exceeding previous expectations, SVs have been found to constitute the major source of variation among human individuals: in the human genome, more base pairs are altered as a result of SVs — including CNVs — than as a result of SNPs (Conrad et al. 2010; Feuk, Carson, and Scherer 2006; Iafrate et al. 2004; Kidd et al. 2008; Levy et al. 2007; Pang et al. 2010; Redon et al. 2006; Sebat et al. 2004; Tuzun et al. 2005). Recently, there have been a few studies attempting to reveal the full spectrum of genetic variation in the human genome, either by studying a single genome (Levy et al. 2007; Pang et al. 2010) or multiple genomes (Kidd et al. 2008, 2010; Mills et al. 2011). Pang et al. (2010) employed whole-genome sequencing (split-read and PEM alignment) and microarray techniques and compiled results of Levy et al. (2007) to produce one of the most complete, up-to-date genetic variation catalogue of the human genome. According to Pang et al. (2010) , genome differs from the consensus reference sequence by

approximately 1.58%: 1.28% when considering indels/CNVs, 0.3% by inversions and only 0.1% by SNPs.

2.1.1 HapMap project

"The HapMap Project" (International HapMap Consortium 2003) was launched with the goal of developing high-density SNP genotyping technology to provide the scientific and medical community with detailed information about common SNPs and identify haplotype blocks for the analysis of human variation and their potential associations with human complex traits and diseases (Hinds 2005; The International HapMap Consortium 2005). From the estimated 15 million places along our genomes where one base can differ from one person or population to another, around three million (3.1×10^6) such locations have already been validated and characterized as SNPs in the second phase of the HapMap project. In addition, they have been charted using genotyping assays in 270 individuals from 4 geographically diverse human populations. The project has continued evolving by extending the reference panel on 7 additional populations, and in the third phase, 1.6 million common SNPs were genotyped in 1184 reference individuals from a total of 11 world-wide populations, and ten regions of 100 kb were sequenced in 692 of these individuals (Altshuler et al. 2010). This resulted in the characterization of population-specific differences among low-frequency variants, and the improvement of imputation accuracy, especially for variants with a minor allele frequency ($\leq$5%).

2.1.2 1000 Genomes Project

The 1000 Genomes Project (1000GP Consortium and others 2010) aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Specifically, the ultimate goal is to characterize over 95% of variants that are in genomic regions accessible to current high-throughput sequencing technologies and have allele frequency of 1% or higher (the classical definition of polymorphism) in each of five major population groups (populations in or with ancestry from Europe, East Asia, South Asia, West Africa and the Americas).

In the pilot phase (1000GP Consortium and others 2010), the location, allele frequency and local haplotype structure of approximately 15 million SNPs, 1 million short insertions and deletions, and 20,000 SVs are described by analyzing 179 unrelated individuals from 5 different populations of different ancestry from Europe, Asia and Africa. Examining the data from a subsequent release, the Structural Variation Analysis Group of the 1000GP found 22,025 deletions and 6,000 additional SVs (including 501 tandem duplications, 5,371 mobile element insertions and 128 novel sequence insertions) by analyzing the same set of individuals as in the pilot phase but increasing the sequencing coverage (Mills et al. 2011). In a posterior release, The 1000GP Consortium (2012) analyzed the genomes of 1,092 individuals from 14 human populations, and provided a validated haplotype map of 38 million SNPs, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. The current release (July 2014), part of the phase III of the project, contains more than 79 million variant sites and includes bi-allelic SNPs but also small indels, complex short substitutions and other SV classes. It is based on data from 2,535 individuals from 26 different populations around the world.

However, in all the 1000GP published studies till date, the 1000GP Consortium has knowingly overlooked inversions as they state the following:

> "… regions of low sequence complexity, satellite regions, large repeats and many large-scale structural variants, including copy-number polymorphisms, segmental duplications and inversions (which constitute most of the "inaccessible genome"), continue to present a major challenge for short-read technologies, so the SV strategies used focus primarily in the detection of other variants, specifically deletions."
>
> (McVean et al. 2012)

Therefore, the 1000GP results are not representative of the whole spectrum of human genomic variation, including inversions.

2.1.3 HGSV project

The Human Genome Structural Variation Project (HGSV) (Eichler et al. 2007) aims at the discovery of SVs through development of clone resources,

sequence resolution and accurate typing of variants in individuals of African, European or Asian ancestry. This project has employed a clone-based method to systematically identify and sequence SVs genome-wide. In addition, HGSV has generated an integrated database of structural variation polymorphisms ascertained by experimental and computational analyses. This database includes large-scale structural variation, copy number polymorphisms and intermediate-sized structural variation mostly determined by fosmid paired-end sequence analysis (Kidd et al. 2008, 2010; Tuzun et al. 2005). The data are represented in the UCSC Human Genome Browser and related with SNPs by using the same DNA samples used by the HapMap Project and 1000GP.

2.1.4 Databases and repositories of structural variation

The increasing number of inversions that are being predicted are currently stored together with the other SVs in different databases that provide stable and traceable identifiers for their analysis. Many of these projects have been developed to store SVs that are linked to different phenotypes and related with diseases, and because of this, inversions are little represented. However, the ubiquity of SVs on human genomes has fostered some other projects that support public access to much broader information on SVs that generally are not known to cause diseases but are very useful for biomedical studies. Some examples of polymorphic SVs databases are described in this section: `DGV` (Iafrate et al. 2004) and `dbVar` (see URLs). Although both resources provide archival, data accessioning and distribution services for all types of SVs in all species, compiled SVs data is derived basically (but not only, as said) from humans, such as control and case populations, tumor samples as well as three large curated studies derived from multiple sources. Finally, `InvFEST` (Martinez-Fundichely et al. 2014), a warehouse specifically devoted to store information related to human polymorphic inversions is also described in this section.

2.1.4.1 dbVar

The `dbVar` database (see URLs) stores all types of SVs and accepts data from all species, including clinical data of human samples of healthy controls and disease patients. It also identifies variant prediction artifacts and provides some

curation through cross-referencing of its data with information from the Genome Reference Consortium (GRC). dbVar is integrated with Entrez and other NCBI resources.

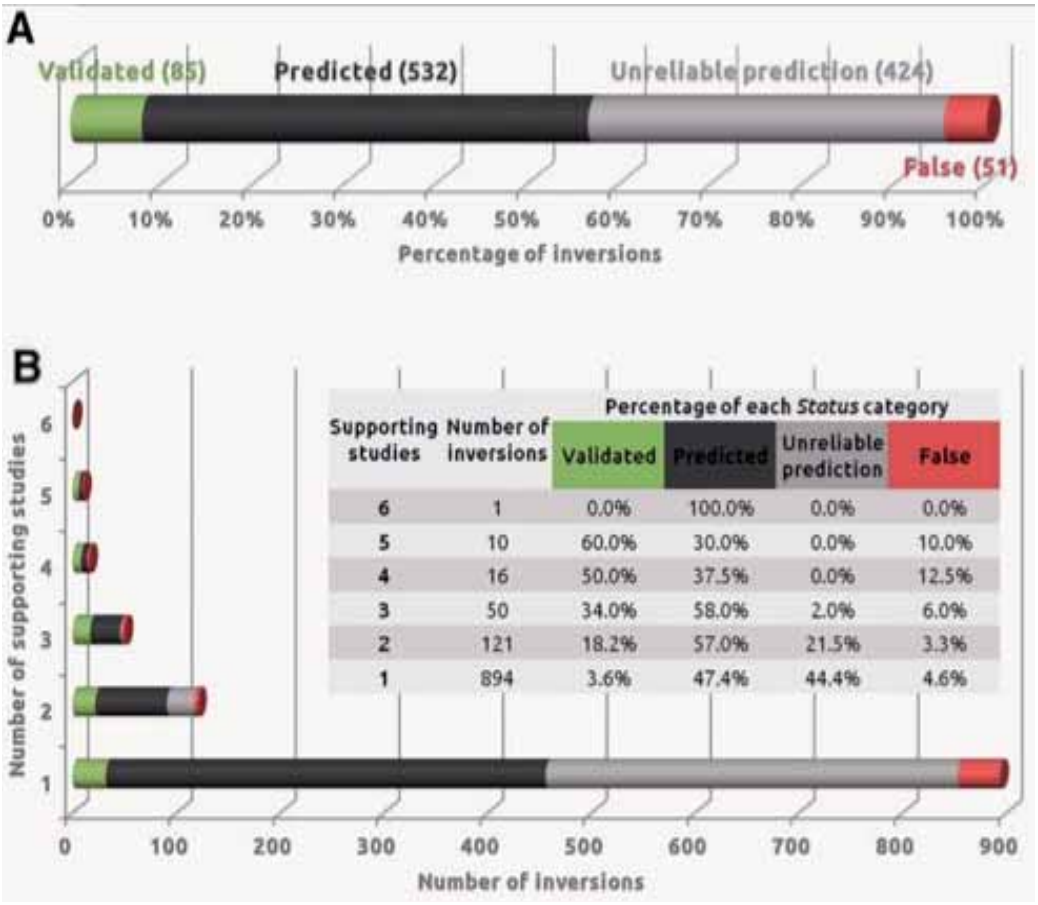
2.1.4.2 DGVa and DGV

DGVa has been designed to facilitate the curatorial work of the major database project focused on human structural variation, the Database of Genomic Variants (DGV) (Iafrate et al. 2004). The main goal of this database project is to provide a useful catalogue of curated data, and to facilitate the interpretation of SVs within the studies aiming to correlate genomic variation with phenotypic data. DGV has served a very important role collecting and analyzing structural variation data. Currently the database stores 55 studies that predict all the variety of SVs, in which there are 2,304,349 predicted events of CNV and 3,380 inversion events (Iafrate et al. 2004; MacDonald et al. 2014). The predictions with similar boundaries across the sample set are merged to form a representative variant that summarizes the common variant found in each study. At this merge level, the number of CNVs gets down to 109,863 and to 238 for inversions events (see section I3.1 for further information).

2.1.4.3 InvFEST

InvFEST (Martinez-Fundichely et al. 2014) is a database focused on inversions in the human genome. The aim of InvFEST is to integrate multiple sources of information to generate a complete catalogue of non-redundant polymorphic inversions in human population; providing the means for subsequent inversion validation and genotyping studies to gain insights about the functional and evolutionary consequences of inversions. This database is part of a larger project to characterize all human polymorphic inversions (see section I3.5). InvFEST inversions are classified according to their reliability through internal processes and exhaustive manual annotation. In addition, information about frequency, mechanisms of formation, functional associations, and breakpoint definitions is provided to the user to get a global picture of each inversion. This data integration and curation effort for inversions is not well represented in other SV databases so far, and, therefore, InvFEST is a useful complement and provides additional value to the DGVa (Iafrate et al. 2004; MacDonald et al. 2014). At the moment of publication

(January 2014) the database reported 1,092 candidate inversions, of which 85 had been validated experimentally. However, if false and unreliable predictions are excluded, the total number of inversions is reduced almost by half, to 617 (**Figure 11**).



Supporting studies	Number of inversions	Percentage of each *Status* category			
		Validated	Predicted	Unreliable prediction	False
6	1	0.0%	100.0%	0.0%	0.0%
5	10	60.0%	30.0%	0.0%	10.0%
4	16	50.0%	37.5%	0.0%	12.5%
3	50	34.0%	58.0%	2.0%	6.0%
2	121	18.2%	57.0%	21.5%	3.3%
1	894	3.6%	47.4%	44.4%	4.6%

Figure 11 - Summary of the invFEST database content - A) Status of the 1092 InvFEST candidate inversions. Numbers in parentheses indicate number of inversions for each status category. **B**) Number of inversions supported by 1, 2, 3, 4, 5 or 6 different studies. Different status categories are shown in colours and its percentage is represented in the table. Taken from Martinez-Fundichely et al. (2014).

2.2 Function of the human genome

One of the main targets after the completion of the Human Genome Project was to find and annotate all functional elements in the human genome. With this

goal, a public research consortium was created which launched "The Encyclopedia of DNA Elements Project" (ENCODE) (ENCODE Project Consortium 2004). The release of the initial results of this project has provided a complete map of the identification and detailed annotation of a wide variety of functional elements in the human genome. The analysis began from a little percentage (1%) of the human genome sequence (ENCODE Project Consortium et al. 2007), but it has scaled up to the study of the entire genome (Dunham et al. 2012).

ENCODE has contributed to expand the catalogue and the knowledge of human genes, annotating 9640 long noncoding RNA loci and reporting 33,977 coding transcripts not represented in other gene annotation datasets such as UCSC genes and RefSeq (Harrow et al 2007). ENCODE results also suggest that gene regulation is far more complex than was previously believed, stating that ~80% of the genome is involved in a biochemical reaction process (e.g. transcription factor binding, RNA polymerase binding, transcription) and has therefore a function. However, this liberal definition of genome functionality has created controversy and is not accepted by a large part of the scientific community (Graur et al. 2013). In brief and despite the polemic, results and data from ENCODE project provide the means for the study of gene functionality, the complexity involved in the regulation of gene expression levels, as well as in the disease association studies that will certainly enable the discovery of potential drug targets and to develop personalized medicine in the future. In addition, other strategies to infer the functionality of genome have become popular in the last decade, specifically on determining the genetic determinants of gene expression (see next section).

2.2.1 Genetic determinants of gene expression: expression Quantitative Trait Loci (eQTLs)

Recently, in the field of transcriptomics, major technological advances, such as the appearance first of high-resolution microarrays and lately next-generation sequencing-based technologies to measure gene expression such as RNA-Seq (Wang, Gerstein, and Snyder 2009) have enabled a reasonably accurate quantification of transcriptomes of multiple samples. The characterization and large-scale genotyping of human genome variation, coupled with genome-wide gene expression measurement has opened a door for a large-scale analysis of genetic variants that may determine gene expression. The analysis of such variants in the

context of gene expression measured in cells or tissues has given place to a big field in human genetics studying expression quantitative trait loci (eQTLs).

An eQTL is a locus that explains a part of the genetic variance of a gene expression phenotype. In its most common form, eQTL analysis is carried out testing the direct association between genetic variation markers (either distal or proximal to the gene) with gene expression levels in a a large number of individuals. Therefore, eQTL mapping permits the identification of new functional loci without requiring any previous knowledge of regulatory regions (*cis* or *trans*). In the eQTL mapping system, regulatory variants are characterized as either *cis* or *trans*-acting and this reflects the predicted nature of interactions depending on how far (in terms of physical distance) the variants are from the gene they regulate. In this thesis, variants within 1 Mb on either side of a gene's transcription-start site (TSS) are called *cis*-acting while those at least 1 Mb downstream or upstream of the TSS or on a different chromosome are considered *trans*-acting.

It is pertinent to mention here that a framework of knowledge that helps the study of genomes in humans has been built by a number of studies that have focused on eQTLs in model organisms such as yeast (Brem 2002; 2005) or rodents.

2.2.2 eQTLs in humans: studies, databases and repositories

According to eQTL studies carried out thus far, most regulatory control takes place locally, in the proximity of genes (Dixon et al. 2007; Göring et al. 2007; Schadt et al. 2008). For example, 831 genes were detected to have *cis* eQTLs in a study performed on 270 LCLs derived from HapMap 2 individuals genotyped for 2.2 million common SNPs (Stranger et al. 2007). As statistical power increases with the availability of larger sample sizes, it is expected that the number of genes detected to have eQTLs will also increase. In addition, correcting for latent confounding factors and use of transcriptome sequencing has allowed a significant increase in statistical power, providing thousands of eQTLs from a few hundred individuals (such as in the results made public by the Genetic European Variation in Health and Disease (GEUVADIS) project (Lappalainen et al. 2013). The study analyzed messenger RNA and microRNA from Epstein-Barr virus-transformed B-lymphoblastoid cell lines (LCLs) of 462 individuals from the 1000 Genomes Project and reported 8,329 genes with QTLs for different transcript traits. For 7,825 genes, a conventional eQTL was

reported (much more than in any other human eQTL study), 639 genes with transcript ratio QTLs (loci associated to different ratios between isoforms) and 60 autosomal miRNAs with an eQTL.

The process of sifting through the whole genome for potential regulatory effects is statistically challenging and hence limited by computational power. Thus, finding *trans* eQTLs has been less successful so far. The question whether the current enrichment of *cis* versus *trans* eQTLs reflects biological reality as opposed to being a result of technical difficulties to detect eQTLs in *trans* is still under debate (Wittkopp, Haerum, and Clark 2008; Wray 2007). It is pertinent to note that however, recent studies have shown that when a reasonable sample size is tested, thousands of replicated tissue specific *trans* eQTLs can be found (Fairfax et al. 2012; Franceschini et al. 2012; Grundberg et al. 2012).

Most human eQTL studies have been performed exclusively on blood-derived cells or cell lines. The ease of access of this cell type has made it very useful in understanding the genetics of gene expression and other cohort studies. However, because gene expression signatures are cell-type specific, the question whether regulatory control of expression is also cell-type-dependent acquires importance. Estimates vary depending on the eQTL methods used and the tissues being compared, but usually, *cis* regulation has extensively been found to be tissue-specific. In a comparison of blood and adipose expression patterns in two Icelandic cohorts, 50% of the *cis* eQTLs detected were shared (Emilsson et al. 2008). Contrarily, a comparison of LCL and cortical tissue regulatory overlap in a European population showed hardly any overlap. However, this difference is probably increased by the different microarrays used in the two experiments (Myers et al. 2007). In another study, Dimas et al. (2012) compared the regulatory landscape in LCLs, fibroblasts and primary T cells derived from the same set of 75 individuals from European ancestry. The authors reported that 69–80% of *cis* eQTLs are cell-type specific, providing an impetus to the study of multiple tissues to determine the full spectrum of regulatory variants. Recently, twin-based eQTL studies analyzing hundreds of individuals have signalled diminishing returns when sample size increases. That is, there is certainly significant tissue specificity, but less than previously estimated. Tissue-dependent effect size can be detected basically when the study is well-powered (Grundberg et al. 2012, 2013; Nica et al. 2011). Finally, a big consortium (GTEx Consortium 2013) joined efforts to discover tissue-specific eQTLs and launched The Genotype-Tissue Expression project (GTEx), which aims to provide a comprehensive atlas of gene expression and regulation across multiple

human tissues. The GTEx project is in the scale-up phase of donor collection and tissue analysis, towards an end goal of 900 donors and around 20,000 tissue samples.

Other studies are investigating context-specific genetic association with differential gene expression (Fairfax et al. 2014; Lee et al. 2014). Fairfax et al. (2014) exposed primary $CD14^+$ human monocytes from 432 European volunteers to the inflammatory proxies interferon-γ (IFN-γ) or differing durations (2 or 24 hours) of lipopolysaccharide (LPS), characterized genotypes and transcriptomes of the samples and performed eQTL mapping. Interestingly, they report hundreds of *cis* associations that reveal only after providing the innate immune stimuli, like the *IFNB1* case, which expression is associated to a SNP in the LPS-stimulated monocytes but not in the non-stimulated cells. Lee et al. (2014) performed eQTL mapping on dendritic cells (DCs) of a cohort of 534 individuals, and similarly to Fairfax et al. (2014), they also stimulated the cells with IFN-γ and LPS. Authors report 121 context-specific eQTLs, associated with variation in the induction of gene expression by one or more stimuli.

To make accessible the results from the large number of eQTL studies published recently, several online databases are available which report eQTL associations based on published datasets. These resources are described briefly below:

`The GTEx portal`, in the framework of the GTEx project, provides preliminary results of the GTEx project and allows users to lookup an eQTL by gene or SNP in all tissues at once and display results in a genome browser. The latest data release (June 2014), makes available data of 2921 RNA-Seq samples across 53 different tissues. GTEx resources have extensively used in this work to look for inversion-QTLs.

`Genevar` (Yang et al. 2010) is a platform of database and web services designed for data integration, analysis and visualization of SNP-gene associations in eQTL studies. The default public server at the Sanger Institute currently contains genetic variation and gene expression profiling data from 5 tissues and cell lines (LCL, fibroblast, T-cell, skin, adipose) from in total 1657 individuals from 3 different sources: HapMap, MuTHER and Geneva GenCord projects (Dimas et al. 2012; Grundberg et al. 2012; Nica et al. 2010; Stranger et al. 2012). Besides eQTL data, Genevar provides a genetic variation and DNA methylation profiling dataset

from adipose tissue collected from 856 healthy female twins of the MuTHER resource (Grundberg et al. 2013).

`The eQTL browser` (see URLs), as part of the tool suite of eQTL resources at the Pritchard lab (see URLs) is a web browser that aims to provide eQTL data from recent studies in multiple tissues (Bell et al. 2011; Degner et al. 2012; Pickrell et al. 2010; Pique-Regi et al. 2011; Veyrieras et al. 2008).

`seeQTL` (Xia et al. 2012) is a comprehensive and versatile eQTL database that includes various eQTL studies and a meta-analysis of HapMap eQTL data. The database presents eQTL association results in a convenient browser, using both segmented local-association plots and genome-wide Manhattan plots. A brief summary of each one of the studies with corresponding references can be found in **Table 2**.

Datasets	**N**	**# of Genes**	**# of cis-eQTL (q-val <0.1)**	**# of trans-eQTL (q-val <0.1)**	**Reference**
Stranger CEU	56	16,925	9,228	11,793	Stranger et al. 2007
Stranger YRI	57	16,925	6,497	11,166	
Stranger CHB_JPT	85	16,925	33,507	16,580	
Price_CEU/ Spielman_CEU	56	8,696	921	1,311	Price et al. 2008
Price YRI	56	8,696	794	3,077	
Choy_CEU	53	13,094	2,667	2,578	Choy et al. 2008
Choy YRI	51	13,094	628	1,704	
Choy CHB_JPT	66	13,094	2,229	3,799	
Spielman CHB_JPT	75	8,696	3,317	3,543	Spielman et al. 2007

Datasets	N	# of Genes	# of cis-eQTL (q-val <0.1)	# of trans-eQTL (q-val <0.1)	Reference
Montgomery1RNA-Seq CEU	55	20,861	1,899	6,339	Montgomery et al. 2010
Montgomery2 CEU	107	16,624	18,747	19,952	
PickrellRNA-Seq YRI	52	20,861	627	6,714	Pickrell et al. 2010
Myers Europeans brain	186	16,819	2,586	23,280	Myers et al. 2007
Zeller Europeans monoctye	1,490	37,808	73,363	6,913	Zeller et al. 2010
HapMap eQTL consensus*	N/A	20,881	51,924	16,699	All HapMap samples

Table 2 - seeQTL summary table of eQTLs in each dataset – Statisitcs of the different studies contained in seeQTL database. N: sample size. *Metanalysis of Stranger et al. 2007 dataset, pooling all populations.

Finally, databases providing QTL information in non-human species, although smaller in number, also exist. An example is `QTL Archive`, a site that provides access to raw data from various QTL studies using rodent inbred line crosses (see URLs).

2.2.3 eQTLs and disease

On account of its relevance to differential disease risk among individuals, genome variability has been the focus of several studies in recent years. An understanding of the specific biological effect such variants have in the cell can help understand the biology of the organismal disease or phenotype and hence drives the need for interpretation of the effects of genome variants. Therefore, from the perspective of the disease, it is very important to document cell-type specific regulatory variation. Integrating the data of expression with results from GWAS is useful for discovering genes and pathways whose disruption may cause disease (Chen et al. 2008; Nica and Dermitzakis 2008; Nica et al. 2010). However, the pursuit of this objective is possible only when the tissue from which the expression

data is derived is relevant to the complex trait under interrogation (Nica and Dermitzakis 2008). For example, eQTLs discovered in LCLs have helped explain GWAS associations with two autoimmune inflammatory disorders namely, childhood asthma (Moffatt et al. 2007) and Crohn's disease (Libioulle et al. 2007). The blood and adipose cohorts analysed by Emilsson et al. (2008) were evaluated for various phenotypes as well, including obesity relevant traits. Significantly, 50% of the *cis* eQTL signals were reported as overlapping between the two cohorts, but a substantial correlation with obesity-related traits was only seen for gene expression measured in adipose tissue (Emilsson et al. 2008). These observations underline the importance of integrating data from a tissue of relevance when trying to interpret GWAS results. However, a caveat exists: the same regulatory region and variant can be linked to different genes in different tissues (Fairfax et al. 2014). This leads to the view that poor tissue sample size will give misleading biological interpretations about the disease risk-increase gene. Having said this, there is no clarity on the extent to which this phenomenon plays out across human tissues and hence it is not known what tissues could be substantially informative in large cohorts. For example, LCLs have been useful in finding candidate genes for associations with autism (Nishimura et al. 2007) or bipolar disorder (Iwamoto et al. 2004).

GWAS has enabled the discovery of disease-susceptibility variants at a great extent. Nevertheless, the understanding of how these loci contribute to disease lags behind, particularly for loci that map to non-coding genomic regions. In addition, it is also common to identify loci that seem to predispose to disease on genomic regions with high degree of LD, which makes it difficult to derive firm conclusions about causality (i.e. which is the causal variant and affected gene). For all these reasons, it is necessary to incorporate additional information (such as eQTL results) for interpreting GWAS results, as it is known that gene expression is an important mechanism underlying complex traits. That is, the integrative approach consists on looking for SNPs that are simultaneously associated with disease status and eQTLs: alleles found to be more frequent in cases than controls that in addition seem to cause an effect on the expression of a nearby gene. If the gene is by itself associated to the disease, then it is probable that causality can be ascertained. Consistently with this idea, several recent studies have integrated eQTL analyses with GWAS results and proposed candidate disease genes. For instance, Moffatt et al. (2007) identified a list of strongly correlated SNPs associated with childhood asthma in a large (200 kb) region of chromosome 17q23 that contained 19 genes. However, none of the genes had an evident disease role. Expression analysis on LCLs derived from the same disease-affected families showed that the most significant GWAS SNPs also

explained approximately 29.5% of the variance in transcript levels of *ORMDL3* (*ORM1*-like 3), one of those 19 genes, which is currently the top candidate for further functional studies.

In addition, Crohn's disease is another example of a disease with reported association signals better understood thanks to expression data. A recent GWAS reported multiple susceptibility loci mapping to a 1.25 Mb intergenic region on chromosome 5 (Barrett et al. 2008). On top of that, eQTL data showed that at least one of these loci act in *cis* as long-range regulators of *PTGER4* (prostaglandin E receptor 4). Interestinlgy, this gene locates 270 kb away from the associated region whose homologue has been implicated in phenotypes similar to Crohn's disease in the mouse (Libioulle et al. 2007). Similar other examples for height, systemic lupus erythematosus, type I diabetes or bipolar disorder support the use of eQTL data in contributing to a better interpretation of GWAS results (Burton et al. 2007; Gudbjartsson et al. 2008; Hakonarson et al. 2007; Hom et al. 2008).

However, there are also exceptions and more complex cases. Willer et al. (2009) found an association between body mass index (BMI) and a non-synonymous SNP in SH2B adaptor protein 1 (*SH2B1*) locus. Besides this association, the variant is reported to be an eQTL for expression of two other genes, *EIF3C* and *TUFM*. In mice, mutations on the S*H2B1* homologue lead to extreme obesity, which reinforces the hypothesis that the SNP is the actual functional variant. However, the SNP is in high LD with a causal regulatory variant of *EIF3C* and *TUFM* changes in expression. This is a frequent example of an overlap of GWAS and eQTL results that can lead to misleading interpretations and needs to be separated from informative, causal cases where both the disease-associated SNP and the eQTL tag the same functional variant. Regulatory variants seem to be very abundant (Stranger et al. 2007) and therefore such non-informative, coincidental overlaps are probable to happen by chance. Therefore, integrative methods pinpointing true causal regulatory effects are desirable (Nica and Dermitzakis 2008).

Finally, it could also be useful to integrate *cis* and *trans* eQTLs together with GWAS analysis to detect previously unknown determinants of disease. In the case of the *KLF14* gene, via eQTL analysis and by adding additional data from mice, it was inferred that *KLF14* is a likely signal from fat to other diabetes-related tissues to induce insulin resistance (Small et al. 2011). Nevertheless, it needs to be taken into account that phenotypes appear in a tissue-specific manner; therefore crossing

GWAS with eQTL data may be only informative if expression measurements from disease-relevant cell-types are compared.

2.2.4 eQTL methods

Mapping eQTLs can be done using standard QTL mapping methods that measure the association between genetic polymorphisms and gene expression. Surprisingly, the approach for eQTL discovery has not significantly change since the earliest eQTL mapping studies (Kendziorski and Wang 2006; Williams et al. 2007). In more recent RNA-Seq based eQTL studies (Montgomery et al. 2010; Pickrell et al. 2010), the authors keep applying simple single-QTL mapping methods to individual expression traits. The only considerable difference between traditional standard QTL mapping and current eQTL studies is the scale: that the latter can involve a million or more expression microtraits. Therefore high-throughput, automated and efficient approaches are needed.

Storey, Akey, and Kruglyak (2005) were pioneers in developing a method specifically designed for eQTL mapping, based on the calculation of *F*-statistics for each marker and trait. The approach identifies a primary locus that corresponds to the one with the maximal F-statistic. A secondary locus is identified as the one having maximal statistic in a second F-test conditional on the first, with permutations used to estimate the posterior probabilities and thresholds for locus-specific and joint linkage. More recently, Zou and Zeng (2009) propose a method based on multiple interval mapping that also combines features of Storey's approach with multiple interval mapping. Nowadays, one of the most popular methods for eQTL identification is `Matrix eQTL` (Shabalin 2012). `Matrix eQTL` can be used to create ANOVA and linear regression based models to test for SNP-gene expression associations. The models can be customized to include several covariates (population structure, gender, and clinical variables) and the software can test heteroscedastic models and models with correlated errors.

Certainly other powerful and effective methods exist (Kao, Zeng, and Teasdale 1999; Sen and Churchill 2001). These approaches are useful to accommodate complex genetic models, and they have proven successful in several studies. However, they are not straightforward to use but require “fine tuning”,

choice of thresholds and extensive parameterization instead, in order to identify single-trait interactions and resolve multiple linked QTLs.

Approaches that model all (Jia and Xu 2007; Kendziorski et al. 2006) or groups of traits (Chun and Keleş 2009) at once, compared to single-trait traits identification also exist. These strategies are based one building one model for the data, which enables the estimation of false discovery rate across markers and transcripts at the same time.

In summary, state-of-the-art QTL mapping methods are useful for the identification of complex genetic architecture, but the several pitfalls still exist. Non-trivial choices on the class of models to consider as well as significance thresholds compromise the easiness of the analysis and its application to studies of high-throughput phenotypes such as expression. However, methods designed specifically for eQTL mapping are popular, successful and are used in numerous studies (Shabalin 2012).

As is true in all QTL mapping studies, the final steps in defining DNA variants that cause variation in traits are usually difficult and require a second round of experimentation. This is especially the case for *trans* eQTLs that do not benefit from the strong prior probability that relevant variants are in the immediate vicinity of the parent gene. Statistical, graphical, and bioinformatic methods are used to evaluate positional candidate genes and entire systems of interactions.

Inversions in the human genome

3.1 Overview

Underrepresentation of inversions among human SVs can be observed in the Database of Genomic Variants (DGV) (Iafrate et al. 2004). Statistics from the last release (July 2013) published along the last version of the database (MacDonald et al. 2014) show that CNVs exceed inversions by far, and the latter constitute only 0.09 % of all variant cells and ~2% of the variant regions (**Figure 12**). So, according to the combined results of the collection of studies aiming to study structural variation that are compiled in the DGV database, roughly 1,149 inversions exist on

the human genome. However, as discussed earlier, the veracity of this figure is not clear since the false positive rate in inversion prediction is very high. According to InvFEST results, only a few hundreds of inversion predictions in the human genome are expected to be true (Martinez-Fundichely et al. 2014) (see section I2.1.4.3).

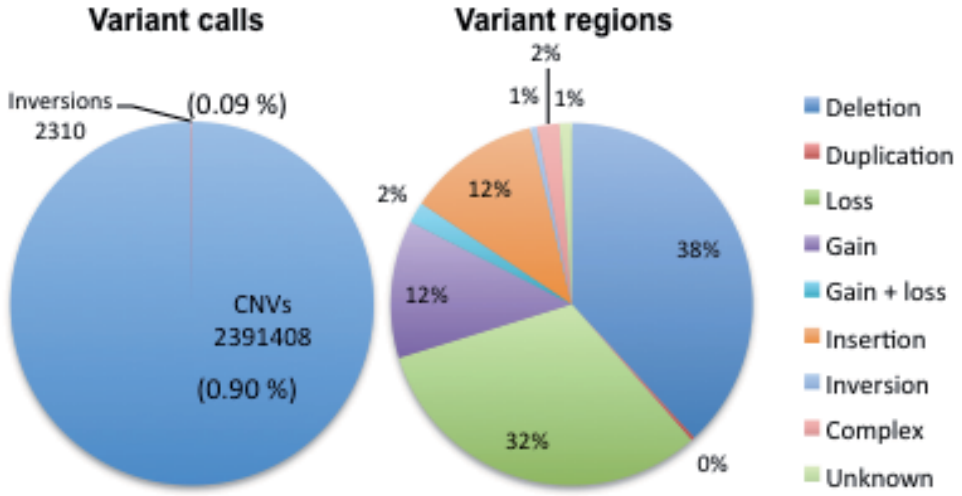


Figure 12 - DGV content - An overall summary of the number of the type of variants reported in the database (July 2013 update, mapped to GRCh37 assembly). Individual variant types are reported highlighting the distribution of SV content in the database. Data obtained from MacDonald et al. (2014).

3.2 Inversion origins in humans: mechanisms of formation and recurrence of inversion mutational events

Regarding the mechanism involved in inversion formation in humans, Aguado et al. (2014) and other studies (Kidd et al. 2008, 2010; Levy et al. 2007) have shown that inversions are generated by two main processes: either through breaks in relatively simple regions that are joined in opposite orientation by non-homologous mechanisms or by non-allelic homologous recombination (NAHR) between inverted repeats (IRs) (either repetitive elements or segmental duplications (SDs)).

The most exhaustive study of the mechanisms of formation of structural variation in the human genome up to date, by Pang et al. (2013), suggests that contrary to CNVs, where NH processes were associated with the majority of variants with a gain or loss of DNA, NAHR has a major role in inversion formation (**Figure 13**). According to their results, NAHR is responsible for only 0.2% of the gains and 0.3% losses, but for more than a half (54.7%) of the inversions. These NAHR-mediated inversions present a median distance of 1.9 kb between homologous copies, which are mainly large SDs or L1 elements. Moreover, the two largest NAHR inversions were L1 and SD associated (87,609 and 68,145 bp, respectively). The authors also found that there is a mild correlation (Spearman's correlation coefficient ρ = 0.34) between inversion size and size of the flanking homologous sequence. Only a few percent of inversions (4.3%) present 1 to 10 bp of inserted sequence at breakpoints, probably formed as a consequence of imperfect NHEJ repairs and 6 times more inversions (23.9%) showed 1–20 bp of flanking homology, probably formed by NH or microhomology based mechanisms.

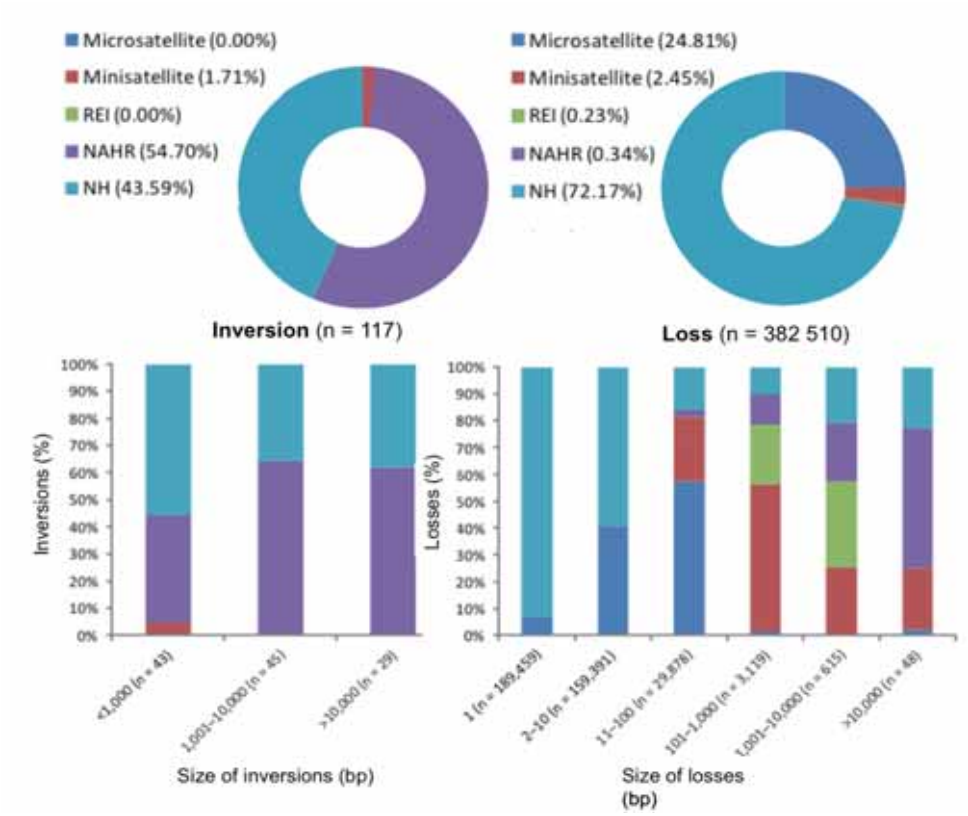


Figure 13 - Mechanisms of formation of inversions – Top panel shows the relative proportion of mechanism of formation for 117 inversions (left) and 382,510 losses (right). Bottom panel shows the relative proportion of mechanism divided by variant size, for inversions (left) and losses (right). Taken from Pang et al. (2013).

Finally, 15.4% of the inversions were simple clean break junctions that had neither additional sequence nor microhomology. However, this figure may be on the lower end as missannotation of one to three base pairs of microhomology can occur as a result of random chance, or could be false positives due to sequencing, alignment or assembly error.

Altogether, the results of the study conducted by Pang et al. (2013) suggest that inversions in the human genome are predominantly mediated by IRs through homology-based mechanisms. This indicates that inversions do not occur randomly in the genome but often between IRs and therefore they can be prone to recurrence events. In fact, several regions of chromosome breakage reiteratively used throughout evolution (i.e. independent breaks occurring at the same chromosomal sites) have been found in mammals. For instance, Murphy et al. (2005) analyzed the

genome organization of 8 mammalian species to identify patterns of chromosome evolution. Using homologous synteny blocks (HSBs) they identified several regions recurrent for rearrangements. Interestingly, the authors also observed that the majority of primate-specific breaks involve inversions that originated via NAHR between duplicated HSBs. Further support for this hypothesis was later provided by Cáceres et al. (2007), who identified an additional example of long-term breakpoint reuse during evolution of mammalian species in a genomic segment containing an X-chromosome polymorphic inversion on the human genome. The authors compared genomic sequences of 28 placental mammals and suggested that at least 10 independent recurrent events occurred and contributed to the present-day genomic structures observed in different species. In addition, recurrent events have also been detected within multiple primate lineages for the 17q21.31 region (Zody et al. 2008). Finally, results from the work of Aguado et al. (2014), based on the study of 17 polymorphic inversions in 68 HapMap individuals of European ancestry, also support a high degree of inversion recurrence during human evolution. However, despite this evidence, further work is needed in order to determine the distribution and rate of these supposedly non-randomly distributed break sites, as studies analyzing in depth the population genetics of inversions are scarce in the literature.

3.3 Human polymorphic inversions: well-studied cases

In humans, several inversion variants of different sizes segregate in populations (Feuk 2010; Hoffmann and Rieseberg 2008; Martinez-Fundichely et al. 2014; Pang et al. 2013). Although the vast majority falls within the 10 to 100 kb size interval, there are a few inversion polymorphisms bigger than 1Mb (Feuk 2010). However, despite the great amount of genetic material contained in these large variants, the chances of directly altering genic sequences are not proportional to the inversion length, as the impact of an inversion is primarily related with the location of its breakpoints. Therefore, if no gene is disrupted, even large inversions may be neutral and consequently spread within and between populations through stochastic processes. However, the larger the inversion is, the higher the probability that it will recombine and be selected against, due to unbalanced gametes produced by recombination.

Albeit in small samples, some indirect studies focusing on human diseases (Antonacci et al. 2009) have shown some examples of inversions well characterized.

In addition, a few studies assessed the frequency of a subset of inversions at a population level (Pang et al. 2013). However, only a few inversions have been extensively characterized at the population level. One example is the 8p23.1 inversion, spanning 4.5 Mb and considered the largest polymorphic inversion known in the human genome (Salm et al. 2012). Another well-studied case is the 17q21.31 inversion, that spans 900 Kb and attains relatively high frequencies in several European populations (Stefansson et al. 2005; Zody et al. 2008). Both inversions are extensively described in this section.

3.3.1 8p23.1 Inversion

The inversion located in 8p23.1 region (8p23.1-inv) was discovered more than 10 years ago (Giglio et al. 2001). In additional to the original one, more recent studies (Antonacci et al. 2009; Bosch et al. 2009) have reported that this variant presents a very complex, repetitive genomic architecture, mainly due to a couple of big blocks of segmental duplications (SDs) embedded in the inverted region.

An important feature of the 8p23.1 inversion is the considerable number of genes encompassed. The region contains at least 50 genes, among which there is the *BLK* (B lymphocyte kinase) gene that has been associated with systemic lupus erythematosus (SLE), rheumatoid arthritis (RA) and other autoimmune diseases (Simpfendorfer et al. 2012). Interestingly, it has been suggested that the risk alleles are specific to the non-inverted configuration (Salm et al. 2012). 8p23.1-inv is considered a neutral polymorphism (Salm et al. 2012). However, it indirectly confers susceptibility to disease due to subsequent rearrangements via NAHR, mediated by the presence of SD highly identical structures. Specifically, 8p23.1-inv is associated to syndromic phenotypes (e.g. microdeletion syndromes) in the offspring of heterozygous mothers. However, the exact molecular mechanisms leading to disease phenotypes remain to be elucidated.

In order to characterize its worldwide distribution, Salm et al. (2012) have recently applied an innovative approach to human SNP-genotype data. The authors have designed an algorithm based on multidimensional scaling (MDS) called PFIDO (Phase-Free Inversion Detection Operator). With this software, they efficiently categorized ~2000 individuals from 56 populations by inversion status, some of which were validated by FISH. According to Salm et al. (2012) results, this inversion

polymorphism displays a worldwide clinal distribution with frequencies ~25-79% (min. freq. in a “Manchu” sample of North-East Asia and max. freq. in a Mozabite sample of North Africa) consistent with demographic models of early human expansions out of Africa.

Salm et al. (2012) reasoned that the 8p23.1 inverted allele seems to have evolved either under very weak selective pressure or neutrally in humans. Furthermore, given the correlation between the inversion status and the genetic substructure, they suggested that recurrent events were infrequent across this region in the Homo lineage. However, non-recurrent inversion features such as presence of tag-SNPs were not found for 8p23.1, as no single SNP was in perfect LD with the inversion status. Therefore, 8p23.1-inv may not act as an absolute recombination barrier and gene flow may have occurred throughout its evolution. This phenomenon can occur in big inversions and it involves double crossover events. In fact, in a recent article, Alves et al. (2014) reported that, in agreement with Salm et al. (2012) results, the accumulation of genetic differentiation between the two inversion haplotypes positively correlates with the variation in recombination profiles. The recombination dissimilarity between inversion haplotypes is consistent across all populations analyzed and cannot be explained only by the effects of geographic structure. This observation suggests that both inversion conformations have evolved independently throughout an extended period of time, despite being subjected to the same demographic history. Nevertheless, the study identified a short segment (350 kb, <10% of the whole inversion) in the central region of the inversion where the genetic divergence between the two structural haplotypes is lower, suggesting possible gene flow events.

3.3.2 17q21.31 Inversion

Another relatively common inversion polymorphism that became the focus of intense research in the last years is located at 17q21.31. In contrast to the 8p23.1 inversion, early studies suggested that the 900 kb inversion polymorphism is undergoing selection in Europeans (Stefansson et al. 2005). The authors interrogated more than ~30,000 Icelandic individuals for a tag-SNP linked to the inversion, and observed that females carrying either one or two copies of the inversion had more children. Therefore, the authors concluded that this trait was being positively selected and acting on 17q21.31-inv allele frequencies.

More recently, Zody et al. (2008) analyzed the 17q21.31-inv evolutionary history in non-human primates. The authors observed that this genomic region suffered a serial of rearrangements throughout primate evolution that contributed to shape a complex, duplicated structural conformation. One or more rearrangement events caused the appearance of directly-oriented segmental duplications (SDs) blocks in the human H2 haplotype (inversion-associated haplotype). These in-face duplicons can mediate non-allelic homologous recombination (NAHR) events that ultimately increase the appearance of microdeletion and microduplication events, often associated with disease (Gu, Zhang, and Lupski 2008; Zody et al. 2008) (see section I1.3.3). On this basis, Zody et al. (2008) proposed that this detrimental feature particular to the ancestral H2 haplotype increased the frequency of the H1 conformation in humans. This hypothesis is in disagreement with the high frequency of the H2 haplotype in some European populations (between 5 and 35%). However, the H2-haplotype high frequency could also be explained by founder effects during the Out-of-Africa human colonization of the European continent.

Donnelly et al. (2010) came also with similar interpretations were subsequently after analyzing a more detailed global distribution of 17q21.31-inv haplotypes. In this study, the authors used not only SNPs but also short tandem repeats (STRs) polymorphisms to genotype the inversion. In opposition to the Out-of-Africa founder effect hypothesis, the authors concluded that the Neolithic transition shaped the present-day distribution of 17q21.31-inv different conformations in Europe, and suggested a complete fixation of the H1 haplotype followed by a *de novo* occurrence in the Homo line.

Recently, two independent studies analyzed the evolution of 17q21.31-inv by studying the duplicated architecture of the region (Boettger et al. 2012; Steinberg et al. 2012). Steinberg et al. (2012) used NGS data from more than hundreds of individuals and applied an integrative strategy that combined BAC-based assemblies, read depth-base copy number estimates, BAC pool sequencing and FISH to derive the structural conformation of 17q21.31-inv. The authors identified distinct copy number polymorphisms (CNPs) associated to H2 and H1 haplotypes: a short duplication (CNP155) associated to H2 and a long duplication (CNP205) exclusively associated to the H1 haplotype. In-depth analysis of the combination of these duplication and inverted allele status reported four main structural haplotypes. Furthermore, the frequency of the 17q21.31 inverted allele in the African continent was inferred by surveying a large collection of new population samples from different sources (e.g. 1000GP). Remarkably in contrast with earlier observations

(Stefansson et al. 2005), it was reported that the different inversion-associated haplotypes were segregating at fairly high frequencies in some populations with African ancestry.

In light of these new results, Steinberg et al. (2012) proposed a complex, novel model to explain 17q21.31-inv evolutionary trajectory. First, an ancestral H2 haplotype appeared in central or eastern Africa and spread to southern regions before the emergence of anatomically modern humans. Then, the region (re-)inverted back to the direct orientation (H1) approximately 2.3 Million years ago and spread throughout the Homo lineage ultimately becoming the predominant haplotype. The authors also note that the other structural haplotypes (H2D and H1D) appeared by convergent evolution and represent younger evolutionary events, as the duplication events in the two major clades (H1 and H2) have occurred independently. Noticeably, Steinberg et al. (2012) also report that there is only one NAHR-mediated haplotype (H2D) that predisposes to the syndromic 17q21.31 microdeletion. H2D structural conformation is characterized by the presence of directly-oriented homologous SDs flanking the disease-critical region and it is associated with a duplication of the *KANSL1* locus (Pagon et al. 2013). H2D appears to be significantly frequent (freq. = 25%) in some European populations.

Similar conclusions were reached in a parallel study by Boettger et al. (2012), where two duplications of the *KANSL1* locus, one in each genomic background (H1 and H2), have also been reported. According to the authors, these architectural changes originated a new transcript of the *KANSL* gene, which may have an impact on female fertility.

In summary, there exist contradictory hypotheses to explain the high genetic divergence observed between H1 and H2 17q21.31-inv haplotypes in modern humans. The hypothesis of selection shaping the region is still a matter of debate, as the observed patterns of variation across the region could be explained by particular demographic events that do not involve selection.

3.4 Inversions and human diseases

That fact that the inversion of a DNA segment could interfere with gene function by disrupting its reading frame or rearranging the position of promoters,

enhancers and other regulatory elements, should not be surprising. However, no inversion has been found to be the only cause of a disease although some are found to be the most common cause of pathology. For instance, a recurrent inversion located on the X chromosome rearranges the coagulation factor VIII and causes hemophilia in 42% of the cases (Antonarakis et al. 1995). Another example is an inversion in chromosome X (Xq28) associated to Hunter syndrome in 13% of the patients (Bondeson et al. 1995). Another inversion in the chromosomal locus 16p11.2 is associated to asthma and obesity (González et al. 2014). In some cases inversions do not cause disease directly but due to the characteristic duplicated architecture of inversion breakpoints. Instead, they increase the probability of disease through the occurrence of unbalanced rearrangements in the offspring. This means that parents who carry the inversions in heterozygosis confer a predisposition to further deletion of the inverted segment in subsequent generations (Feuk 2010; Gu, Zhang, and Lupski 2008; Zody et al. 2008) (see section I1.3.3).

The mechanistic details by which inversions contribute to complex genomic disorders are still unclear. Even with novel technologies allowing the refinement and characterization of inversion breakpoints, the understanding of how inverted rearrangements represent a source of genetic variation and simultaneously cause human disorders is still an important challenge to the research of human genetics.

3.4.1 Genome-wide association studies

The large efforts on the application of GWAS for the analysis of genome function, especially in the context of studies of genome variation has also allowed the discovery of regions of the genome that harbor genetic variants that confer risk to different types of complex diseases (Kingsmore et al. 2008). Theoretically, this technique can be applied to scan for inversion association to diseases. However, the applicability is subjected to a) high or perfect LD between the inversion and a surrogate SNP and b) the presence of the SNP in the genotyping microarray. In addition, it is not possible to look for direct association between inversions and disease due to lack of high-throughput genotyping methods. Despite all these limitations, there have been some successful attempts in correlating inversion to disease: by means of association studies, an inversion that provides a genetic basis for the joint susceptibility to asthma and obesity in European populations has been found recently (González et al. 2014). The authors genotyped the inversion via SNP

array data and performed association analysis in a combined sample set of 317 cases and 543 controls, finding significant (OR = 0.48, p-value = $5.5x10^{6}$) between disease phenotype and inversion genotype.

3.5 Expanding knowledge of inversions in the human genome: the InvFEST project

As we have discussed in previous sections, the map of human inversions is still quite limited. Further, our understanding of the number of inversions, the size distribution and the frequency distribution is probably incomplete due to biases in the approaches used for variation identification. The INVFEST (INVersion Functional & Evolutionary Studies) project was designed to address this deficiency and to expand the knowledge of polymorphic inversions in humans (see URLs). INVFEST aims to do a global study of polymorphic inversions in the human genome starting with the creation of an exhaustive and accurate catalogue of inversions. It aims to then study their effect on nucleotide variation, their evolutionary history and finally examine their functional consequences. To date, results of the project are regularly updated to `InvFEST` database (Martinez-Fundichely et al. 2014) (see section I2.1.4.3). In addition, in the framework of the project two techniques have been developed aiming to genotype inversions in a high-throughput manner: an optimization of the iPCR protocol (Aguado et al. 2014) (see section I1.4.2.5) and another approach based on probe hybridization and multiplex amplification (Villatoro et al., in preparation). In addition, an exhaustive comparison of the HuRef genome (Levy et al. 2007) with the HG18 reference genome assembly has been carried out and results compared to the original work, to determine, among other goals, the percentage of true and false positives inversions previously reported by Levy et al. (2007) (Vicente et al., unpublished results). Also `GRIAL`, a novel method based on PEM, aiming to specifically predict inversions, has been developed (Martínez-Fundichely et al., in preparation). PEM data from 8 individuals (Kidd et al. 2008) has been used as input for `GRIAL` to predict inversions. Results have been compared, merged if necessary with inversions in the literature and uploaded to `InvFEST` (Martinez-Fundichely et al. 2014, see URLs). In addition, also as part of the project, Lucas-Lledó and Cáceres (2013) simulated PEM data to quantify and track down the origin of false positives and negatives along sequencing, mapping, and downstream inversion predictions. They show that PEM is very appropriate to detect a wide range of inversions, even with low coverage data. However >80% of inversions located between segmental

duplications are expected to go undetected by the most common sequencing strategies. The last INVFEST published method is `svgem` (Lucas-Lledó et al. 2014) an expectation-maximization implementation to estimate allele and genotype frequencies, calculate genotype posterior probabilities, test for Hardy-Weinberg equilibrium and identify population differences.

We consider the analysis of INVFEST inversion frequencies a piece of helpful data to provide an overview of inversions that change frequencies across populations, as it is reasonable to hypothesize that they may be having an impact on gene expression. Therefore, in the following paragraph and figure (**Figure 14**) there is a descriptive overview of the frequency patterns of a subset of 41 inversions in 550 Hapmap individuals from different populations, genotyped by an approach based on probe hybridization and multiplex amplification (Villatoro et al., in preparation). Global inverted allele (with respect to human reference genome HG19) frequencies of the 41 inversions vary between 0.45% to 98.81% and derived allele (with respect to ancestral conformation) frequencies vary between 0.45% to 97.35%, although for some inversions the ancestral allele cannot be determined, mainly due to the lack of genome reference sequence in primates for the inverted region. We observe that a few inversions are either present in low frequencies in all populations or even absent in certain continents (e.g. HsInv0284, HsInv097, HsInv0790, HsInv0061) suggesting a recent origin or a detrimental effect that keeps the inverted rearrangement low in frequency (**Figure 14**). In addition, in most cases the oldest arrangement is the one highest in frequency, although in many inversions (13/40) the derived allele surpasses in frequency the ancestral one. Furthermore, in some cases allele frequencies greatly differ across different populations (**Figure 14**). In general, we observe that in most cases the oldest arrangement is the one highest in frequency. Interestingly, we observe that in many inversions (13/40) the derived allele surpasses in frequency the ancestral one. Furthermore, in some cases allele frequencies greatly differ across different populations (**Figure 14**). For example, in HsInv0114 case the ancestral allele (*Std*) is highly frequent (>75%) in African populations, but the derived allele (*Inv*) is the most common one in European and Asian populations, with frequency >60%. Contrarily, in some cases (e.g. HsInv0041) the supposedly ancestral orientation shows a low frequencey in Africans but a high frequency in non-African populations. Finally, the frequency of some inversions is significantly different among Eurasians (e.g. HsInv0124, highly frequent in Europeans but not in Asians) or even different among populations from the same continent (e.g. HsInv0095, frequent in CHB and JPT populations but not in GIH). Combined with additional evidences, these frequency patterns suggest that some inversions could be

under selection in the human lineage (D. Castellano and M. Cáceres, unpublished results).

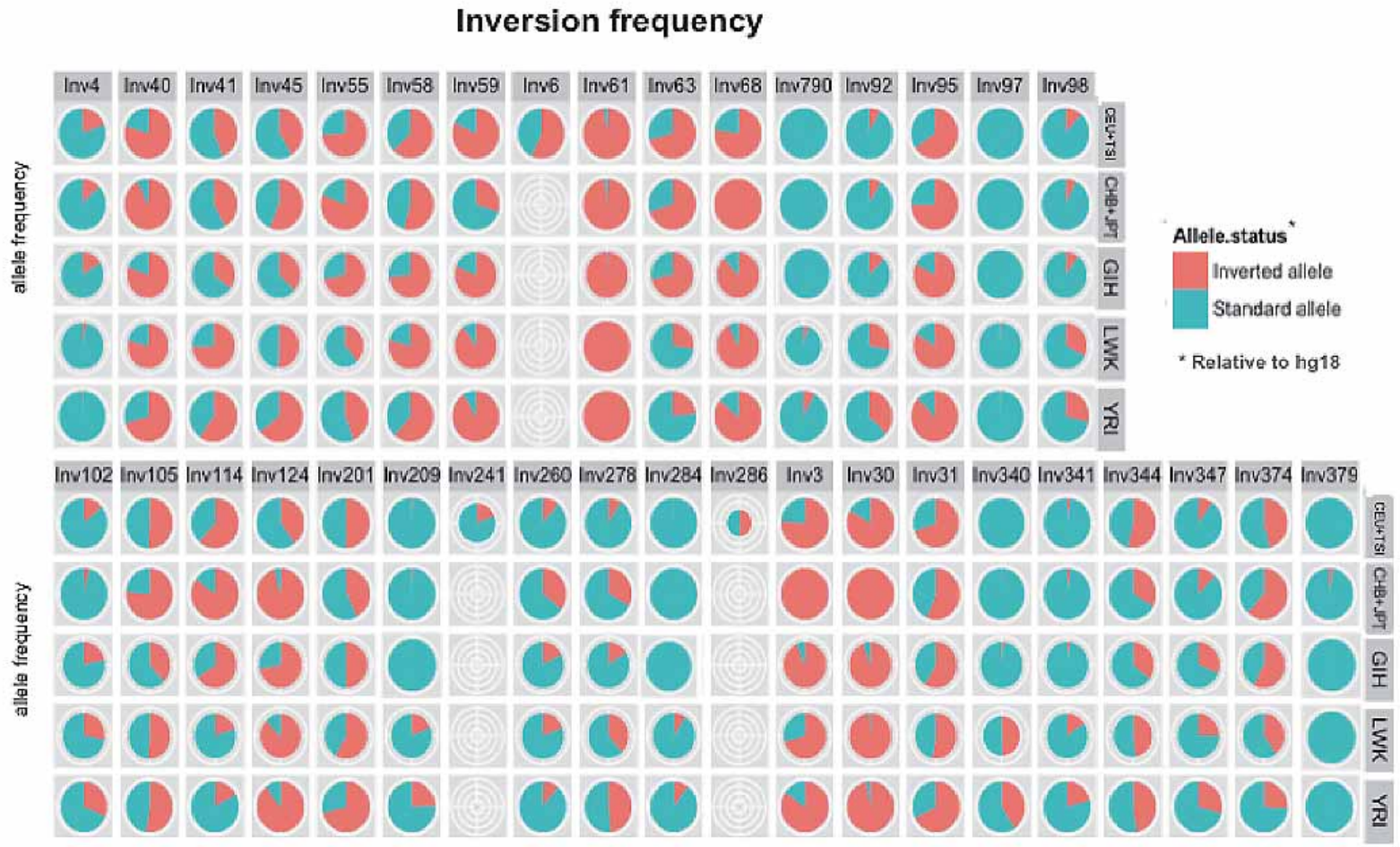


Figure 14 – Frequency of inversions in HapMap populations – Frequencies of autosomal inversions of the 44 inversions set in 7 HapMap populations (CEU, TSI, GIH, CHB, JPT, YRI, LWK). Inverted and standard alleles are indicated in red and blue, respectively, whereas circle sizes are proportional to the sample size (number of genotyped individuals). Inversion genotypes have been obtained by MLPA (S. Villatoro and M.Cáceres, unpublished results) except for HsInv0006, that has been genotyped by conventional PCR in 60 HapMap samples from CEU population (D. Vicente and M.Cáceres, unpublished results); HsInv0241 and HsInv0286, that have been genotyped by iPCR in 92 and 72 samples from CEU population, respectively (Aguado et al. 2014).

II OBJECTIVES

The objectives of this PhD can be summarized as follows:

1. Benchmark the performance of `GRIAL` against other inversion-detection PEM based methods in the literature.

2. Develop scoring systems and strategies to filter false positive inversion predictions and obtain an accurate catalogue of human polymorphic inversions.

3. Gain further insight on the functional impact of polymorphic inversions in the human genome, particularly aiming to identify inverted rearrangements that modulate gene expression, by studying the effect of 44 inverted rearrangements in blood cell lines of 550 individuals from European, Asian and African ancestry.

4. Apply simple inversion genotyping-free strategies to scan for inversion-eQTL associations and carry out a proof of concept by detecting inverted rearrangements affecting gene expression in non-blood tissues.

5. Elucidate the mechanisms by which inversions modify gene expression, either by mutational effects, positional effects or by association with a causal haplotype.

6. Find possible associations of inversions with disease and complex phenotypes, by surveying a comprehensive compendium of genome-wide association studies (GWAS).

III MATERIALS AND METHODS

URLs section

The URLs for databases, software and other sources information and tools used and referenced in the document are as follows:

Human Genome Structural Variation Project (HGSV)
http://hgsv.washington.edu/

INVFEST project
http://grupsderecerca.uab.cat/cacereslab/content/invfest

InvFEST database
http://invfestdb.uab.cat/

GRIAL
http://grial.uab.es

Expression Atlas
http://www.ebi.ac.uk/gxa/home

dbVar
http://www.ncbi.nlm.nih.gov/dbvar

SMALT v.0.6.1
http://www.sanger.ac.uk/resources/software/smalt/

Online Mendelian Inheritance in Man (OMIM)
http://www.omim.org/

HapMap project
http://hapmap.ncbi.nlm.nih.gov/

1000 genomes project
http://www.1000genomes.org/

Genome browser
http://genome.ucsc.edu/cgi-bin/hgGateway

Functional genomics database
http://www.ebi.ac.uk/arrayexpress/

Geuvadis repository
ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/

GWAS Central
http://www.gwascentral.org

GTEx
http://www.gtexportal.org/home/

MalaCards
http://www.malacards.org/

NCBI eQTL browser
http://www.ncbi.nlm.nih.gov/projects/gap/eqtl/index.cgi

seeQTL browser
http://gbrowse.csbio.unc.edu/cgi-bin/gb2/gbrowse/seeqtl/

eQTL resources at the Pritchard lab
http://eqtl.uchicago.edu/Home.html

eQTL browser (Prichard lab)
http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/

QTL Archive
http://databib.org/repository/555

Generation of GRIAL inversion predictions

Paired-ends mapped to HG18 from fosmids of 9 individuals (ABC7, ABC8, ABC9, ABC10, ABC11, ABC12, ABC13, ABC14 and G248) (Kidd et al. 2008) were obtained from the Human Genome Structural Variation Project (see URLs) and were subsequently analyzed by GRIAL to derive the inversion predictions (Martínez-Fundichely et al., in preparation). Libraries were merged into a large and unique dataset, with average fosmid length of 39,222 bp, standard deviation of 2,691 bp, and minimum and maximum lengths of respectively 25,163 bp and 49,224 bp. Duplicated fosmids were identified when distance (bp) between the 2 mapped ends of fosmids of the same individual and with the same orientation was ≤17 bp for G248 and ≤50 bp for ABC7-14 libraries. Applying this criterion, we identified 1,624 fosmids as potentially duplicated. Fosmids with this error were weighted down proportionally to the number of fosmids that could have been duplicated. Discordant fosmids in which the best mapping of the 2 ends overlapped by >50% (3,366 fosmids) were excluded from the input dataset. In total, 12,162 inversion discordant-in-orientation fosmids were selected. For inversion prediction at least 2 fosmids support of one or both breakpoints from the same or different individuals was required.

Generation of VH and PEMer inversion predictions

The dataset composed of 12,162 inversion discordant fosmid PEMs used as input for GRIAL to generate inversion predictions was also used to benchmark the results against VariationHunter (VH) (Hormozdiari et al. 2009) and PEMer (Korbel et al. 2009). For both programs, the default parameters were used with a minimum support of 2 fosmids to predict an inversion. For VH and PEMer inversion predictions missing the BP1, the breakpoint was inferred by taking the 3' extreme of the right outer-most read of the 5' cluster as the beginning of the BP1 and adding the maximum insert size length to the 5' extreme of the right inner-most read to define the end of the BP1 (**Figure 15**). The same procedure is applied to predict the 3' BP (BP2) of an inversion defined by a 5' cluster: the 3' extreme of the right outer-most read of the 5' cluster is considered as the beginning of the BP2 and the maximum insert size length is added to the 5' extreme of the 5' cluster right inner-most read to define the end of the BP2.

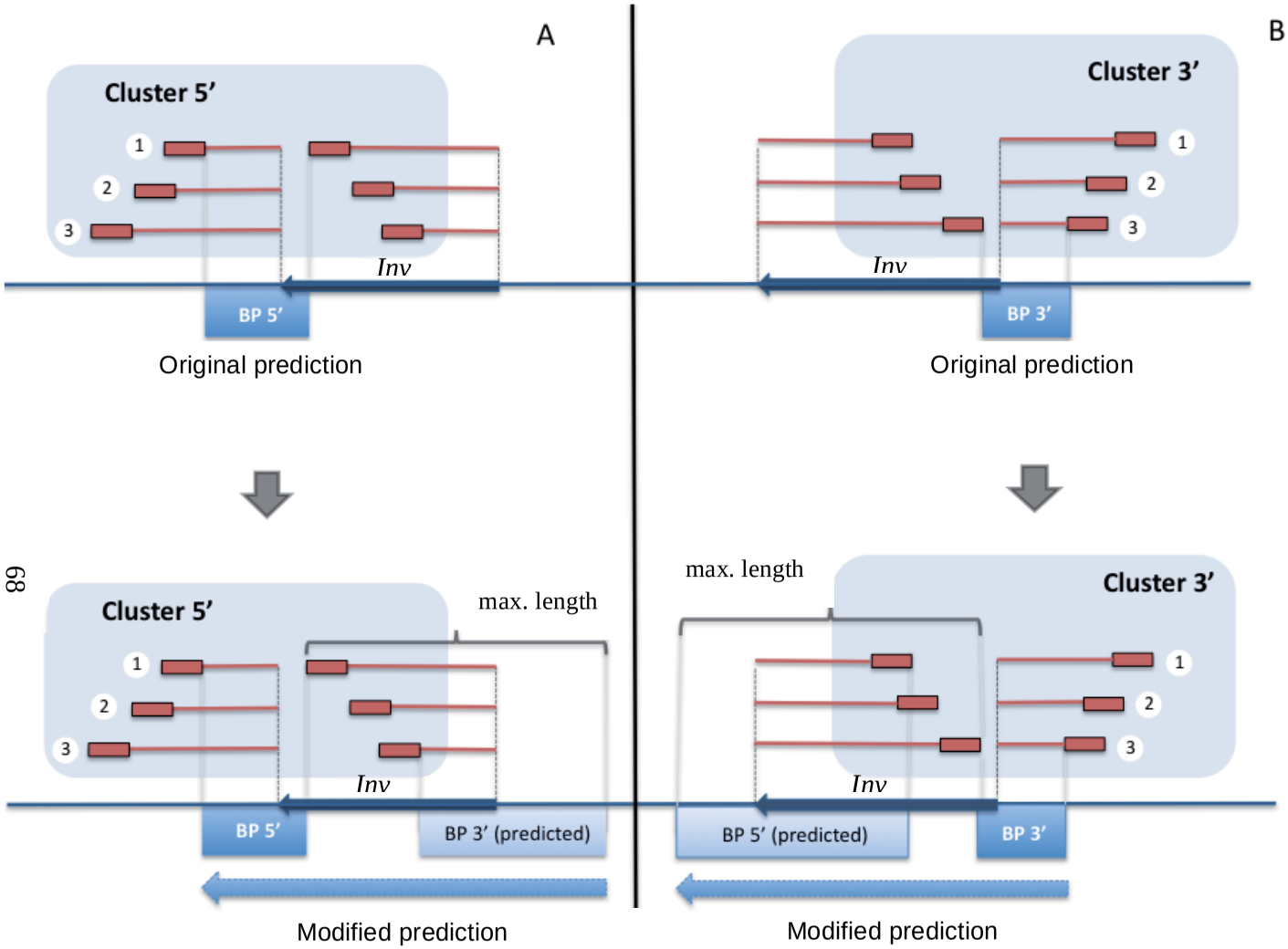


Figure 15 – VH and PEMer inversion BP inference - A) PEM read pairs (red boxes) form a cluster targeting the 5' extreme of an inversion (*Inv:* represented by a dark blue arrow) depicted in a reference genome (blue line). The inversion prediction (blue box) given by the SV method is not accurate, as it only spans the BP 5' (BP1) range and underestimates the real length of the inversion by missplacing the 3' extreme. In the bottom part, a more accurate inversion prediction (light blue arrow) is given: to infer the 3' BP (BP2) the 3' extreme of the right outer-most read of the 5' cluster is considered as the beginning of the BP2 and the maximum insert size length is added to the 5' extreme of the 5' cluster right inner-most read to define the end of the BP2. **B)** The same procedure is applied to predict the 5' BP (BP1) of an inversion defined by a 3' cluster: the 5' extreme of the left outer-most read is considered the beginning of the BP and the maximum insert size length is subtracted from the 5' extreme of the left inner-most read to define the end of the BP1.

Benchmarking of PEM methods across inversion prediction datasets

The degree of overlap between inversion predictions across methods (`GRIAL`, `PEMer`, `VH`) was calculated in a pairwise fashion, considering that 2 predictions overlap only if both BPs overlap. For instance, consider two inversion predictions: inversion "x" predicted by `GRIAL` and inversion "y" predicted by `PEMer`. We define that inversion candidate regions "x" and "y" overlap if BP1 of inversion prediction "x" overlaps with BP1 of prediction "y" and BP2 of inversion prediction "x" overlaps with BP2 of prediction "y". Then, in case of overlap, we assume that both predictions define a single inverted region. For `VH` and `PEMer`, original predictions were modified (missing BPs are inferred) as previously described.

Experimental validation of `GRIAL` inversion predictions

Experimental validation of some of the predicted inversions was carried out by PCR. Two pairs of primers were designed at each side of the 2 predicted breakpoints using `Primer3` (Rozen and Skaletsky 2000) and the 2 orientations, HG18 reference (standard or *Std*) and inverted (*Inv*), were tested by PCR amplification of DNA from the same 9 individuals from which the fosmids were derived (ABC7, ABC8, ABC9, ABC10, ABC11, ABC12, ABC13, ABC14 and G248) (Kidd et al. 2008) and HuRef (J. Craig Venter) DNA (Levy et al. 2007). DNA was extracted from Epstein-Barr virus-transformed B-lymphoblastoid cell lines of each individual as previously described (Aguado et al. 2013) or obtained directly from Coriell Cell Repositories (Camden, New Jersey, USA). PCR was performed in 25 µl reactions with 50-100 ng of DNA, 1.5 U of Taq DNA polymerase (Biotherm or Roche), 0.4-0.8 µM of each primer, 0.8 mM dNTPs, 1.5 mM $MgCl_2$, and 1x Taq DNA polymerase buffer by an initial denaturation of 5 min at 95°C, followed by 35 cycles at 95°C for 30 s, 58-62°C for 30 s, and 72°C for 30-120 s depending on the template size and a final extension at 72°C for 7 min. PCR products were analyzed by gel electrophoresis on 1.5-2% agarose gels stained with ethidium bromide and for those inversions in which the *Inv* orientation sequence was not available, the amplification products were purified and sequenced to determine the exact location of the breakpoints.

Building the gold-standard inversion dataset

The gold-standard inversion data set was generated from all the validated inversions found in the literature and a set of inversions validated in our laboratory (Martínez-Fundichely et al., in preparation). Due to possible assembly errors or chimeric fosmids, no inversions were considered validated based only in sequence information such as those of the HuRef genome (Levy et al. 2007) or whole-sequenced fosmids (Kidd et al. 2010) without additional independent experimental validation. In addition, only previously PCR validated inversions with sequence support for the breakpoints were considered (Korbel et al. 2007; Lam et al. 2010). Breakpoint positions were refined by the comparison of the *Std* and *Inv* sequences. For inversions with inverted repeats (IRs) at the breakpoints, the IR sequences in *Std* and *Inv* orientation were aligned using `Muscle` (Edgar et al., 2004) to identify sequence exchanges between the paralogous copies of the repeats in *Std* and *Inv* chromosomes and the point where these variants got exchanged due to the inversion. In these cases, the breakpoint intervals were defined by 3 or more consecutive paralogous sequence variants (PSVs) indicating a recombination between the IRs in the inverted sequences. Finally, 4 identified assembly errors in HG18 were also included in the comparison to increase the sample size. All together, we consider 59 inversion regions as part of the gold-standard inversion dataset.

Benchmarking of PEM methods on gold-standard inversion dataset

To calculate the overlap between inversion predictions by the 3 different PEM methods (`GRIAL, PEMer` and `VH`) and real inversions, the same criteria used previously to compare predictions across the different methods is applied: `GRIAL` original predictions and `VH` and `PEMer` modified predictions (with missing BPs inferred) are used; and we consider that a real inversion is detected by a PEM method only if both predicted BPs of the candidate inverted region overlap with the real inversion BPs. Alternatively, to calculate the overlap between BPs of real inversions and inversion predictions by `GRIAL, PEMer` and `VH`, only BPs originally predicted (not inferred) by the 3 different methods are considered. To compute the inversion detection efficiency, the ratio between the number of detected inversions and the total number of generated predictions to detect them is calculated. To determine the accuracy of the different methods in detecting inversion BPs, we selected two different datasets of real inversion BPs: the first one (“all BPs”) is

composed of all identified BPs per PEM method and the second one ("common BPs") is composed of a collection of 68 BP loci derived from the set of 39 predicted inversion regions detected in common by all methods. Then, we calculated the difference (number of bps) between BP coordinates of inversion predictions and real inversions for "all BPs" and "common BPs" datasets. Only the minimum difference between the 2 (start and end) real inversion boundaries was considered. If a given BP was identified by more than 1 prediction, measurements were averaged.

Misspriming analysis

During the detailed analysis of the 12,162 available fosmid sequences supporting `GRIAL` predicted inversions (Martínez-Fundichely et al., in preparation), it was found that many end-reads mapped not to the extremes but to the inner part of their original fosmid sequence which mapped concordantly to HG18 (suggesting the absence of any SV). These cases were not isolated but found in clusters: a single inversion region usually presented more than one overlapping PEM mapping to the inner part of its correspondent fosmid. Therefore, based on the hypothesis that there could be some end-reads not belonging to the extreme of fosmid but to a fosmid inner region where the sequencing primer paired, we consider as a candidate of this type of sequencing error all PEMs that have in common at least 1 read mapping with same features in same region. That is, we selected from the set of inversion discordant-in-signal fosmids supporting `GRIAL` inversion predictions, clusters of fosmids mapping to the same chromosome and strand, sharing the same sequencing orientation and with differences in the coordinates of the 5' end of the reads (sequence start) being less than 400 bp. According to these criteria, we selected 3,685 paired reads affecting 444 inversions predictions that correspond to 260 inversion regions. Then, for each problematic fosmid read, we added 100 bp to the length of the read and we extracted this number of nucleotides from the reference genome (HG18) region adjacent to the extreme coordinate of the read mapping locus; upstream for reads mapping to positive strand and downstream from reads mapping to negative strand. Then, to see if sequence could be generated due to a misspriming of the sequencing primer on the human genome, we mapped (`blastn -task blastn-short –gapopen 0`) the following set of primers to the extracted regions depending on the fosmid vector used for the construction of the library:

Primers for sequencing of ABC7-14 library fosmid ends (Kidd et al. 2008):

> pCC2 TM Forward Sequencing Primer
5′ - GTACAACGACACCTAGAC - 3′

> pCC2 TM Reverse Sequencing Primer
5′ - CAGGAAACAGCCTAGGAA - 3′

Primers for amplification of G248 library fosmids ends (Tuzun et al. 2005):

> pCC1 TM / pEpiFOS TM Forward Sequencing Primer
5' - GGATGTGCTGCAAGGCGATTAAGTTGG - 3'

> pCC1 TM / pEpiFOS TM Reverse Sequencing Primer
5' - CTCGTATGTTGTGTGGAATTGTGAGC - 3'

> T7 Promoter primer
5′ - TAATACGACTCACTATAGGG - 3′

> pCC1 TM / pEpiFOS TM RP-2 Reverse Sequencing Primer
5′ - TACGCCAAGCTATTTAGGTGAGA - 3′

We selected all hits with at least the last 8 bp of the 3’ extreme of the primer mapping with maximum 1 gap or mismatch, resulting in 295 possible mispriming cases. For G248 library, hits were obtained from 2 out of the 4 available primers: T7 Promoter primer and pCC1 TM / pEpiFOS TM RP-2 Reverse Sequencing Primer.

Paired ends remapping analysis in HG18 and HG19 reference assemblies and HG19 patches

Re-mapping of the fosmid paired-end reads to HG18, HG19 or individual genome patches was carried out using the `SMALT` v.0.6.1 program (see URLs). Reads were mapped independently and sequences shorter than 150 nucleotides and/or with less than 90% identity were filtered out. For each read, we kept track of the top 10 alternative mappings hits differing less than 0.05 from the top hit (hit with highest score). This value corresponds to the difference between the top hit score and the alternative hit score normalized by read effective length (number of read

nucleotides with good-quality PHRED score). Mapping quality for a read is considered to be good if the ratio between mapping score and effective read length is higher than 0.5. Fosmids with 2 mapped reads were classified into 3 categories: concordant, discordant and ambiguous (**Table 3**). We consider as concordant pairs those paired reads that map uniquely, with the expected orientation (+/−) and insert size (labeled as "CONC UNIQUE" in **Table 3**). These PEMs are not indicative of any inversion or other SV. Expected insert size (Insize) is defined as MinSize < InSize < MaxSize; where MinSize = 25,163 bp and MaxSize = 49,224 bp. These values correspond to the mean insert size of the concordant fosmids set minus or plus 3 standard deviations. Discordant pairs are candidates to target an inversion breakpoint as one of the reads maps in the opposite orientation than expected (+/+ or −/−). The discordant pair can map uniquely with good mapping quality for both reads (DISC ++).

Category	Orientation top hit	InSize top hit	Mapping quality top hit	Multiplicity	Alt. CONC conf.	Alt. DISC conf.
CONC UNIQUE	Expected	Expected	Good	Unique	-	-
CONC AMBIGUOUS	Expected	Unexpected	Good	Unique	-	-
	Expected	Irrelevant	Bad	Unique/ multiple	Irrelevant	No
	Expected	Irrelevant	Good	Multiple	Irrelevant	No
DISC ++	Unexpected	Irrelevant	Good	Unique	-	-
DISC +	Unexpected	Irrelevant	Good	Multiple	No	Irrelevant
	Unexpected	Irrelevant	Bad	Unique/ multiple	No	Irrelevant
DISC AMBIGUOUS (top DISC)	Unexpected	Irrelevant	Good	Multiple	Irrelevant	Yes
DISC AMBIGUOUS (top CONC)	Unexpected	Expected	Good	Multiple	Yes	Irrelevant

Table 3 - Scheme of fosmid PEMs classification after remapping - Concordant pairs that map uniquely, with good quality and expected orientation and insert size (InSize). In discordant pairs (DISC) one of the reads map in the opposite orientation than expected. In DISC ++, the reads pair map uniquely with good mapping quality for both reads. In DISC -, the reads map multiply and/or with bad quality, albeit with no alternative concordant conformation mapping. Ambiguous pairs map concordantly in orientation but discordantly in size and/or with bad mapping score (CONC AMBIGUOUS), or present both discordant and concordant conformations with similar mapping scores (DISC AMBIGUOUS); with the top hit presenting a discordant conformation (top DISC) or concordant conformation (top CONC).

Alternatively, if at least one of the reads of a discordant pair maps with bad quality, either uniquely or multiply, or both reads map with good quality but at least one maps multiply; but if in none of the cases an alternative concordant conformation exists, the pair is labeled as "DISC +" (**Table 3**). Ambiguous pairs for inversion present both discordant and concordant conformations with similar mapping scores (DISC AMBIGUOUS). Finally other read pairs not indicative of an inversion are labeled as CONC AMBIGUOUS. Only paired ends classified in DISC ++ or DISC + categories are considered to be reliable for inversion prediction. In HsInv0710 and HsInv0306 cases, we consider a pair to be ambiguous when 1 or both reads map to a copy of a highly-identical IR region (defined using `MEGABLAST` as sequences spanning ≥1 kb and ≥97% identity) and it has both concordant and inversion discordant alternative mappings (Aguado et al. 2014).

Normalization and filtering of LCL expression datasets

For the dataset Stranger 2007 (Stranger et al. 2007), we obtained the processed expression values for lymphoblastoid cell lines (LCLs) of 270 HapMap individuals from the Gene Expression Omnibus database (accession number GSE6536). The dataset derives from Illumina's human whole-genome expression array (WG-6 version 1). The gene expression dataset is available background-corrected, quantile-normalized (Bolstad et al. 2003) across replicates of a single individual and median-normalized across individuals of a single population. Expression values from sex chromosomes are provided. Principal component analysis (PCA) and multidimensional scale analysis (MDS) were performed to identify and discard possible outliers due to batch effects or other confounding factors. All samples passed the quality control and had available inversion genotypes. Then, array probes were mapped to Entrez gene identifiers. Probes not mapping to any Entrez gene or mapping to multiple genes were excluded from further analyses. For genes with several mapping probes, the probes with most variable expression across samples was kept. Genes with very low expression values (below 1st decile of the distribution) were filtered out and only genes analyzed in all populations were retained. Downstream differential expression analyses were based on 270 samples and 19,573 genes.

For the dataset Stranger 2012 (Stranger et al. 2012), we obtained raw expression values for 730 HapMap individuals from the Expression Altlas Database

(accession number E-MTAB-264, E-MTAB-198). The dataset derives from Illumina's human whole-genome expression array (WG-6 version 2). For each population, MA plots were generated and 4 distance and dependence based measures between each pair of sample replicates were calculated: a) root-mean-square deviation (Euclidean distance) for the top 500 genes with largest standard deviations between each pair of samples (function:plotMDS, R package: `limma`); b) Pearson correlation (function:cor, R package: `stats`); c) MA plot interquartile range (IQR); and d) MA plot median (functions:ma.plot, R package: `affy`). Afterwards, sample replicate outliers were identified. We consider an outlier an array that fell outside 1.5 times the interquartile range above the upper quartile and below the lower quartile of any of the 4 measure distributions (Euclidean distance, Pearson correlation, MA median or MA IQR distributions). Samples with both replicates identified as outliers were filtered out. For samples with 1 replicate identified as an outlier, only the non-outlier replicate was kept. After filtering 2.5% of the samples (18/730), individuals for whom inversion genotypes were available were kept. Then, the expression dataset was quantile normalized, averaged between replicates, median scaled across all populations and log2 transformed (function:normalizeBetweenArrays, R package: `limma`). Afterwards, gene expression values were mapped to Entrez gene identifiers and filtered as in the Stranger 2007 dataset. Subsequent differential expression analyses were based on 366 samples and 19,565 genes.

For Geuvadis project, bam files were downloaded from Geuvadis repository (see URLs) and filtered with `bamtools` (Barnett et al. 2011) to remove reads that were not uniquely mapped. Then, reads were mapped to Ensembl v.73 gene annotations with `HTSeq` (Anders, Pyl, and Huber 2014), with default parameters (L. Pantano, unpublished results). Genes expressed in less than 25% of the samples were filtered out, and individuals for which inversion genotypes were available were kept. Then, genes and exons analyzed in all populations and with non-null expression values in at least 1/3rd of the sample set (1/3*175 = ~58 samples) were retained. In summary, 14,950 protein-coding genes, 304,374 protein-coding gene exons and 7,895 RNA genes for 175 individuals were finally used in differential expression analyses. Expression values were transformed to log2-counts per million (function:voom, R package: `limma`).

Differential expression analysis of LCL expression datasets

For each one of the 44 genotyped inversions, differential expression (DE) analysis with respect to inversion genotype was performed in each expression dataset separately (Stranger 2007, Stranger 2012, and Geuvadis). For the Geuvadis dataset, the DE analysis was also performed separately for the protein-coding genes, protein-coding gene exons and RNA genes datasets. In all cases, the DE analysis was performed both for the pooled sample set (pooled-population) and for each population separately (population-specific), in order to detect possible differences among populations. A linear model, with gene expression being the response variable and inversion genotype the modeled variable of interest, was fitted for every gene (function:lmFit, R package: `limma`). Inversion genotype variable was either coded as number of inverted alleles {0,1,2} to be treated as a discrete quantitative variable (additive model) or coded as a categorical variable for pairwise comparison between different genotype groups. In each of the 3 possible pairwise comparisons, 2 genotype groups were pooled and compared to the remaining one: INV vs. HET+STD (recessive model), HET vs. INV+STD (overdominant model), STD vs. HET+INV (dominant model). For inversions located in chromosome X, only homozygotes STD vs. INV were compared for males and females separately. For the inversion located in chromosome Y, inversion carriers were compared against non-carriers. We filtered out all comparisons with inversion MAF < 5% or/and minimum number of samples per group < 4. For the pooled-population approach, population and sex factors were taken into account in the model, whereas for the population-specific approach only sex was modeled. In both approaches, confounding factors represented as surrogate variables were identified, estimated (R package: `sva`) and incorporated in the linear model. The statistical significance of the effect of inversion genotype on the expression of a particular gene is tested by Empirical Bayes moderated t-statistics test (function:eBayes, parameters:trend = TRUE, R package: `limma`). P-values were corrected for multiple testing by controlling the expected false discovery rate (FDR) to be below 10%. Two multiple testing corrections differing in stringency were applied: *cis* correction (considering only genes inside the inversion and up to +/− 1 Mb) and genome-wide correction (considering all genes tested).

Differential expression analysis of 8p23.1 and 17q21.31 inversions

For inversion 17q21.31**,** information about structural haplotypes and inverted allele genotypes was obtained from Boettger (2012) and Steinberg (2012). For Geuvadis, Stranger 2007 and Stranger 2012 datasets, 175, 206 and 535 samples were analyzed, respectively. For inversion 8p23.1**,** information about inverted allele genotypes was obtained from Salm et al. (2012). For Geuvadis, Stranger 2007 and Stranger 2012 datasets, 113, 191 and 348 samples were analyzed, respectively. In both cases the DE analysis was carried out analogously to the set of 44 genotyped inversions.

Association of inversions with known eQTLs (inversion-eQTL analysis)

In total, we collected eQTL information from 15 eQTL studies covering 11 different tissues. For each different tissue study, information relative to eQTL-gene location, significance of association (p-value), magnitude and direction of expression change (when available) was recorded. Only SNP-gene eQTL associations reported as significant in the correspondent eQTL study were selected, except for the ones obtained from Genevar and NCBI eQTL browsers (see URLs), where associations with original p-value < 10E-03 and < 10E-05 were obtained, respectively. Then, for each inversion, 1000GP SNPs in *cis* with the inverted region (inside the inversion or in the upstream and downstream 1 Mb flanking region) were selected. For each of CEU, TSI, CHB, JPT, LWK and YRI populations, linkage disequilibrium analysis between the inversion and the *cis* SNPs set was performed with PLINK to identify inversion tag-SNPs, that are defined as polymorphisms in high linkage ($r^2 \geq 0.8$) with inversions. Ultimately, eQTL datasets were surveyed for inversion tag-SNPs to identify inversion-eQTLs (that is, it was checked if inversion tag-SNPs where eQTLs in some of the studies) and results were interpreted.

Association of inversions with diseases and phenotypes from GWAS studies

For each inversion, 1000GP SNPs in *cis* with the inverted region (inside the inversion or in the upstream and downstream 1 Mb flanking region) were selected. For each of CEU, TSI, CHB, JPT, LWK and YRI populations, linkage

disequilibrium analysis between the inversion and the *cis* SNPs set was performed as described above. Then, the resultant SNP set was provided as input to GWAS Central database (Beck et al. 2014) to retrieve associations of each SNP with diseases or phenotypes. An association was considered significant if the original p-value was < 0.01 in the corresponding GWAS study. Subsequently, the SNPs of the candidate associations were queried for being genetic determinants of gene expression by crossing the information with the resulting eQTL database build for the inversion-eQTL analysis, and the results were interpreted.

IV RESULTS

Chapter 1

Refining a catalogue of human polymorphic inversions

SUMMARY - In this chapter, we refine an extensive catalogue of inversion predictions in the human genome that constitutes the foundation of subsequent downstream analyses aiming to determine the functional impact of a subset of validated inversions derived from this dataset. To address that, we first benchmark GRIAL (Geometric Rule Inversion Algorithm; Spanish for Grail), an algorithmic method developed in our laboratory focused on predicting inversions out of PEM data. Here, we first compare GRIAL performance with other structural variants detection methods based on PEM data, in order to assess the robustness, accuracy and sensitivity of GRIAL. Then, we reassess the validity of GRIAL predictions in updated and alternative human genome sequences and apply additional filters to discard artifacts to reduce the false positive predictions burden at a minimum level. Finally, we experimentally validate 2 inversion predictions using standard molecular biology techniques and characterize the breakpoints (BPs) of a particular inversion at the nucleotide level.

1.1 Benchmarking of GRIAL against alternative PEM-based methods for inversion prediction

GRIAL is a specialized algorithm developed in our laboratory that aims to define inversions accurately from PEM data by identifying and analyzing inversion-specific characteristics. The method is based on geometrical rules derived from expected inversion PEM patterns, used to cluster individual mappings belonging to each inversion breakpoint (BP), merge clusters into inversions and refine BP location. In addition, it tries to eliminate most false positives through a prediction scoring system, and also refines the BPs to the minimum interval. The method is already publicly available (see URLs) and will be published soon (Martínez-Fundichely et al. in preparation).

In order to benchmark GRIAL performance, we selected 2 widely used, PEM based methods: VariationHunter (VH) (Hormozdiari et al. 2009) and PEMer (Korbel et al. 2009). These methods differ in the strategies used to handle PEM data. PEMer selects the best mapping for each read pair, if various exist. VH allows as input multiple mappings for the same read, computes different combinations of SVs and selects the most parsimonious set, which is usually the one with the lowest number of predicted variants. It does this by trying to merge compatible clusters targeting both extremes of the rearranged region to give as result a set inversions defined by both BPs, similarly to GRIAL. Contrarily, PEMer does not perform this step and outputs inversions defined by a cluster corresponding to 1 of the 2 BPs. GRIAL and VH only use this strategy (predicting inversions by targeting only 1 BP) for predictions based on clusters for which no compatible counterpart is found. In addition, VH does not output predictions spanning more than 1 Mb and does not accept as input paired ends mapping at random chromosomes.

GRIAL was used to build a refined catalogue of inversions predictions based on one of the largest existing PEM data sets (see Materials and Methods). In brief, 12,162 (8,797 after filtering) inversion discordant fosmids from 9 individuals belonging to the HGSV were used as GRIAL input to predict ~700 inversions distributed genome-wide. This set was subsequently refined by applying several filters and scoring systems to obtain the most reliable and accurate catalogue of human polymorphic inversions to date (Martínez-Fundichely et al. 2014). However, to benchmark GRIAL performance against other methods, here we use raw predictions.

VH and PEMer were run with the same original fosmid data set used by GRIAL as input, providing fosmid library minimum, maximum and mean insert size as well as minimum support per prediction as parameters (see Materials and Methods). In addition, since each method predicts BP coordinates differently and inversions can be predicted by targeting 1 or both of its BPs, the predictions were modified in order to be comparable across different methods. In predictions derived only from 1 BP, the missing one was inferred by adding the maximum fosmid length to create an interval where the BP should be located (see Materials and methods).

First, we observe that all methods predict a similar number of inversions. Considering a minimum support of 2 discordant fosmids, GRIAL predicted 690 inversions located at 324 regions, which results in 636 inversions located at 306 regions if predictions placed in random chromosomes are excluded (Martínez-Fundichely et al., in preparation). Of those 636 predictions, 220 (34.6%) have PEM support for both BPs, while for 416 (65.4%) only 1 BP has been detected. In addition, 201 inversions are supported by a single PEM in each BP, or a single PEM in 2 individuals, and are identified just by the merging of all the information. VH predicts 633 inversions located in 364 regions, but only 28 (4.4%, ~30% less than GRIAL) have PEM support for both BPs, while the rest is detected by only 1 BP. We observe that although PEMer predicts more inverted rearrangements compared to the other methods (772, 720 if predictions placed in random chromosomes are excluded), this prediction set corresponds to a number of inversion regions (361, of which 342 in non random chromosomes) very similar to VH and GRIAL. This is caused because PEMer predicts inversions on the basis of one of the 2 BPs and it does not merge predictions in any case, contrarily to VH and GRIAL. Therefore, two different PEMer inversion predictions may correspond to the same inversion region.

To assess the agreement between the different algorithms in predicting equivalent inversion regions, we measured the degree of overlap between inversion predictions across methods. We considered that 2 predictions overlap only if both BPs overlapped (Materials and Methods). The results show that there exists a high overlap of GRIAL predictions with the tested PEM methods set (**Figure 16**).

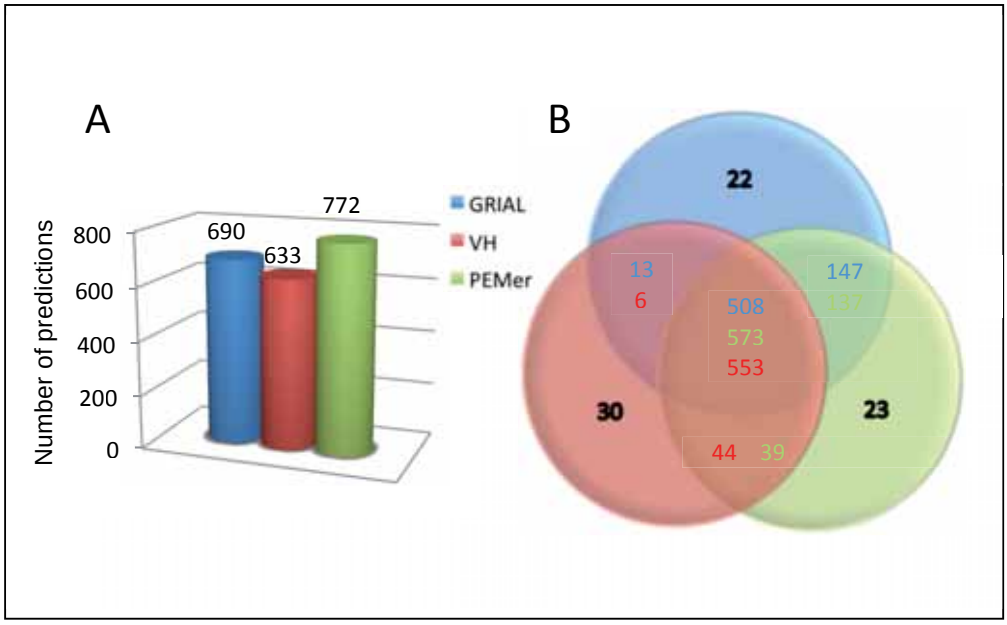


Figure 16 - Comparison of GRIAL with PEM methods inversion predictions – A) Number of inversion predictions per PEM method (GRIAL, VH, PEMer), including predictions in random chromosomes. B) Overlap of GRIAL, VH, PEMer inversion predictions.

Interestingly, we observe that GRIAL uniquely identifies a subset of 22 inversions that is not predicted by any of the other benchmarked methods (**Figure 16**, **Table 4**). This subset is mainly composed (17 out of 22) by inversions supported by only 2 paired-ends: 1 supporting BP1 and another one supporting BP2 (1+1−). Both VH and PEMer do not predict these inversions due to insufficient signal, as for both methods the minimum required support for an inversion prediction is at least 2 paired-end supporting each BP, and they cannot be parameterized to accommodate a lower fosmid support. Despite the low support, we demonstrate that 4 of these predictions are real (HsInv0409, HsInv0410, HsInv0201, HsInv0301), as 3 of them have been experimentally validated by PCR (2 as part of this thesis, see below) and the remaining one is a well-known inversion involved in further rearrangements that cause Williams syndrome (Osborne et al. 2001). We observe that 4 out of 22 unique inversions of GRIAL are not predicted by PEMer because the fosmids supporting the predictions do not satisfy a particular rule of this method (Lmax-Lmin rule), which does not allow 2 paired ends to be clustered if the differences between the coordinates of the reads of the different pairs is bigger than the difference between the maximum and the minimum insert size. However, we consider that paired ends that do not fulfill this condition can support true inversion predictions; therefore

GRIAL has not implemented this rule. Among this subset of GRIAL unique predictions, VH also misses 14 inversion predictions spanning more than 1 Mb, which are filtered. This rule may reduce the number of false positive predictions, because evidences for polymorphic inversions in the human genome spanning more than 1 Mb are scarce, with few exceptions. However, to our knowledge there is no biological reason for a particular limit on inversion size, consequently GRIAL does not apply any filter based on this feature. In fact, 1 of the 4 validated inversions (HsInv0301) spans more than 2.4 Mb. Finally, in a couple of cases (HsInv0437, HsInv0441) a GRIAL prediction is missed by the other PEM methods due to paired ends being differently clustered, resulting in the prediction of different inversion regions (**Table 4**).

InvFEST	Sup.BP1	Sup.BP2	Chr.	Size (kb)	VH filtered	PEMer filtered
HsInv0437	1	1	1	4.6 Mb	>1 Mb	diff. cluster
HsInv0238	1	1	1	4.5 Mb	1+1-| >1 Mb	1+1-
HsInv0273	2	1	5	1.9 Mb	>1 Mb	Lmax-Lmin rule
HsInv0201	1	1	5	1058 bp	1+1-	1+1-
HsInv0281	1	1	5	1.9 Mb	1+1- | >1 Mb	1+1-
HsInv0301	1	1	7	2.4 Mb	1+1- | >1 Mb	1+1-
HsInv0313	1	1	8	337 Kb	1+1-	1+1-
HsInv0116	1	1	9	1.7 Mb	1+1- | >1 Mb	1+1-
HsInv0343	1	1	13	40.4 Kb	1+1-	1+1-
HsInv0358	1	1	15	1.0 Mb	1+1- | >1 Mb	1+1-
HsInv0560	1	1	16	1.8 Mb	1+1- | >1 Mb	1+1-
HsInv0152	1	1	16	1.1 Mb	1+1- | >1 Mb	1+1-
HsInv0384	1	1	22	2.01 Mb	1+1- | >1 Mb	1+1-
HsInv0386	1	1	22	2.9 Mb	1+1- | >1 Mb	1+1-
HsInv0387	1	1	22	2.6 Mb	1+1- | >1 Mb	1+1-
HsInv0388	1	1	22	2.6 Mb	1+1- | >1 Mb	1+1-
HsInv0409	1	1	X	1345 bp	1+1-	1+1-
HsInv0410	1	1	X	3346 bp	1+1-	1+1-
HsInv0412	1	1	X	45.3 Kb	1+1-	1+1-
HsInv0441	3	0	1	149.9 Kb	diff. cluster	Lmax-Lmin rule
-	3	0	8	96.8 Kb	chr random	Lmax-Lmin rule
HsInv0646	0	4	1	4.3 Mb	>1 Mb	Lmax-Lmin rule

Table 4 - GRIAL unique inversion predictions – The list of inversion regions uniquely predicted by GRIAL (not by VH nor PEMer) is provided. Inversion identification (InvFEST id) corresponds to the key id for the GRIAL predicted inversion in InvFEST database (Martinez-Fundichely et al. 2014) and experimentally validated inversions are indicated in bold. Inversion size is indicated in kilobases unless specified otherwise.The number of supporting paired ends per inversion BP is indicated (Sup.BP1 for leftmost BP, Sup.BP2 for rightmost BP). VH filtered and PEMer filtered columns specify the reason for inversion prediction being filtered by corresponding method. Lmax-Lmin rule: PEMer clustering strategy does not allow 2 paired ends to be clustered if the differences between the coordinates of the reads of the different pairs is bigger than the difference between the maximum and the minimum insert size. Predictions supported by 1 fosmid per BP are indicated by 1+1-.

Then, to assess the performance of GRIAL in identifying real inversions and provide refined BP regions, we benchmarked GRIAL, VH and PEMer predictions to a collection of validated inversion regions. This gold-standard inversion data set was generated from all the validated inversions found in the literature and those validated in our laboratory, both in this work (HsInv0409, HsInv0410, see Materials and Methods) or elsewhere (Martínez-Fundichely et al. in preparation), and is composed by 59 inversions (Materials and Methods). Here, we apply the same criteria used for the overlapping predictions analysis (Materials and Methods).

Results (**Table 5**) show that GRIAL and PEMer perfom better than VH by predicting a superior percentage of real inversions (GRIAL: 83.0%, PEMer: 83.0%, VH: 71.2%). Moreover, GRIAL clearly outperforms the rest when the number of originally predicted real BPs is accounted for (that is, when original instead of modified VH and PEMer predictions are considered). In this case, we observe that GRIAL predicts ~13-23% more BPs than PEMer and VH, respectively. Finally, if we consider the set of real inverted regions predicted by the different algorithms, we observe that GRIAL is able to identify them by providing a lower number of predictions compared to the other methods. Considering the detection efficiency as the ratio between the number of detected inversions and the total number of generated predictions to detect them, GRIAL is clearly superior to the other methods, especially compared to PEMer (detection efficiency of 0.80 versus 0.53 for GRIAL and PEMer, respectively).

Feature	GRIAL	PEMer	VH	Common
Detected inversions	49 (83%)	49 (83.0%)	42 (71.2%)	39 (66.1%)
Detected BPs	98 (83%)	83 (70.3%)	71 (60.2%)	68 (57.6%)
Predictions	61	92	68	-
Detection efficiency	0.8	0.53	0.62	-
Common BPs average dist. error (bp)	4021	13461	12776	-
Common BPs median dist. error (bp)	594	10061	9278	-
Common BPs std. dev. dist. error (bp)	2654	9728	9832	-
All BPs average dist. error (bp)	14195	36508	13739	-
All BPs median dist. error (bp)	815	11672	9492	-
All BPs std. dev. dist. error (bp)	39282	102614	14486	-

Table 5 - Benchmarking of different inversion prediction methods against gold-standard inversion dataset – Per each method, the number and percent of detected real inversions (detected inversions), number and percent of real BPs (detected BPs), total number of predictions (predictions) and the ratio between number of detected inversions and the produced predictions to detect them (detection efficiency) is shown, as well as number and percent of inversion regions and BPs detected by all methods (Common). The common 68 BPs detected by all methods have been used to calculate the average, median and standard deviation distance error for BP prediction (and also for all BPs).

Finally, we checked the performance of the different SV methods in accurately predicting inversion BP regions (Materials and Methods). For that, we use all identified BPs per method (labeled as “all BPs” in **Table 5** and **Figure 17**) and a collection of 68 BP loci derived from the set of 39 predicted inversion regions by all methods (labeled as “common BPs” in **Table 5** and **Figure 17**). As a measure of BP detection accuracy, we considered the differences between real and predicted BP boundaries coordinates, (the number of superfluous or missing nucleotides predicted per BP). For each BP prediction, the minimum difference between the 2 (start and end) real inversion boundaries was considered. If a given BP was identified by more than 1 prediction, measurements were averaged. Results (**Table 5**, **Figure 17**) show that GRIAL outperforms the other PEM methods in refining BP regions by providing, on average, 4021 extra bps per BP, 3 times less than PEMer and VH (13461 and 12776 extra bps, respectively). However, the difference is less pronounced when we consider the commonly predicted BPs set, where the average error distance (14,195 bp) is similar to the one predicted by VH (13,739 bp) but still smaller than PEMer (36,508 bp). We observe that the distributions of the error distances show a positive skew and high variance with presence of outliers corresponding to BPs located in big and highly repetitive regions such as in the case of inversions 15q13.3, 8p23.1 and 17q21.31-21.32. The presence of these outliers heavily affects the mean, especially when all BPs are considered (**Figure 17**). Hence, instead of the mean, a parameter more robust to outliers should be employed. If we compare the median of distance errors across methods for the common BPs set, we observe that the error for GRIAL is 16 times lower compared to the other methods (594 bp compared to 10,061 and 9,278 bp for PEMer and VH, respectively). Therefore, this indicates that GRIAL is a particularly good method for predicting inversion BPs with high accuracy.

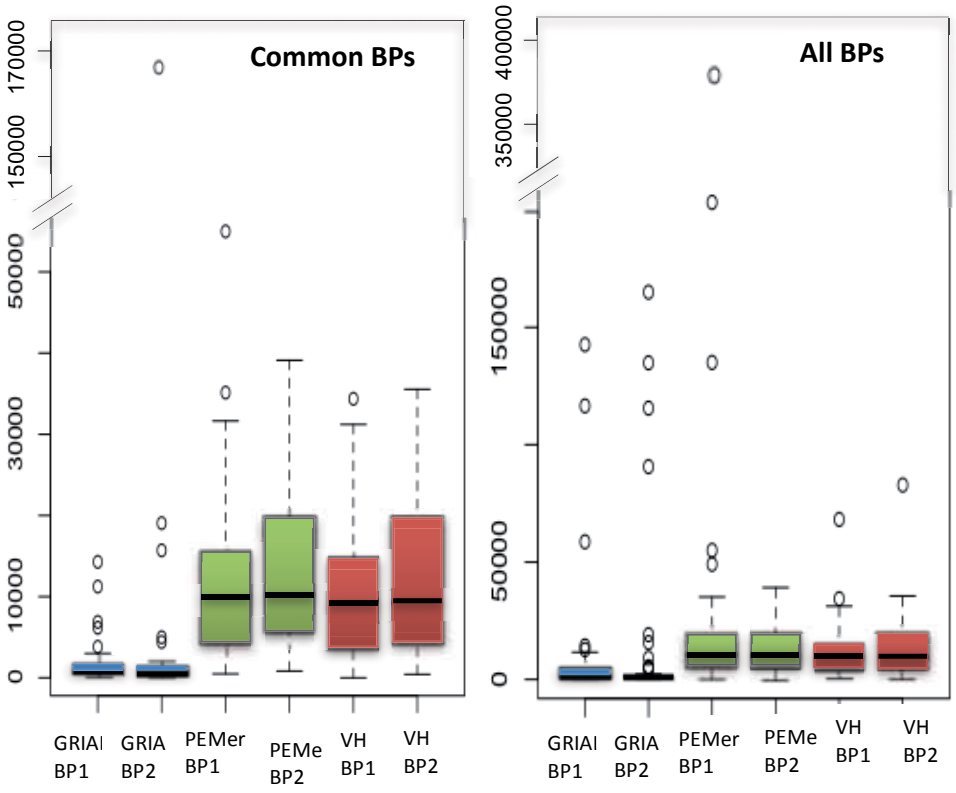


Figure 17 - BP detection accuracy by SV prediction methods - Differences (in bp) between real and predicted BP boundaries coordinates by `GRIAL`, `VH` and `PEMer` are shown, considering all identified BPs per method (All BPs) and the common set (Common BPs).

1.2 Refinement of the inversion catalogue: filtering false positive predictions and validation of inverted region candidates

False predictions arise from technical issues in several steps of the inversion prediction process, both at early steps such as in the paired ends library generation and also in final stages as in PEM mapping, clustering and filtering. For example, some false positive predictions arise from fosmid chimeras that are generated by coligation of unrelated DNA fragments during the fosmid library generation (Tuzun et al. 2005). However, most of the spurious predictions are generated during the mapping step and are associated with incorrect mapping of the fosmid paired ends (Lucas Lledó and Cáceres 2013) due to sequence differences between individuals or errors and gaps in the genome reference sequence. In addition, a common source of

mapping errors is sequence divergence or gene conversion between paralogous IRs, as they can produce PEM signal characteristic of inverted regions. Another type of mapping error is due to the mixing of 2 haplotypes in the HG18 assembly (Antonacci et al. 2010). Lastly, it has been observed that inverted duplications in tandem also create spurious inversion PEM patterns. These and other inversion false prediction sources have been taken into account to filter and label inversion predictions (Martinez-Fundichely et al. 2014, Martinez-Fundichely et al., in preparation).

Here, both by manual inspection and systematic bioinformatic analysis of many available sequences (including remapping of paired-end sequences and analysis of fully sequenced fosmids or other available human sequences) and also by PCR amplification of predicted inversion BPs, we identified a series of false discordant PEMs that have originated incorrect GRIAL inversion region predictions. Altogether, this collection of problematic artefactual fosmids was employed to filter 387 GRIAL false positive predictions in order to build a reliable catalogue of inversions. The different filtering processes are explained in this section.

1.2.1 PCR amplification of inversion predictions supported by duplicated fosmids.

It has been shown that during the construction of a the PEM library, fosmids can undergo artifactual duplication (Tuzun et al. 2005). Regarding inversions, discordant in signal duplicated fosmids can generate false positive predictions as they artificially increase the discordant signal. Fosmid duplication events are not rare: from a total of 12162 inversion discordant fosmids, GRIAL identifies 13.4% (1624/12162) duplicated fosmids in a pre-filtering step of the inversion prediction analysis (Martínez-Fundichely et al., in preparation). To overcome false positive events, GRIAL weights down duplicated fosmids when included in an inversion prediction (Martínez-Fundichely et al., in preparation) and discard predictions only supported by duplicated fosmids.

Here, we aim to check the veracity of a few inversion regions supported solely by duplicated fosmids. We have chosen 2 inversions supported by 4 fosmids in total (**Table 6**). To validate them, we have selected for PCR amplification the BPs of both inversions. Remarkably, 2 out of the 4 duplicated fosmids, one per each inversion, are fully sequenced and indicative of inverted rearrangement when mapped to HG18 and HG19 human genome assemblies. In addition, one of the 2

fosmids (Genbank Id: AC226692.2) was identified as targeting an inverted region in Kidd et al. (2010). For both inversions, amplification of predicted BPs was performed using the same 9 individuals and protocol used to validate GRIAL inversion predictions (Material and Methods).

HSVD id	Genbank id	PEMs	Predicted inversion BPs[1]	Duplicated
ABC8_000040890500_F15	AC226692.2	chr14 30829354-30856268	chr14: 30855719-30870075	ABC8_40905100_L15
ABC8_2146440_O20	AC231266.2	chr3 170335138-179489584	chr13: 170316698-170317453	ABC8_2146940_O22

Table 6 - Duplicated fosmids features – Coordinates of fosmid PEM reads (PEMs) obtained from Kidd et al. (2008). Predicted inversion breakpoints were inferred by mapping fully sequenced fosmids AC226692.2 and AC231266.2 to HG18. [1] Amplification of predicted BPs (Materials and Methods) showed standard orientation (*Std*) in all tested samples.

PCR results indicate that for the 2 putative inversion regions, all tested samples present standard conformation for both alleles (data not shown). Thus, it seems probable that an artifact occurred during fosmid library generation that resulted in the formation of duplicated chimeric products that produce false inversion mapping patterns. These results support the elimination of putative duplicated fosmids as a conservative strategy in PEM inversion prediction based on fosmid libraries.

1.2.2 Artefactual fosmid paired-end detection by misspriming analysis

Among the whole set of GRIAL predictions, we have detected by manual inspection several cases of inversions with support of only 1 BP by a cluster of fosmid paired-ends that share the location of 1 end, which is extremely unlikely to happen by chance. A subset of these fosmids are fully sequenced and publicly available as part of the HGSV project (see URL section). By analysis of whole sequenced fosmids it was found that they mapped 100% concordantly in the human genome and that the end read was actually generated from an internal region of the fosmid. This creates a fictitious inversion signal, although mapping of the entire ~40 kb sequence provides no evidence of a rearrangement occurring in the corresponding genomic region (**Figure 18**).

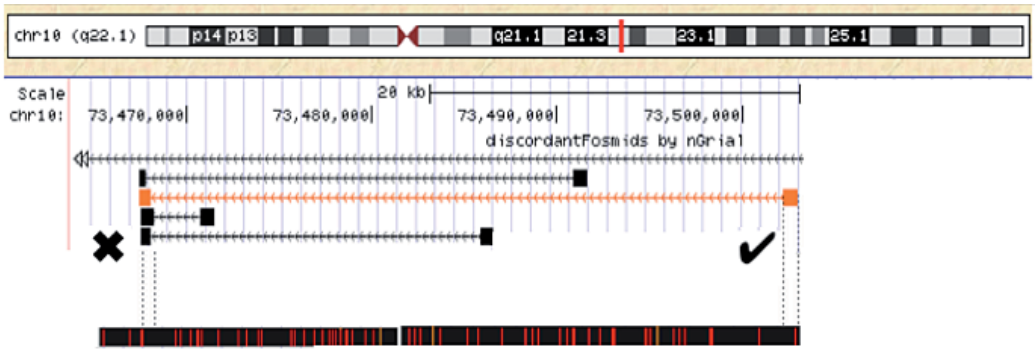


Figure 18 - Artefactual paired-end fosmid mapping – Scheme of a genomic region of chromosome 10 with a cluster of 4 PEMs (boxes linked by arrowed line) showing inversion mapping patterns (ends with discordant orientation), which are indicative of the presence of an inversion in the region. However, all fosmids share the location of 1 end. Whole sequence from a single fosmid (in orange) mapped 100% concordantly (black and red box). Notice that the sequenced end was actually generated from an internal region of the fosmid, not from the real extreme (indicated by a cross). Contrarily, the counterpart paired read belongs to the real end of sequenced fosmid (indicated by a tick).

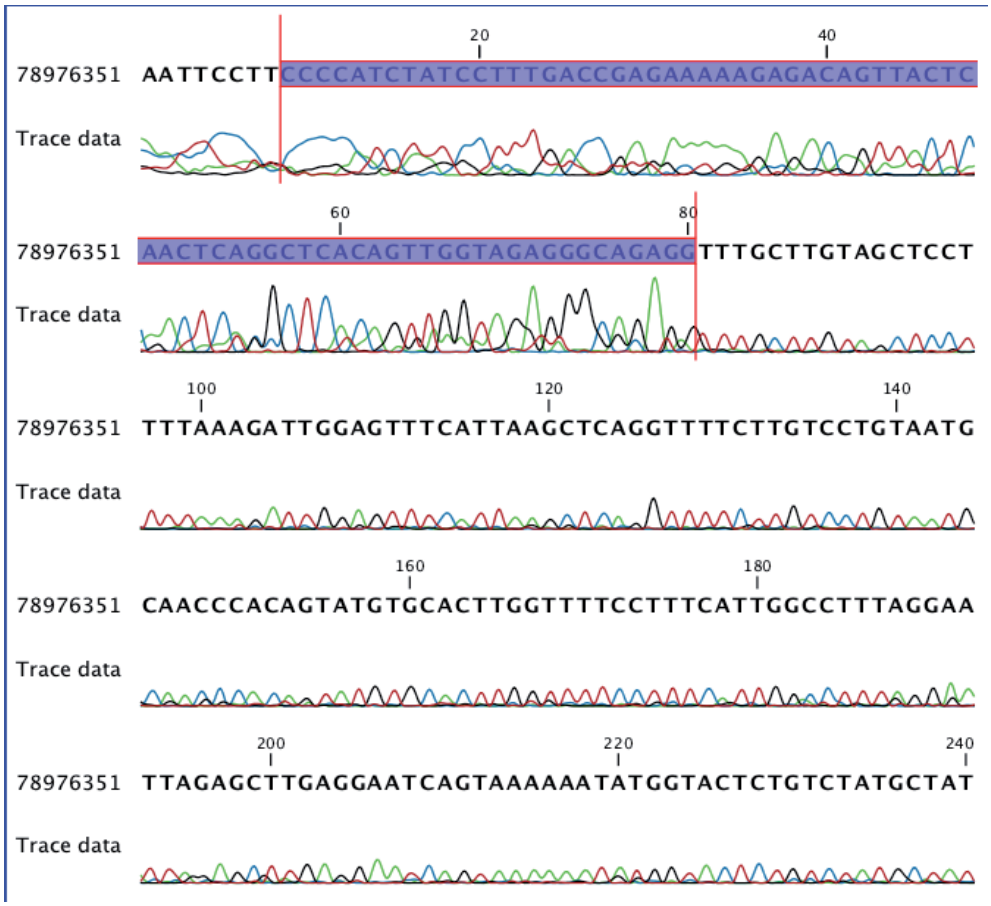


Figure 19 - Problematic fosmid paired-end read sequence – First 240 nucleotides of read 78976351 (TRACE id) belonging to HGSV discordant fosmid ABC10_44554900_B24 are shown. The electropherogram corresponding to the first 80 nucleotides (highlighted in blue) of the read shows conflicting signal that may correspond to 2 sequences coming from different amplifications.

We also observe that the sequence of many of these almost perfectly overlapping reads seems to be an admixture of different amplified products (**Figure 19**). Analysis of the sequence quality of a subset composed by 64 problematic fosmid paired ends shows a clear bias for low quality of the putatively artefactual read set compared to the 64 counterpart reads and to 64 randomly picked reads of fosmids showing no artefactual signal (**Figure 20**). These differences are especially notorious in the 5' extreme of the read sequence (from the 5' first nucleotide up to 10% of total read sequence length), for which median quality of the problematic set is 3 times lower compared to the non-problematic set. This is in agreement with the observed noise in problematic reads electropherograms, as the differences are significant for all tested bins (p-val<0.01, t-test), but particularly for the 5' extreme segment of the read (0-10% bin : p-value = 2.2E-16, t-test).

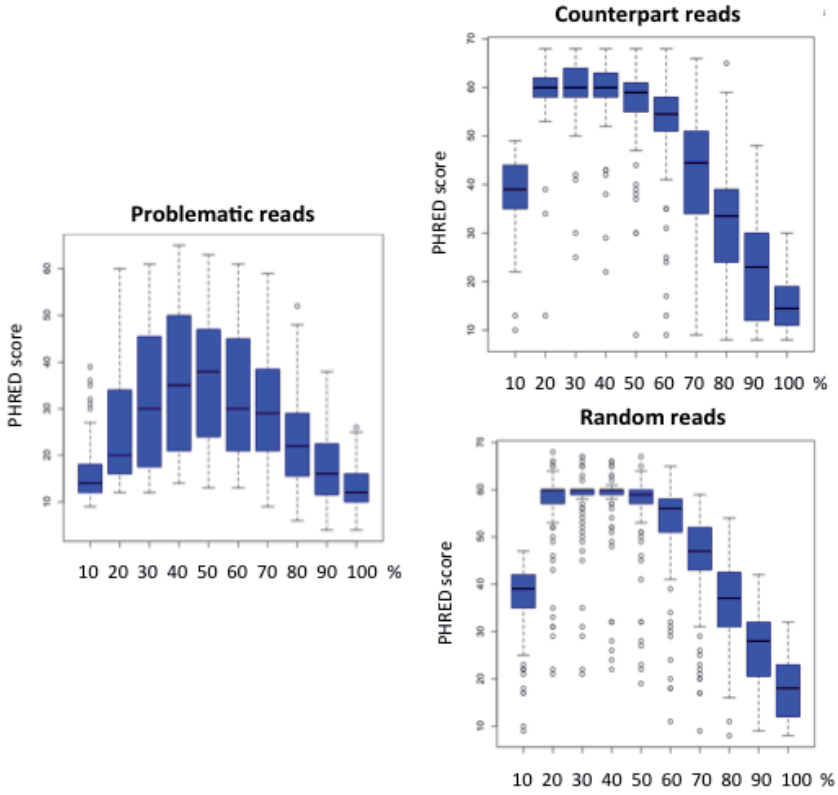


Figure 20 – Sequence quality of problematic reads – Sequence quality (PHRED score) per sequence fragments of 64 problematic reads, its counterpart paired ends and 64 random fosmid paired-end reads. Bins correspond to windows of 10% of the read sequence, labelled by the outer limit. For example, label "10" corresponds to nucleotides contained in the first 10% of the sequence total length, label "20" to nucleotides between 11% and 20% of sequence total length, and so on.

We hypothesized that this phenomenon could be due to primer missprimingh happening in the amplification step during the fosmid library paired-end sequencing, causing artefactual sequence products from the adjacent, leading missprimed fosmid region (**Figure 21**).

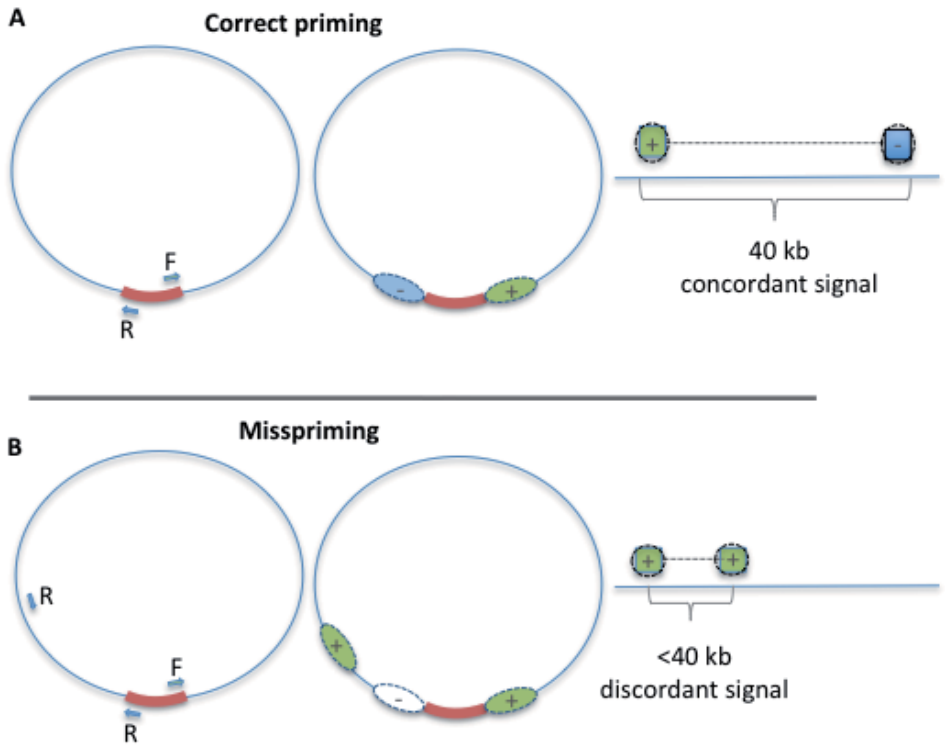


Figure 21 - Fosmid library paired-end sequencing – A) Primers forward (F) and reverse (R) prime correctly to the amplification vector (in red) containing the circularized fosmid (blue line) and in this case, they amplify positive and negative strand, respectively (it could be the opposite way, it depends on how the fosmid circularizes). As the fosmid contains no SV, mapping signature of paired ends is concordant in size (~40 kb) and orientation. B) Primer reverse missprimes to an internal region of the fosmid, generating an artefactual amplification of a fragment on the same strand amplified by primer forward (positive strand, in this case). When the sequenced products are mapped, they produce a mapping signal characteristic of inversions (discordant in orientation). Therefore an inversion can be erroneously predicted despite the fact that the fosmid targets a region without any SV.

To check if this artifact was potentially a source of a great number of inversion false positives predictions, we looked for fosmids with problematic signature and found 3721 instances corresponding to 444 `GRIAL` predictions (see Materials and Methods) and we decided to further explore this issue. First, for each problematic fosmid read, we extracted from the reference genome (HG18) a

nucleotide sequence adjacent to the start coordinate of the read-mapping locus (upstream for reads mapping to the positive strand and downstream from reads mapping to the negative strand). The total length of the extracted sequence equaled to the read length plus 100 bp. Then, we mapped the set of sequencing primers used to generate the paired-end fosmid library sequences by Kidd et al. (2008) and Tuzun et al. (2005) to the extracted genomic regions (see Materials and Methods). We obtained 295 regions with a mapping hit for the primer sequence, affecting 101 inversion predictions. From this set, 88 inversions were supported entirely by fosmids affected by misspriming and 5 additional inversions were supported completely by affected fosmids with the exception of one. These 93 inversion predictions were considered as false positives and were filtered out. In most of the cases the conflicting sequence was generated with the pCC2 reverse primer compared to the pCC2 forward primer (4:1 ratio) with only 4 cases showing a hit for the pCC1 reverse primer. This can be explained by the sequence of the pCC2 reverse primer being more abundant than pCC2 forward genome-wide. Finally, although not relevant in this work, we point out that this type of misspriming can be also responsible for some apparent deletion calls in the fosmid PEM data.

1.2.3 Detection of false positive inversion predictions by remapping of fosmid paired ends

For simplicity, GRIAL predicts inversions from pre-mapped paired ends and does not account for multiple mappings. However, in order to discard false positive predictions, we have done a posterior remapping analysis. Remapped paired ends results allow us to distinguish inversions entirely supported by unambiguously mapped paired ends from the ones presenting ambiguous mapping signal at some extent. We then used this information to filter out dubious predictions and therefore increase the reliability of our inversion catalogue.

To address that, first, we remapped the entire set of fosmid paired-ends supporting any of the 597 GRIAL inversion predictions (the original 690 set filtered for misspriming cases) to HG18 and HG19 human genome reference assemblies. Second, we developed a score system based on the quality and multiplicity of the mapped reads set in order to categorize the paired ends according to their level of reliability for inversion prediction (Materials and Methods). For instance, a discordant in orientation read pair with both ends displaying a good mapping score is categorized as reliable (supporting an inversion event) as long as an alternative

mapping conformation concordant in size and orientation does not exist; if that is not the case, then the paired-ends are labeled as ambiguous and therefore unreliable for inversion prediction. Similarly, a read pair now mapping uniquely, concordantly and with a good mapping score would be also considered not indicative of an inversion prediction. Results indicate that only 73-63% of the fosmids present reliable discordant in orientation signal (classified as DISC+ or DISC++): 5172/7087 in HG18 and 4452/7087 in HG19 (**Figure 22**).

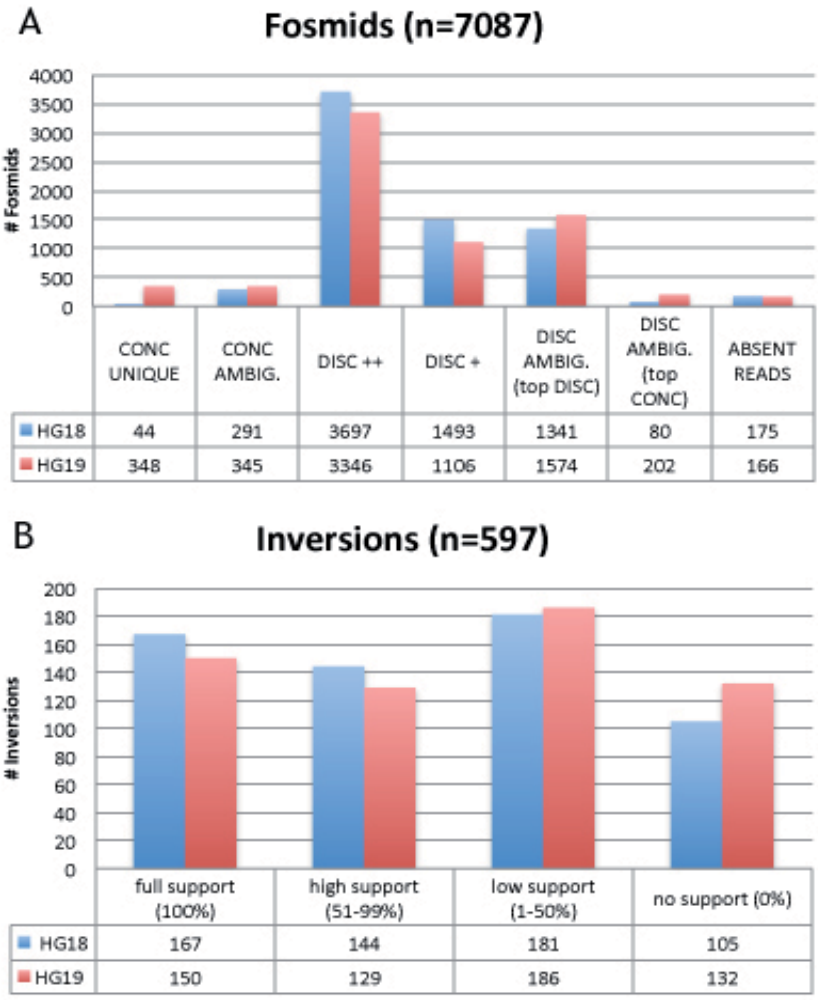


Figure 22 - Remapped discordant in signal fosmids – A) Classification of fosmids into different categories depending on statistics obtained from mapping fosmid PRs to HG18 and HG19. This fosmid set (N = 7087) was used as input for `GRIAL` to generate inversion predictions (Materials and Methods). DISC+ and DISC++ are considered reliable categories for inversion prediction. **B)** Inversion support for reliable discordant in orientation signal. The percentage values correspond to the ratio of the sum of DISC+ and DISC++ fosmids divided by the total number of fosmids supporting the inversion.

Therefore, in HG19 compared to HG18, 10% of discordant signal is lost. The opposite trend occurs for concordant-in-orientation signal: in the latest assembly, HG19, the signal increases (x2) compared to HG18 (291+44 = 335 fosmids in HG18 and 345+348 = 693 fosmids in HG19, CONCORDANT UNIQUE and CONCORDANT AMBIGUOUS categories considered).

In most cases, signal change can be attributed to different and more accurate conformation of the region (gap filling, sequence addition, change of orientation, etc.) in the latest assembly. Therefore, relying on HG19 instead of HG18 fosmid mappings results in a decrease of false positive inversion predictions. However, in some cases, the difference is explained by the replacement of HG18 genomic sequence by a clone corresponding to a different allele in HG19. For instance, for a certain inversion, *Std* conformation is represented in HG18 and *Inv* conformation in HG19 (Aguado et al. 2014). Therefore the discordant signal in HG18, although genuine, is lost in the posterior assembly. Here, we follow a conservative strategy and consider HG19 mapping results to filter out unreliable inversion predictions.

We observe only 150 predictions entirely supported by paired ends showing reliable inversion mapping signal when mapped to HG19, compared to 167 predictions in HG18 (**Figure 22**). The remaining set is supported by ambiguous or unreliable mapping signal to some extent. We detect, according to our scoring strategy, 132 predictions that are not supported by any reliable read pair in HG19 (compared to 105 predictions in HG18), and 17 additional predictions that lose all their support in HG19 assembly with respect to HG18. We also identify 64 inversions that present less support in HG19 compared to HG18. In addition, we detect 186 predictions that present a low degree of reliable paired ends support (**Figure 22**); in all cases the prediction is mostly supported by ambiguous pairs (>50%) and in 103 instances only a single reliable pair supports the inversion. Out of these, 43 correspond to inversions supported by only 2 PRs (1+/1− cases). Altogether, considering the results derived from HG19 assembly, there are 318 predictions (49 in random chromosomes) that could be filtered out or considered unreliable due to the lack of consistent signal. Contrarily, we observe that in 32 cases the predictions increase the reliable/unreliable paired ends ratio in HG19 compared to HG18 indicating that the remapping on the current version of the genome has been useful not only for detecting false positive cases, but also for strengthening confidence on a few inversion predictions.

Some inversion predictions are located in genomic regions that show a different and more complex organization in subsequent releases of the human genome sequence. That is the case for 4 GRIAL inversion predictions merged to 2 INVFEST predictions (HsInv0710 and HsInv0306), analysed in this work by bioinformatical means, experimentally tested by iPCR and described in Aguado et al. (2014). HsInv0710 and HsInv0306 are 2 overlapping inversion predictions of different length supported by many discordant fosmids. This region has been updated with a new sequence patch (GL949743.1), which has 75 kb of extra sequence that transforms the 15-kb inverted SDs found in HG18 and HG19 into 2 SD blocks of 109 kb and 95 kb (**Figure 23**, figure included as part of the supplementary material of Aguado et al. (2014) article). To determine if the 2 inversion candidates are still valid in this context, we re-mapped a set of 1,725 concordant and discordant fosmid paired-end reads with mappings spanning the region of interest +/−50 kb (Material and methods).

We obtained a total of only 20 discordant paired-end reads, and most of the fosmids originally supporting the inversion predictions mapped in highly identical regions within the new SD blocks and were not informative. For HsInv0710, only 1 of the 53 fosmids still supports the inversion, although it maps within the inverted SDs with just slightly higher score in the discordant than the concordant orientation. For HsInv0306, 19 of the 65 fosmids continue to map as discordant in orientation in the HG19 patch (**Figure 23**). However, a similar amount of fosmids from all the individuals also support the reference orientation and the 19 apparently discordant fosmids are explained by a ~16 kb polymorphic deletion of part of SD2 (**Figure 23**). Thus, there is not reliable PEM evidence that these inversions exist. Due to the size of the new SDs it was not possible to interrogate the presence of the inversion by iPCR. Nevertheless, several iPCR primers were designed to confirm the organization of the genomic region (**Figure 23**). For HsInv0710, 6 of the 9 individuals are heterozygous for AB and BD (the other 3 being homozygous for AB) and the fragments AC and CD were never amplified. Similarly, for HsInv0306, the 9 individuals showed AB and AC amplification. These results support the existence of big SDs and indicate that the sequence of the new patch is probably correct.

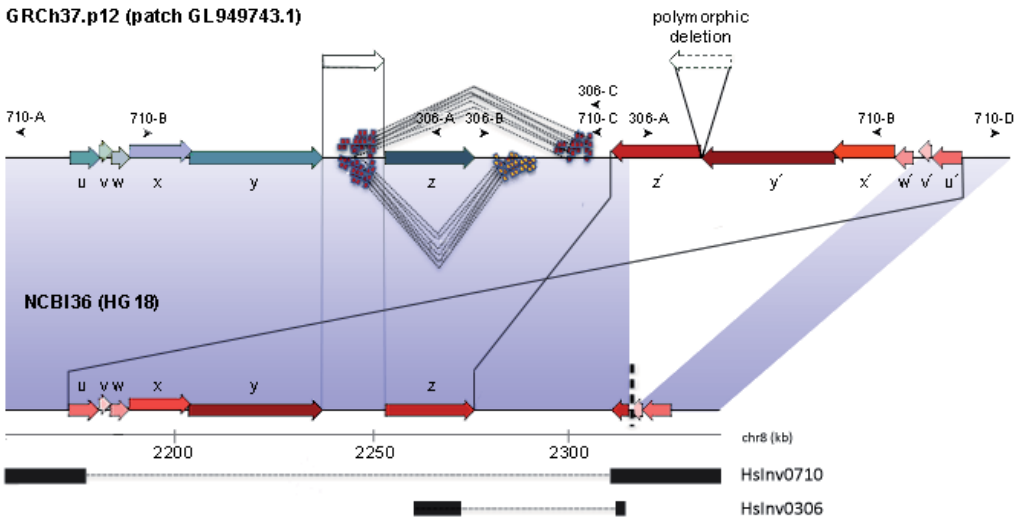


Figure 23 - HsInv0306 and HsInv0710 region - Schematic representation of the HsInv0306 and HsInv0710 inversion prediction region in HG18 (bottom) and GL949743.1 patch (top). Collinear blocks between the 2 sequences are depicted on a purple background. The new inverted duplication in the patch is indicated by solid lines and regions with more than 97% identity between the duplications are labelled as u, v, w, x, y, and z (SD1) and u′, v′, w′, x′, y′, and z′ (SD2). An additional 15.8 kb region that is deleted in SD2 between z′ and y′ in some individuals is represented on top of the diagram. Unique discordant-in-orientation and concordant PRs from the remapping of the fosmid data in the patch are linked by dashed lines, with reads mapping to the negative strand as yellow boxes (concordant PRs, below) and reads mapping to the positive strand as red boxes (discordant-in-orientation PRs, above). HsInv0306 inversion corresponds to original `GRIAL` predictions HsInv0306 and HsInv0312, whereas HsInv0710 inversion corresponds to original `GRIAL` predictions HsInv0710 and HsInv0311 (Martínez-Fundichely et al. 2014). In the remapping analysis, inversion HsInv0306 is supported by 19 unambiguously discordant fosmid paired-end reads, but this mapping profile is compatible with the polymorphic deletion of a genomic fragment between duplications z′ and y′, which causes that the end reads that should map concordantly within this region map within SD1 instead. The existence of this polymorphic indel was confirmed by the analysis of the mapping distance of the fosmid ends across this region (with only 3 individuals having fosmids consistent with the deleted form of SD2) and additional available human BAC sequences (AC245519 and AC245187, both including the SD2 extra sequence). According to this scenario, the presence of HsInv0306 and HsInv0710 is not supported anymore on the basis of the PEM data.

1.3 `GRIAL` inversion candidates validation

In this section, using PCR we aim to validate a subset of `GRIAL` inversion predictions (Material and Methods). We have prioritized the selection of inversions uniquely predicted by `GRIAL` and also inversions predicted by clustering 2 fosmid paired ends showing characteristic inversion mapping signal, with 1 paired-end supporting each BP (1+1− cases). We hypothesize that although lowly supported, some 1+1− predictions may be real. Thus, we have chosen to validate 2 predictions with clear, unambiguous discordant signal that have passed all filters previously described. Both inversions are small (<3.5 kb), which may explain the lack of

supporting signal: due the big insert size of the fosmid library (~40 kb) it is improbable to capture small inversions (Lucas Lledó and Cáceres 2013).

1.3.1 HsInv0409

HsInv409 is a small (~1.3 kb) inversion localized in an intronic region of gene *NLGN4X* in chromosome X (**Figure 24**). The inversion was first identified in a Korean individual (Ahn et al. 2009) by means of PEM, and is extensively described in section IV3.3. As previously mentioned, HsInv0409 is part of the uniquely predicted inversion set by `GRIAL`. This is because other benchmarking tested PEM based methods (`VH`, `PEMer`) did not predict it due to insufficient signal as they require more than 1 fosmid supporting each inversion BP.

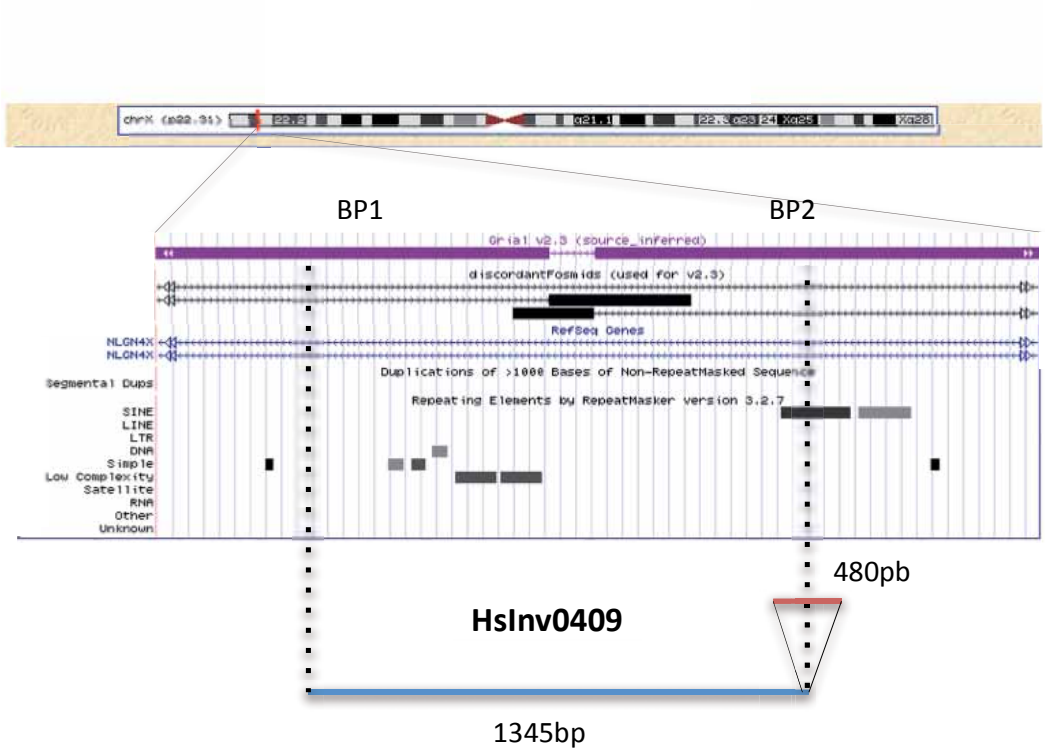
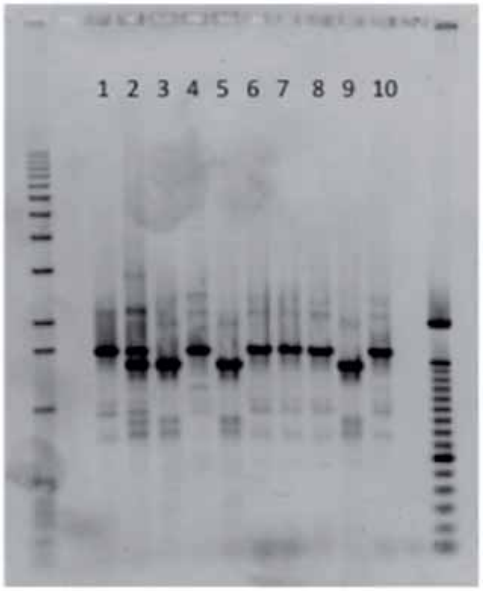


Figure 24 - HsInv0409 inversion region – `GRIAL` predicted inverted region in purple with BP regions schematized by purple boxes. Below, discordant fosmids are represented in black with ends schematized by black boxes. The exact inverted region (blue line) spans 1,345 bp and is flanked by a 480 bp sequence not present in the reference genome (red line). SVs < 10 bp (8 bp deletion, 6 bp duplication) not shown. We observe that HsInv0409 inversion locates in *NLGN4X* intronic region.

The inversion was experimentally tested in a panel of 10 individuals (**Figure 25**) by PCR (see Material and Methods). Results validate the inversion candidate region and show that 6 samples are homozygous for the standard rearrangement (2 Asians, 2 Africans and 2 Europeans), 1 European is heterozygous and 3 are homozygous for the inverted rearrangement (2 Africans, 1 European) (**Figure 25**).



Lane	Genotype	Sample	Pop	HGSV id
1	STD/STD	NA12156	CEU	ABC14
2	INV/STD	NA12878	CEU	ABC12
3	INV/INV	NA15510	CEU	WI2 (G248)
4	STD/STD	NA18507	YRI	ABC8
5	INV/INV	NA18517	YRI	ABC7
6	STD/STD	NA18555	CHB	ABC11
7	STD/STD	NA18956	JPT	ABC9
8	STD/STD	NA19129	YRI	ABC13
9	INV/INV	NA19240	YRI	ABC10
10	STD/STD	HuRef	EUR	

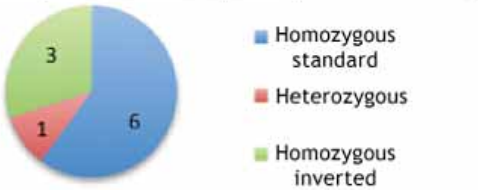


Figure 25 - HsInv0409 validation – Results of HsInv0409 multiplex PCR. From the 10 samples tested, 6 are homozygous for the standard rearrangement (2 Asians, 2 Africans, 2 Europeans), 1 European is heterozygous and 3 are inverted homozygous (2 Africans, 1 European). Lateral lanes correspond to PCR DNA ladders. Sample ids correspond to HapMap ids.

Then, to characterize HsInv0409 BP regions at a nucleotide level, the amplification product corresponding to the inverted allele of sample NA12878 was purified and sequenced. Mapping of the sequenced fragment shows that HsInv0409 spans 1,345 bp and is flanked by 2 small sequences (8 bp and 480 bp) not present in the reference genome as well as a 6 bp duplication (**Figure 24**). Microhomology signal indicates that the inversion was probably generated by several FoSTeS events or NHEJ followed by a deletion. Alignment of the region with chimpanzee, gorilla and orangutan genomes indicates that the inverted allele is ancestral. Although no genotyping of HsInv0409 has been performed in primates, the inverted

rearrangement is not expected to be polymorphic in these species as the lack of IRs at the BP regions suggests a unique, non-recurrent origin of the inversion in the human lineage.

This inversion was subsequently genotyped in a bigger sample set and its functional impact on gene expression and association to disease have been explored in this work (section IV3.3)

1.3.2 HsInv0410

HsInv0410 is a relatively small (~3.3 kb) inversion also localized in an intergenic region of chromosome X, with BP regions contained in 2 IRs spanning 1.7-1.9 kbs. To our knowledge, this inversion has been first identified in our lab by means of PEM. As in the HsInv0409 case, the inversion is predicted by `GRIAL` by clustering of 2 paired ends, 1 supporting each inversion BP and it has been missed by the other PEM methods for this reason. In addition, there is another fosmid with PEM signal discordant in orientation present in the inverted region although it has not been included in HsInv0410 `GRIAL` prediction (**Figure 26**). However, 2 of these fosmid paired ends map to the IR in inversion rightmost BP. Hence, the mapping signal is ambiguous and not completely reliable. Moreover, there are many other fosmid paired ends indicating the possible presence of further structural variants in the region that are probably independent of the inverted rearrangement.

PCR results validate the inversion region and confirm the presence of the inverted allele in Asian HapMap sample NA18956 in heterozygosis (**Figure 27**). However, PCR results for the other 2 samples in which discordant fosmids targeting the inversion region were found (NA19129, NA12156) show no evidence for the presence of HsInv0410 inverted allele (*Std* conformation appears to be in homozygosis). In addition, the PCR failed for individual NA18555.

PCR conditions for HsInv0410 genotyping have been improved and the inverted region is currently being reanalysed in the same sample set; results corroborate observed genotypes and show that sample NA18555 is homozygous for the standard rearrangement (D. Izquierdo, personal communication). However, due to the lack of an appropriate high-throughput genotyping protocol, no further analysis has been performed for this inversion.

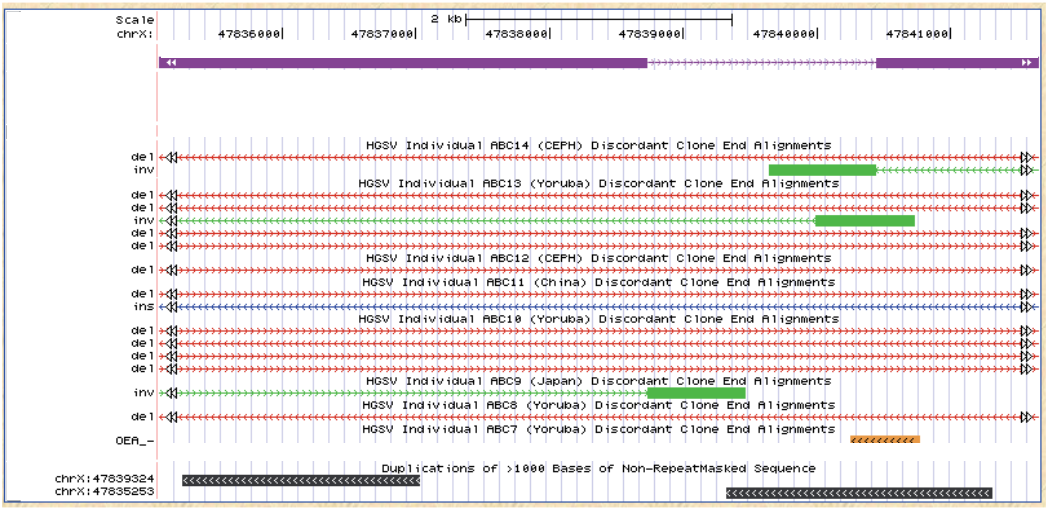


Figure 26 - HsInv0410 inversion region – Scheme of HG19 chrX:47,838,700-47,843,540 region is shown. HsInv0410 inverted rearrangement predicted by `GRIAL` in displayed in purple with BP regions schematized by purple boxes. Below, discordant fosmids showing inversion mapping signal are labelled as "inv" and are represented in green with ends schematized by green boxes. Notice that 2 of these paired ends map to an IR (black boxes). Fosmids showing deletion ("del") and insertion ("ins") mapping signal are represented in red and blue, respectively, and fosmids with missing ends ("OEA") are represented in orange.

Lane	Genotype	Sample	Pop	HGSV id
1	STD/STD	NA12156	CEU	ABC14
2	STD/STD	NA12878	CEU	ABC12
3	STD/STD	NA15510	CEU	WI2 (G248)
4	STD/STD	NA18507	YRI	ABC8
5	STD/STD	NA18517	YRI	ABC7
6	ND	NA18555	CHB	ABC11
7	STD/INV	NA18956	JPT	ABC9
8	STD/STD	NA19129	YRI	ABC13
9	STD/STD	NA19240	YRI	ABC10

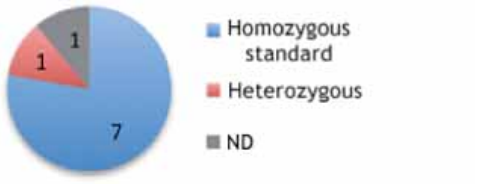


Figure 27– HsInv0410 validation – Results of HsInv0410 multiplex PCR. From the 9 samples tested, 7 are homozygous for the standard rearrangement (4 Africans, 3 Europeans) and 1 Asian is heterozygous. PCR failed for sample NA1855. Lateral lane corresponds to PCR DNA ladder.

Chapter 2

Functional impact of polymorphic inversions on gene expressions

SUMMARY – In this chapter, we have performed a global exploratory functional analysis with the objective to reveal associations of inversions with changes in gene expression in *cis* and *trans*. To achieve this goal, we used previously obtained genotypes of a set of 44 polymorphic inversions from a total of 551 samples from 7 HapMap populations of European, Asian and African ancestry. Then we looked for positive and negative correlations between the presence of inverted alleles and gene expression levels by means of two different approaches in a wide spectrum of different tissues. First, we performed a linear regression analysis in three different expression datasets derived from LCL cell lines (LCL DE analysis). Second, we interrogated blood and non-blood tissues by contrasting expression quantitative trait loci (eQTL) data with inversion tag-SNPs (inversion-eQTL analysis). The results show that a few inversions affect gene expression in *cis* and *trans* both by positional effects or linkage to associated haplotypes.

2.1 Selection of candidate polymorphic inversions and predicted functional effects

To interrogate the functional impact of a collection of human polymorphic inversions located throughout the genome, as part of the INVFEST project a list of candidates from the filtered and refined catalogue of predictions of inverted regions generated by GRIAL were selected (see previous chapter). The criteria applied for the selection was based both on feasibility of genotyping (as inversions with very big (>25 kb) repetitive regions at the BPs present technical difficulties to be genotyped in a high-throughput manner) and also by potential functional impact. Thus, we favored the selection of inversions adjacent to or breaking genic sequences compared to others located in intergenic regions. The final set is composed of 44 inversions ranging from ~140 bp to ~ 415 kb (**Figure 28**) with available genotyping data. While ~41% of the inversions present either entirely clean BPs (with absence of repetitive elements) or small microhomology signatures, the remaining set is mediated by inverted repeats (IRs) formed by repetitive elements or segmental duplications spanning from ~255 bp to ~ 32.2 kb (**Table 7**).

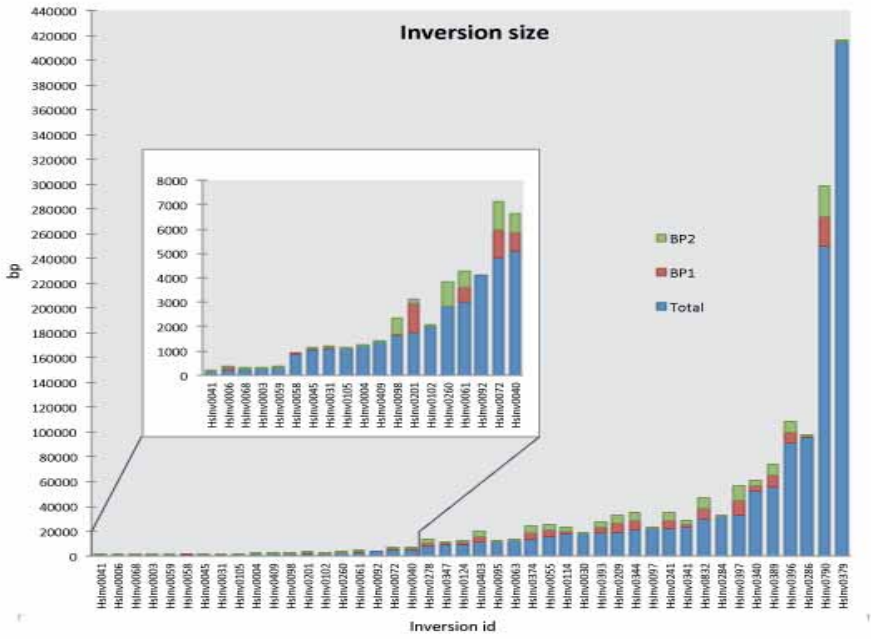


Figure 28 – Inversion and BP size distribution – The total length of the 44 inverted regions is displayed, including the length of the leftmost (BP1) and rightmost (BP2) BP predicted region. Inversion identifiers (ids) correspond to those of invFEST (Martínez-Fundichely et al. 2013).

Inversion id	Chr.	Size (bp)	IR1 / IR2 size (kb)	Inverted allele freq.	Derived allele freq.	Ancestral conformation
HsInv0003	1	255	-	82.40	82.40	Std
HsInv0004	1	1197	-	11.55	88.45	Inv[4]
HsInv0006[1]	1	157	-	59.17	40.83	Inv
HsInv0040	2	4346	754 / 755	77.28	22.72	Inv
HsInv0041	2	140	-	50.27	50.27	Std
HsInv0241[1]	2	16232	13283 / 12836	16.33	16.33	Std[5*]
HsInv0260	2	2335	-	17.06	17.06	Std
HsInv0097	3	21760	-	0.54	0.54	Std
HsInv0098	3	1308	296 / 296	17.51	17.51	Std
HsInv0095	4	11536	-	78.13	21.87	Inv[4]
HsInv0102	4	2033	-	19.24	19.24	Std
HsInv0055	5	11135	5998 / 6001	64.70	ND	ND
HsInv0201	5	1059	-	56.62	56.62	Std
HsInv0278	5	5083	2830 / 2791	25.32	74.68	Inv[5]
HsInv0058	6	876	-	66.00	34.00	Inv
HsInv0059	6	316	-	76.86	23.14	Inv
HsInv0061	6	2330	653 / 655	98.81	ND	ND
HsInv0092	6	4111	-	16.97	16.97	Std
HsInv0284	6	32390	-	3.45	3.45	Std
HsInv0105	7	1095	-	51.09	48.91	Inv
HsInv0286[1]	7	95055	15887 / 15878	48.53	48.53	Std[5]
HsInv1053	7	15308	-	53.09	53.09	Std[4, 6]
HsInv0068	9	253	-	86.91	13.09	Inv
HsInv0114	9	14667	2733 / 2733	51.64	48.36	Inv[5]
HsInv0124	11	7721	4516 / 8882	71.39	71.39	Std[5]
HsInv0209	11	11935	7028 / 7027	9.41	9.41	Std[5]
HsInv0340	13	48350	10585 / 10585	16.35	83.65	Inv[5*]
HsInv0341	13	20831	6169 / 6158	8.00	8.00	Std[5*]
HsInv0344	14	14409	7189 / 7201	44.39	ND	ND[5*]
HsInv0347	14	7138	1588 / 1588	19.71	19.71	Std[5*]
HsInv0030	16	17120	1547 / 1545	93.10	6.90	Inv[4]
HsInv0031	16	1065	296 / 288	62.41	37.59	Inv[6]
HsInv0374	17	8865	5752 / 5760	46.62	ND	ND
HsInv1051	17	225827	24247 / 24322	2.31	2.31	Std[5]
HsInv0379	19	415128	-	0.45	0.45	Std
HsInv0045	21	994	255 / 256	51.47	51.47	Std
HsInv0072	X	3879	1447 / 1443	97.35	97.35	Std
HsInv0389[3]	X	46822	11328 / 11335	51.55	48.45	Inv[5]
HsInv0393	X	14059	4689 / 4689	44.24	55.76	Inv[5*]
HsInv0396	X	81700	9495 / 9478	20.35	20.35	Std[5*]
HsInv0397	X	21499	32159 / 15096	38.96	61.04	Inv[5*]
HsInv0403	X	7212	4323 / 4329	47.44	ND	ND[*]
HsInv0409	X	1345	-	49.27	50.73	Inv
HsInv0832[2]	Y	40248	8722 / 8725	44.24	ND	ND

Table 7 - Summary of characteristics of the inversion set – Inversion identifier (Inversion id) corresponds to `InvFEST` id (Martinez-Fundichely et al. 2014). Inversions were genotyped by MLPA/iMLPA (S. Villatoro and M.Cáceres, unpublished results), unless specified otherwise. Inversion size is calculated as the distance between the two BP intervals midpoints. Inverted allele is defined with respect to the conformation present in HG19, referred as standard. The ancestral allele has been inferred by alignment of the inverted region with chimp, gorilla, orang and macaque genome, unless specified otherwise.[1]HsInv0006 has been genotyped by conventional PCR in 90 (60, excluding childs of trios) HapMap samples from CEU population (D. Vicente and M.Cáceres, unpublished results), whereas HsInv0241 and HsInv0286 have been successfully genotyped by iPCR in 77 and 55 samples mostly but not only from CEU population, respectively (Aguado et al. 2014). [2]HsInv0832 showed different orientations in chimpanzees and gorillas (Aguado et al. 2013). [3]HsInv0389 showed different orientations in orangutans compared to other primates (Cáceres et al. 2007). [4]Ancestral allele determined in Pang et al. (2013) [5]Ancestral allele determined in Aguado et al. (2014). [6]Ancestral allele determined in Feuk et al. (2005). *Inverted allele is polymorphic in some primates (Aguado et al. 2014)

Genotypes of 41 inversions were obtained by an MPLA derived protocol in 551 samples from 7 HapMap populations of European (CEU, TSI), Asian (GIH, CHB, JPT) and African (YRI, LWK) ancestry with genotypes consistent with Hardy-Weinberg equilibrium, concordant inheritance patterns checked in 60 family trios and also 99.8% concordance with previous iPCR based genotyping results (S. Villatoro and M.Cáceres, unpublished results). Regarding the remaining 3 inversions, one was genotyped by conventional PCR in 90 HapMap samples from CEU population (D. Vicente and M.Cáceres, unpublished results), and 2 were successfully genotyped by iPCR in 77 and 55 samples mostly but not only from CEU population (Aguado et al. 2014).

To investigate the possible functional consequences of the inversions and to have a rough estimate of the number of genes potentially affected in *cis*, we analysed the gene content (Ensembl v.75 genes, see URLs) of the inverted regions. For this analysis, we have considered all transcripts predicted in Ensembl v.75, independently of their support, including not only experimentally validated but also predicted isoforms, as well as pseudogenes (**Table 29**, Appendix). We observe that 23/44 inversions are located outside gene domains whereas the remaining set (21/44) overlaps with genes in different manners (**Table 9**), with predicted functional effects ranging from null to severe. For instance, 9 inversions contain complete genes and therefore the inversion changes the original gene orientation without modifying the genic sequence (**Table 8**). In total, 39 genes are entirely contained by inversions.

Inversion id	Gene	Type
HsInv0124	***IFITM1***	protein coding ***
HsInv0209	*KRTAP5-14P*	pseudogene
HsInv0241	*AC011298.1*	protein coding **
	AC011298.2	lincRNA **
HsInv0278	*FOXO1B*	pseudogene
HsInv0340	***OR7E156P***	pseudogene
	AL445989.1	protein coding **
	RP11-473M10.3	lincRNA **
HsInv0374	*SH3GL1P2*	pseudogene
HsInv0379	***RP11-420K14.2***	pseudogene
	AC092364.4	miRNA [NA]
	RP11-420K14.3	pseudogene
	VN1R84P	pseudogene
	ZNF100	protein coding ***
	RP11-420K14.6	pseudogene

Inversion id	Gene	Type
HsInv0379	*AC092364.2*	miRNA [NA]
	RP11-420K14.8	lincRNA *
	RP11-420K14.7	pseudogene
	ZNF43	protein coding ***
	CTD-2607J13.1	pseudogene
	ZNF208	protein coding ***
	AC003973.1	pseudogene
	AC003973.6	pseudogene
	AC003973.4	lincRNA **
	AC003973.5	lincRNA *
HsInv0389	***FLNA***	protein coding ***
	EMD	protein coding ***
HsInv1051	***TBC1D28***	protein coding ***
	RP11-815I9.3	pseudogene
	AC026271.5	pseudogene
	ZNF286B	protein coding ***
	RP11-815I9.4	lincRNA [NA]
	FOXO3B	pseudogene
	UBE2SP2	pseudogene
	TRIM16L	protein coding ***
	RP11-815I9.5	pseudogene
	FBXW10	protein coding ***
	TVP23B	protein coding ***
	CTD-2145A24.3	lincRNA [NA]

Table 8 – Genes contained in inversions - Inversion identifier (Inversion id) corresponds to `InvFEST` id (Martínez-Fundichely et al. 2013). Gene ids and annotation (Type) obtained from Ensembl v.75 and GENCODE. The asterisk marks refer to the reliability of the gene, according to GENCODE analysis. For RNA genes, one asterisk indicates that the only support is from a single EST, the best supporting EST is suspicious to be an artefact or no single transcript supports the model structure. Two asterisks indicate that the best supporting mRNA is flagged as suspect, is supported by multiple ESTs or all splice junctions of the transcript are supported by at least one non-suspect mRNA. NA indicates that the reliability of transcript has not been analysed. For protein-coding genes, one asterisk indicates a predicted putative protein, with no experimental evidences supporting the prediction. Two asterisks indicate that the protein is supported at a transcript level. Three asterisks indicate that the protein is supported at a protein level. In bold, genes with average expression > 0.5 FPKMs based on 27 tissues expression data (Fargerberg et al. 2013).

On the other hand, a considerable percentage of the inversions (14/44, 32%) clearly affect genic sequences: 8 inversions affect introns whereas the other 6 invert gene exons (**Table 9**). In addition, some inversions can potentially affect more than one gene. Although the ratio of genes overlapped by inversion BPs is 1:1 for 10 inversions, another 11 overlap with multiple genes with the extreme case of HsInv0344, for which BPs are located in duplicated regions and overlap with at least 6 annotated genes. In 9 instances, the inversion BPs overlap with genes situated partially or totally within IRs. In some of these cases, due to the high degree of

sequence identity between duplicated regions or to the lack of inversion BP resolution, the functional impact of inversion on the paralogous genes should be null or remains unclear (**Table 9**). From the gene viewpoint, 42 genes overlap with inversion BPs: mainly protein-coding genes (22/42, 52.3%) but also non-coding genes (14/42, 33.3%) and pseudogenes (6/42, 14.29%). However, only 6 out of these 42 genes (4 protein-coding, 1 pseudogene and 1 lincRNA) are affected by inversions on the exonic region.

Inversion id	Gene	Type	Effect
HsInv0006	***DSTYK***	protein_coding ***	inversion in intron
HsInv0030	***CTRB1***	protein coding ***	overlapping BP
	CTRB2	protein coding ***	overlapping BP
HsInv0059	***GABRR1***	protein_coding ***	inversion in intron
HsInv0061	***RP1-60O19.1***	lincRNA **	inversion in intron
HsInv0098	***ULK4***	protein_coding ***	inversion in intron
HsInv0102	***RHOH***	protein coding ***	inverted internal exon
HsInv0105	***C7orf10***	protein coding ***	inversion in intron
HsInv0124	***IFITM2***	protein coding ***	inverted 3' exon
	IFITM3	protein coding ***	overlapping BP
	RP11-326C3.7	antisense *	overlapping BP
	RP11-326C3.11	antisense *	overlapping BP
HsInv0201	***SPINK14***	protein coding ***	deleted exon
HsInv0209	*KRTAP5-10*	protein coding ***	overlapping BP
	KRTAP5-11	protein coding ***	overlapping BP
	AP000867.14	pseudogene	overlapping BP
	AP000867.1	protein coding *	overlapping BP
HsInv0241	***AQP12A***	protein coding ***	overlapping BP
	AQP12B	protein coding ***	overlapping BP
HsInv0278	*TRNA_Val*	tRNA	overlapping BP
	TRNA_Leu	tRNA	overlapping BP
HsInv0340	***LINC00395***	lincRNA*	inverted 5' exons
HsInv0344	*RP11-671J11.7*	processed transcript **	overlapping BP
	RP11-671J11.6	lincRNA **	overlapping BP
	RNU1-27P	snRNA [NA]	overlapping BP
	RNU1-28P	snRNA [NA]	overlapping BP
	RP11-671J11.4	antisense **	overlapping BP
	SNX6	protein coding ***	overlapping BP
HsInv0374	***AC005562.1***	processed_transcript [NA]	inversion in intron
	LRRC37BP1	pseudogene	overlapping BP
HsInv0379	***ZNF257***	protein coding ***	inverted 5' exons
	RP11-420K14.1	pseudogene	inversion in intron
HsInv0389	***RPL10***	protein_coding **	overlapping BP

Inversion id	Gene	Type	Effect
HsInv0393	***RP4-545K15.3***	pseudogene	overlapping BP
	ARMCX6	protein_coding ***	overlapping BP
HsInv0396	*RP11-493K23.1*	antisense *	overlapping BP
	RP11-493K23.4	antisense *	overlapping BP
	PABPC1L2B	protein_coding ***	overlapping BP
	PABPC1L2A	protein_coding ***	overlapping BP
HsInv0409	***NLGN4X***	protein_coding ***	inversion in intron
HsInv1051	***CCDC144B***	pseudogene	inverted 5' exons
	PRPSAP2	protein coding ***	inverted 5' exons
	AC107982.4	pseudogene	overlapping BP

Table 9 – Genes overlapping inversion BPs - Inversion identifier (Inversion id) corresponds to InvFEST id (Martínez-Fundichely et al. 2013). Gene ids and annotation (Type) obtained from Ensembl v.75 and GENCODE. The asterisk marks refer to the reliability of the gene. For RNA genes, one asterisk indicates that the only support is from a single EST, the best supporting EST is suspicious to be an artefact or no single transcript supports the model structure. Two asterisks indicate that the best supporting mRNA is flagged as suspect, the support is from multiple ESTs or all splice junctions of the transcript are supported by at least one non-suspect mRNA. NA indicates that the reliability of transcript has not been analysed. For protein genes, one asterisk indicates a predicted putative protein, with no experimental evidences supporting the prediction. Two asterisks indicate that the protein is supported at a transcript level. Three asterisks indicate that the protein is supported at a protein level. In bold, genes that with average expression > 0.5 FPKMs in (Fargerberg et al. 2013).

2.2 LCL differential expression analysis

Three different expression datasets (**Figure 29**) have been analysed to identify associations of inversion genotypes with changes in gene expression in LCLs. Two of them are based on Illumina arrays (Stranger et al. 2007, 2012) whereas the third one was generated by RNA-Seq (Lappalainen et al. 2013). The commonality of all them is that the expression data derives from transformed LCLs from HapMap individuals. For each expression dataset, we selected the individuals for whom the 4 inversions were genotyped and also a set of CEU individuals with HsInv0241, HsInv0286, and HsInv0006 inversions genotyped (77, 55 and 90 samples, respectively). In total, the selected sample set consists on 527 different individuals of 7 HapMap populations with African, European and Asian ancestry. Interestingly, there is a considerable degree of overlap between these datasets (**Figure 29**) as 72 samples are common in the three studies and 212 are shared in at least two studies. This overlap allows us to assess the robustness of our findings and to assess if the same effects are found in independent studies.

Dataset	Platform	HapMap Populations	Tissue	# genes*	# samples**
Stranger 2007	Illumina Human WG 6 v1	CEU, CHB, JPT, YRI	LCL	19565	270
Stranger 2012	Illumina Human WG 6 v2	CEU, CHB, JPT, YRI, LWK, GIH	LCL	19573	266
Geuvadis	RNASeq	CEU, TSI, YRI	LCL	14950	175

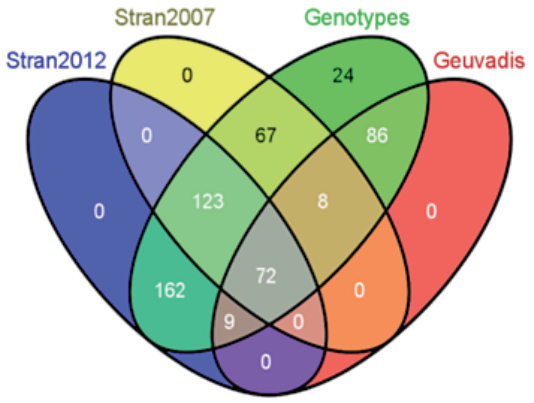


Figure 29 - LCL expression datasets - Gene expression datasets summary is shown. *Genes analysed after filtering (see Materials and Methods). **Samples analysed after filtering (see Materials and Methods).

We are interested in determining the influence of inversions on gene expression. Therefore, for each inversion, we built linear models in which gene expression is explained by inversion genotype and tested the statistical significance of the variable (see Materials and Methods). Although inversion genotype is the primary variable of interest, we incorporated in the model additional known geographic and biological factors such as the gender and the population of the individuals that can also account for variance in gene expression. However, it is known that major components of gene expression variability can often be attributed to technical factors (e.g. batch effects) that introduce unwanted, systematic variability in data (Leek and Storey 2007) leading to spurious correlations among genes or samples and ultimately resulting in both false positive (Kang et al. 2008; Listgarten, Kadie, and Heckerman 2010) and false negative associations (Stegle et al. 2012).

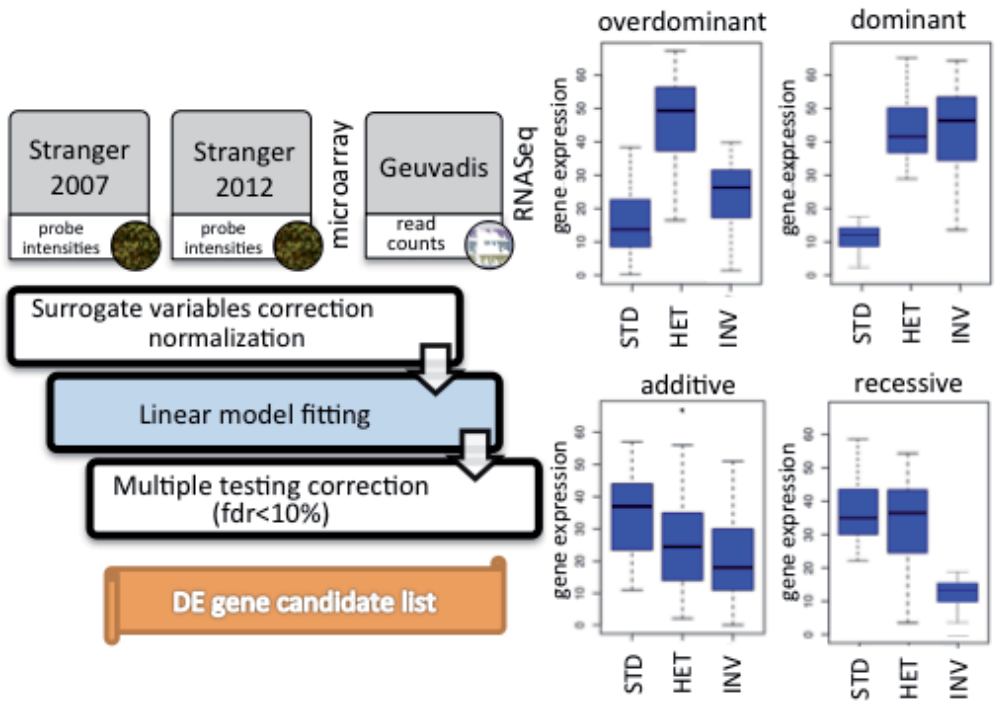


Figure 30 - LCL DE pipeline - Scheme of the different steps performed in the inversion DE analysis on LCL expression datasets. On the right, different inversion inheritance models fitted to the gene expression values in the linear fitting model step. Each model represents a different comparison: overdominant model compares inversion heterozygotes versus a pool of standard and inverted homozygotes, dominant model compares standard homozygotes versus a pool of inversion heterozygotes and homozygotes, recessive model compares inverted homozygotes versus a pool of inversion heterozygotes and standard homozygotes, and the additive model tests the additive effect of inverted allele.

Thus, we identified surrogate variables accounting for the systematic unwanted variability in both RNA-Seq and microarray expression data and incorporated them in the linear model (Materials and Methods). Although there is substantial evidence indicating that the majority of genetic determinants of gene expression act in a strictly additive fashion (Veyrieras et al. 2008), we aim to test alternative scenarios and look for variants acting not only in an additive fashion but also in an overdominant, dominant and recessive fashion. Hence, depending on the genetic inheritance model that we want to test, we modeled the inversion genotype differentially (**Figure 30**).

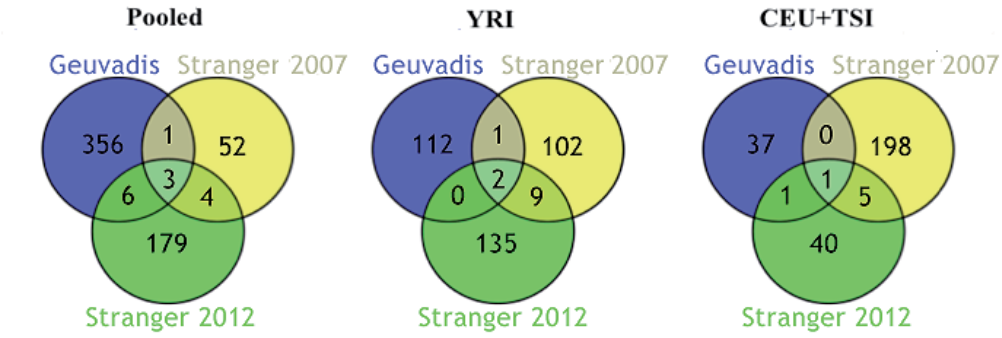
In addition, we aimed to look for eQTLs within and across different populations, so we performed the DE analysis pooling all available samples (population-unspecific) and for each population separately (population-specific) (Materials and Methods).

First, we detect that different studies diverge in the number of identified candidate genes with inversion effects in expression (**Table 10**) and it seems that this figure does not correlate with the sample size of the expression dataset: Stranger 2012 dataset, with a sample size of N = 266, reports 909 genes in total, Stranger 2007 (N = 270) reports 267 genes in total and Geuvadis (N = 175) reports 497. We also observed that the level of redundancy between populations is low: only 6-19% (depending on the expression dataset) of the reported genes are found in more than one population and this may indicate that there is a high number of false positives or that most of the DE genes are population-specific.

Dataset	Pooled	Africans	Europeans	East Asians	West Asians	TOTAL	NR
Geuvadis	457	116	39	-	-	612	497
Stranger 2007	60	115	47	66	-	288	267
Stranger 2012	194	314	204	101	160	973	909

Table 10 – Candidate inversion DE genes per population – Total number (summatory of all inversion results) of DE genes with respect to inversion genotypes per population per expression dataset. Pooled: all individuals analysed together; in the rest population factor included in the linear model and DE genes in each population are shown. NR: non-redundant reported genes across all populations. East Asians: CHB, JPT HapMap populations. West Asians: GIH HapMap population.

To further explore this issue, we identified the genes commonly reported in the different expression datasets across different populations (**Figure 31**). As Asian populations are missing in Geuvadis dataset, we only performed the comparison with European and African populations. We observe that very few candidate DE genes are shared by the different datasets and this may indicate a high false positive rate in the results. We also compared our results for the Geuvadis dataset with an alternative analysis on the same dataset (Lorena Pantano, personal comunication) also aiming to find DE genes associated to inversion genotypes. Again, we find a poor overlap: only 16 genes for 8 inversions are found in common (HsInv0347: *ENSG00000267998*; Hsnv0003: *ZAP70*, *PHYHIP*, *DYRK4*, *FBXO44*, *MB21D1*, *LAPTM4B*; HsInv0124: *IFITM3;* Inv209: *TFPI2*, *GNGT2;* Inv102: *LHX2*, *GPM6A*, *CD44;* HsInv92: *PPM1H;* HsInv58: *HCG22*; HsInv374: *KLHL14*). Among all, there are only two associations in *cis* (HsInv58 and *HCG22* and HsInv0124 and *IFITM3*), which are also the only ones that have been reported in the microarray-derived datasets Stranger 2007 and Stranger 2012.



	Common	Stranger2012 Stranger2007	Stranger2012 Geuvadis	Stranger2007 Geuvadis
Pooled	HsInv0284: *RASIP1* HsInv0058: *HCG27,HCG22*	HsInv0041:*CUL3* HsInv0284:*ADHFE1* HsInv0389:*EMD* HsInv0072:*C21orf33*	HsInv0003: *PTGS2* HsInv0030: *ZNF185,KCNK13* HsInv0340: *TALDO1* HsInv0058: *TCF19* HsInv0389: *PLXNA3*	HsInv0124: *IFITM2*
YRI	HsInv0063: *AUTS2* HsInv0058: *HCG22*	HsInv0284: *WDR66,RASIP1* HsInv0058:*HCG27* HsInv0393:*TCEAL2* HsInv0072:*C21orf33* HsInv0340: *EFCAB4A,IRF7, TSPAN4,TMEM80*	-	HsInv396: *PHKA1*
CEU TSI	HsInv058: *HCG22*	HsInv0209: *UTS2* HsInv0284: *VNN1,ALDH1A1, MRPS22,TIMM10,*	-	-

Figure 31 – Replicability of LCL DE results across datasets – Top panel: overlap of genes common in different expression datasets (Geuvadis, Stranger 2007, Stranger 2012) and different populations (YRI, CEU+TSI). Table shows the identity of common genes found in either two or all three studies.

If we focus on each of the 44 inversions separately and compare the DE genes reported in the different studies, we observe inconsistent results (**Figure 32**) in some cases, mainly for candidate genes in *trans*. For instance, more than 20 DE genes in *trans* are found to be associated to HsInv374 in Stranger 2012 study, whereas the other studies report less than 4; this trend is also found for HsInv0102, HsIn0097, HsInv0072 and others. This lack of consistency may be explained by several factors, such low population size, low inversion frequency and absence of genes detected by RNA-Seq but not in microarrays (mainly long non-coding genes), among other factors. In fact, bias is found when comparing low versus highly populated genotype groups (data not shown), and this is a very common situation given the low frequency of some inversions (see **Figure 14**). In this scenario, a high number of false positives is expected. For all these reasons, we decided to implement stringent filters based on a minimum MAF (5%), population size (N>20), minimum bin size (N>4 or Minimum Genotype Frequency >5%), minimum expression change (log2-fold-change: LogFC>0.20) and replication level (presence in Geuvadis and at least one Stranger study) to discard spurious candidates. This step dramatically reduced the figures, resulting in 11 significantly associated DE candidate genes for 9 inversions (**Table 11**).

Inversion	Gene	Effect	Avg. exp.	LogFC	Study & population	Model
HsInv0105	*INHBA*	*cis*	-1.12	0.59	Stranger 2012 (Pooled) Geuvadis (YRI)	additive
HsInv0124	*IFITM2**	*cis*	6.67	0.18	Stranger 2007 (Pooled) Geuvadis (Pooled)	additive
HsInv0201	*JAKMIP2*	*cis*	-1.91	0.94	Stranger 2012 (GIH) Geuvadis (YRI)	additive
HsInv0340	*IFITM3*	*cis*	1.46	1.00	Stranger 2012 (YRI) Geuvadis (Pooled)	overdominant, dominant
HsInv0389	*PLXNA3*	*cis*	1.00	4.09	Stranger 2012 (Pooled, CEU, CHB+JPT,GIH) Geuvadis (Pooled)	STD-INV, females, males
HsInv0396	*PHKA1*	*cis*	1.00	1.29	Stranger 2007 (YRI) Geuvadis (YRI)	STD-INV, females, males
HsInv0397	*NUP62CL*	*cis*	1.00	0.60	Stranger 2012 (LWK) Geuvadis (Pooled)	STD-INV, females, males

Inversion	Gene	Effect	Avg. exp.	LogFC	Study & population	Model
HsInv0003	*OGDHL*	*trans*	2.11	1.33	Stranger 2012 (YRI) Geuvadis (Pooled)	additive
HsInv0003	*PTGS2*	*cis*	-1.26	0.39	Stranger 2012 (Pooled) Geuvadis (Pooled,YRI)	additive
HsInv0058	*HCG22*	*cis*	4.60	1.45	Stranger 2012 (Pooled, YRI, CEU, CHB+JPT, GIH, LWK) Stranger 2007 (Pooled, CEU, CHB+JPT, YRI) Geuvadis (Pooled, CEU+TSI,YRI)	additive
HsInv0058	*HCG27*	*cis*	1.41	0.34	Stranger 2012 (Pooled,YRI) Stranger 2007 (Pooled, CEU, YRI) Geuvadis (Pooled, CEU+TSI)	additive

Table 11 – Filtered LCL DE candidates - Candidate DE genes after filtering. If the association was found for several models, additive model is shown. Average gene expression and LogFC values obtained from Geuvadis study. LogFC: log2-fold-change. LogFC signal is unspecified. * *IFITM2* has been included despite not passing the filter LogFC<0.20. Transformation of expression data (Avg. exp.) can give place to negative expression values (Materials and Mathods).

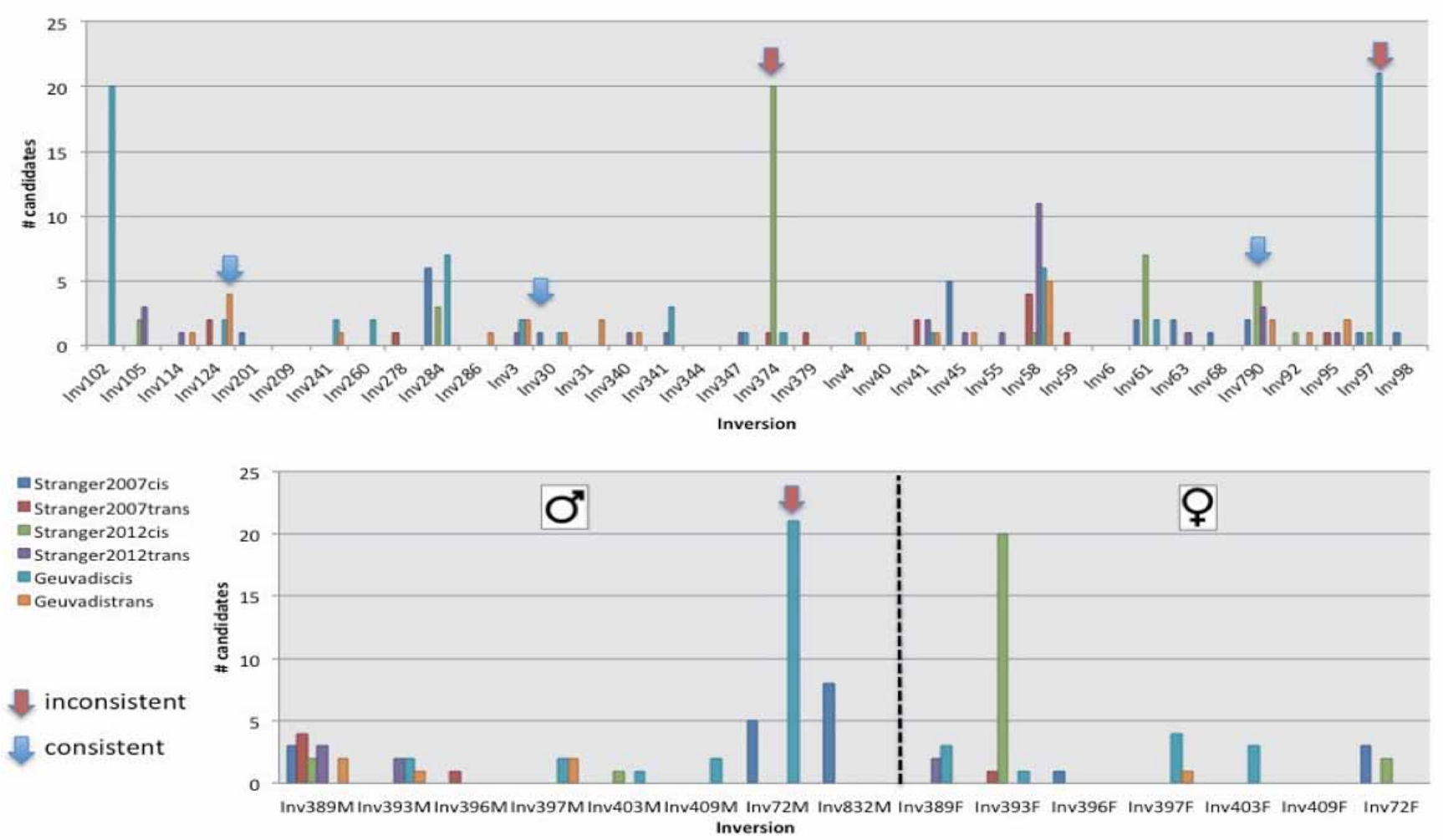


Figure 32 - Overview of LCL DE analysis results – Top panel shows the number of DE genes per inversion located in autosomal chromosomes, bottom panel show results for inversions on sex chromosomes (M, males; F, females). Red arrow shows inconsistent cases: the number of candidate genes differ vastlt between datasets. Blue arrow shows consistent cases: the number of candidate genes is quite homogeneous accross different datasets.

2.3 LCL DE analysis on well-studied inversions

To validate our methodology we benchmarked our pipeline on a gold-standard dataset composed by two well-studied inversions (see Materials and Methods). These inversions, located in chromosomal regions 17q21.31 (Stefansson et al. 2005) and 8p23.1 (Giglio et al. 2001) have been found to be associated with gene expression changes in the literature (de Jong et al. 2012; Salm et al. 2012) and their distribution, origin, functional impact and linkage with disease have been described in the introductory chapter.

In the case of the 17q21.31 inversion, our results are partially in agreement with de Jong's (2012) results. The authors examined the influence of the inversion on 56 transcripts in the 17q21.31 *cis*-region; defined as located inside or within 1 Mb on either side of the inversion. They examined 5 microarray-derived expression datasets: one from blood (sample size = 437, H12.v3 Illumina beadarray) and 4 from brain tissues consisting of frontal cortex, temporal cortex, cerebellum and pons (sample size = 144, H8.v2 Illumina beadarray). For samples of both datasets, inversion genotypes were reconstructed by PCA-analysis of 38 SNPs. In blood, probes were available for 42 genes out of which 22 were expressed above minimum threshold and in the brain 36 genes were available out of which 33 were condirered as expressed in the tissue. Using a linear regression model, authors found 17q21.31 *Inv* haplotype (H2) is associated with decreased expression of *LRRC37A4* (blood), *PLEKH1M* (brain) and *MAPT* (brain) and to increased expression of *MGC57346* (brain and blood), *LRRC37A* (brain) and *CRHR1* (brain).

We corroborated the association of this inversion with decreased expression of *LRRC37A4P*, and increased expression of *CRHR1*, *CRHR1-IT1* (identified as MGC57346 in de Jong et al. study) and *PLEKHM1* (**Table 12**). Interestingly, we observe that *CRHR1-IT1* also seems to be differentially expressed between sexes (**Figure 33**). To our knowledge, this feature is not described in the literature and should be further investigated. On the other hand, we do not detect any association with *MAPT*, as this gene is not expressed in blood. As mentioned, de Jong et al. (2012) found *LRRC37A* differentially expressed with respect to the inversion. When aligning the 50 nucleotide *LRRC37A* probe sequence (identical on H12.v3 and H8.v2 Illumina beadarrays) to Refseq RNA sequences, it was found that this probe aligns not only to *LRRC37A* coding sequence, but also to *LRRC37A2* (100% identity), *LRRC37A3* (100% identity) and *LRRC37A4* (94% identity) genes. Therefore the

contribution of each particular gene to the probe measured expression levels cannot be determined and the strong association can be the result of non-specific binding to more than one target gene in this gene family. However, in our results emanating from the analysis of the RNA-Seq derived expression dataset (Geuvadis), we detect a specific association of 17q21.31 inverted allele with high expression levels of not *LRRC37A* but *LRRC37A2* (**Table 12**). Considering the entire coding sequence (CDS) region, *LRRC37A2* is 99.7% identical to *LRRC37A.* Due to this sequence divergence and contrarily to the microarrays, RNA-Seq mapping is able to assign the expression of the *LRRC37A* paralogous gene family specifically, so we hypothesize that 17q21.31 is actually associated to *LRRC37A2* expression instead of *LRRC37A*. Moreover, we find 17q21.31 inverted allele associated with high expression of *CRHR1* and *PLEKHM1* in blood, even though in de Jong et al. (2012) original work the associations between the inversion and these genes are only found in brain. This could be due to *CRHR1* low expression in blood tissues, yielding poor signal/background ratio in microarray expression measurements and therefore confounding the association analysis. However, this hypothesis does not apply on *PLEKHM1* case, as this gene is highly expressed in blood (>10 log(RPKMs), GTEx data).

Besides the set of genes reported by de Jong et al. (2012), we also found that 17q21.31 inverted allele is associated with high expression of 5 additional genes (*KANSL1, KANSL-AS1, ARL17B, WNT3, NSF*) and to low expression of gene *ARHGAP27*. Some of them (*KANSL-AS1, WNT3, NSF*) were not studied due to the lack of available probes in the work of de Jong et al. (2012). With respect to the remaining three (*KANSL1, ARL17B, ARHGAP27*), although they were expressed above the minimum threshold, our results indicate that in all instances the fold change attributable to the inverted allele is low (LogFC<0.20), so these cases may be false positives.

Gene	LogFC	Avg. Expr.	Expression Dataset	Adj. P-val cis
LRRC37A4P ^{NC}*****	-1.57	3.71	Geuvadis	9.83E-30
KANSL1-AS1 $^{NI\ NC}$	1.15	3.25	Geuvadis, Stranger 2007	1.41E-25
KANSL1 ^{PC}*	0.19	6.29	Geuvadis	8.49E-12
CRHR1-IT1 ^{NC} ***(MGC57346)****	0.15	8.19	Geuvadis, Stranger 2007, Stranger 2012	5.82E-08
ARHGAP27 ^{PC}*	-0.13	6.08	Geuvadis, Stranger 2012	2.71E-05

Gene	LogFC	Avg. Expr.	Expression Dataset	Adj. P-val cis
CRHR1 ^{PC}*****	0.21	1.93	Geuvadis	4.34E-05
ARL17B ^{PC}*****	-0.13	6.08	Geuvadis	5.78E-05
LRRC37A2 ^{PC} ***(LRRC37A)***	0.40	-1.02	Geuvadis	8.13E-03
WNT3 $^{NI\ PC}$	0.55	-1.10	Geuvadis	1.12E-02
NSF $^{NI\ PC}$	0.10	5.6	Geuvadis, Stranger 2012	1.90E-02
PLEKHM1 ^{PC}	0.07	4.52	Geuvadis, Stranger 2007	2.27E-02

Table 12 - 17q21.31-inv differentially expressed genes - Summary of the differentially expressed genes with respect to 17q21.31 genotype in LCLs. In bold, genes found to be differentially expressed in blood in de Jong et al. study (2012). In bold and underlined, genes found to be differentially expressed in brain in de Jong et al. study (2012). The remaining genes have been only found to be associated to 17q21.31 in our study. P-values shown from Geuvadis expression dataset study, adjusted for multiple testing of genes in *cis*. *Detectable gene (expressed above min. threshold) in whole blood in de Jong et al. study. ^{NI} Gene not included in de Jong et al. study expression results. ^{PC} protein-coding. ^{NC} non protein-coding. LogFC: log2-fold-change. Sign of LogFC indicates direction of change. Transformation of expression data (Avg. exp.) can give place to negative expression values (Materials and Mathods).

Interestingly, we have obtained novel findings: *KANSL1-AS1* shows a dramatic increase of expression (LogFC = 1.1) linked to the inverted allele, which seems very unlikely to be a false positive. To our knowledge, this association has not been described in the literature yet. Even more interestingly, further analysis (see Materials and Methods) on 17q21.31 structural rearrangements (**Figure 34**) revealed that some genes are more strongly associated to the number of copies of certain repeats (CNP155 = alpha, CNP205 = beta, CNP210 = gamma) contained in the 17q21.31 inverted region than to 17q21.31 inverted allele (*Inv*) (**Table 13**). These repetitive regions are part of different 17q21.31 structural haplotypes (Boettger et al. 2012; Steinberg et al. 2012). The region consists of three large copy-number polymorphic (CNP) segmental duplications which include short, CNP155 = alpha, (155-kb) and long, CNP205 = beta, (205-kb) duplications corresponding to the promoter and first exon of *KANSL1* associated with *Inv* and *Std* haplotypes, respectively. The third polymorphism, CNP210 = gamma, is 210 kb in length and spans most of the *NSF* gene upstream of *KANSL1* (**Figure 34**).

Variant	**Alpha (α)**	**Beta (β)**	**Gamma (γ)**	**Inv**
Genotype counts	**114:53:08 (0:1:2)**	**136:29:10 (0:1:2)**	**96:66:12 (0:1:2)**	**108:52:12 (STD:HET:INV)**
Gene	**p-val / LogFC**			
KANSL1	9.73E-10 / 0.18	1.05E-03 / 0.12	3.69E-08 / 0.16	**8.49E-12 / 0.19**
ARL17B	**3.33E-09 / 1.31**	6.13E-03 / -0.89	NS	5.78E-05 / 0.96

Variant	Alpha (α)	Beta (β)	Gamma (γ)	Inv
CRHR1	9.09E-04 / 0.20	NS	1.60E-04 / 0.20	**4.34E-05 / 0.21**
ARHGAP27	1.84E-03 / -0.1	NS	3.58E-03 / -0.10	**2.71E-05 / -0.13**
LRRC37A2	1.53E-02 / 0.38	NS	7.89E-02 / 0.31	**8.13E-03 / 0.40**
WNT3	**2.13E-03 / 0.64**	NS	4.19E-03 / 0.60	1.12E-02 / 0.55
NSF	**2.13E-03 / 0.13**	NS	7.89E-02 / 0.09	1.90E-02 / 0.10
PLEKHM1	9.13E-02 / 0.06	NS	7.89E-02 / 0.06	**2.27E-02 / 0.07**
LRRC37A4P NC	3.55E-15 / -1.44	8.32E-03 / 0.58	3.55E-21 /-1.33	**9.83E-30 / -1.57**
KANSL1-AS1 NC	**1.05E-31 / 1.24**	NS	6.92E-16 / 0.93	1.41E-25 / 1.15
CRHR1-IT1 NC	1.12E-07 / 0.15	NS	2.86E-07 / 0.14	**5.82E-08 / 0.15**

Table 13 - 17q21.31 rearrangements differentially expressed genes – Original p-values corresponding to gene-variant associations are shown for the alpha, beta and gamma repeats and for 17q21.31 inverted allele in Geuvadis dataset for pooled sample set. In bold, p-value corresponding to the strongest association. For each variant, the genotypes counts are shown, from less to more alleles/copies. Variant "Inv" correspond to the number of inverted alleles, for the remaining variants genotype counts correspond to number of repeats (in parentheses). For repeat gamma, individuals with 4 copies of the repeat are excluded. NS: non-significant. NC: not coding. LogFC: log2-fold-change. Sign of LogFC indicates direction of change.

For instance, *ARL17B,* and *KANSL-AS*1 are several orders of magnitude more strongly associated to alpha repeat than to 17q21.31 inverted allele (H2) or any other repeat. In some cases, like *WNT3, NSF, PLEKHM, LRRC37A2* and *CRHR1-IT1* the differences are not so clear between the associations with 17q21.31-inv and alpha repeat, although they are much stronger than beta and gamma repeat associations. Thus, it seems that the number of copies of the latter kind of repeats are not responsible for changes in the expression of these genes. In some other cases, like *ARHGAP27, KANLS1* and *LRRC37A4P*, the association with 17q21.31 inverted allele appears to be stronger than with any other variant type, particularly in *LRRC37A4P*. Further investigation should be carried out to elucidate the contribution of 17q21.31 rearrangements to gene expression in *cis* and *trans*, to further understand their functional consequences and associations with disease.

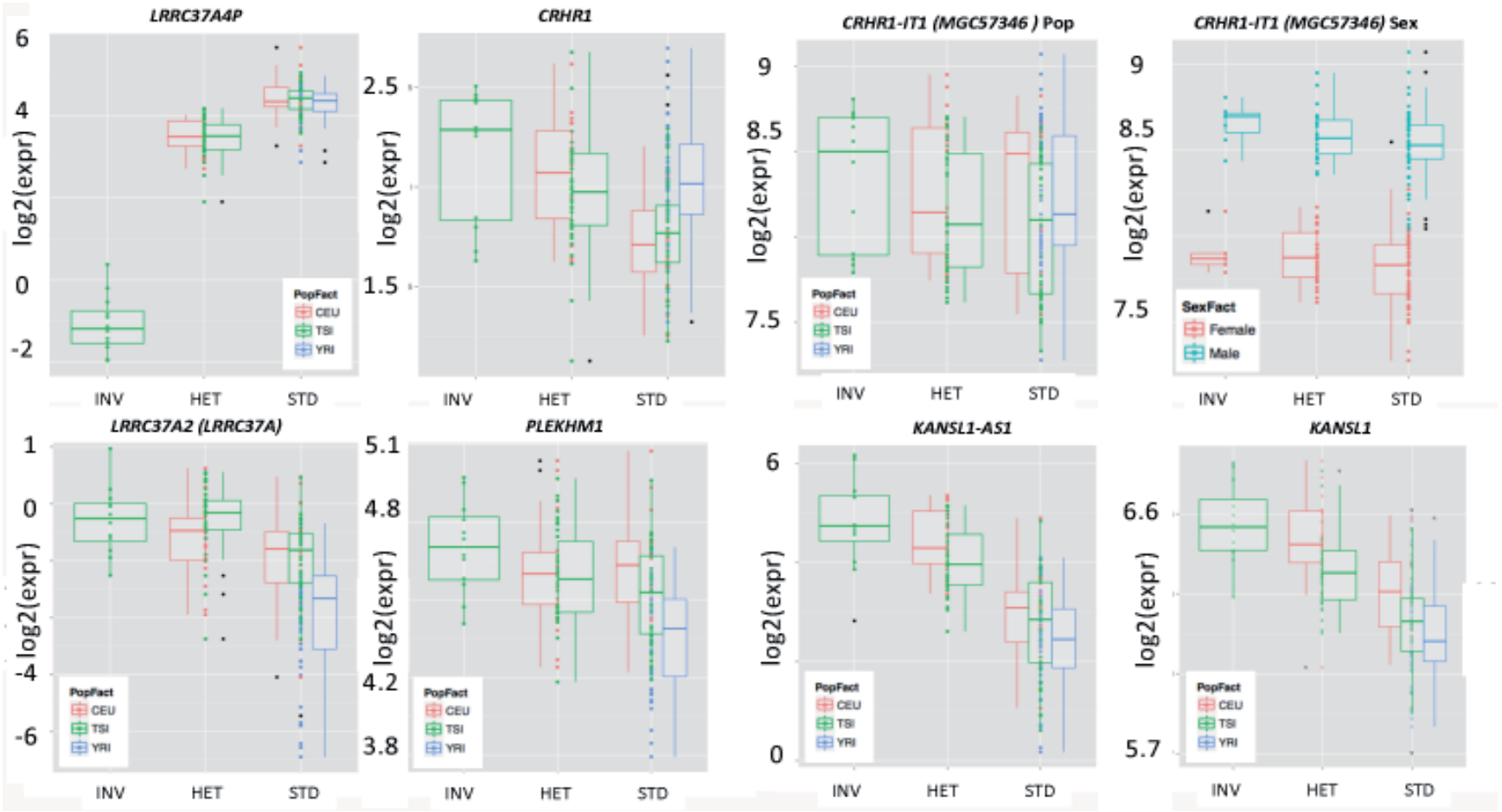
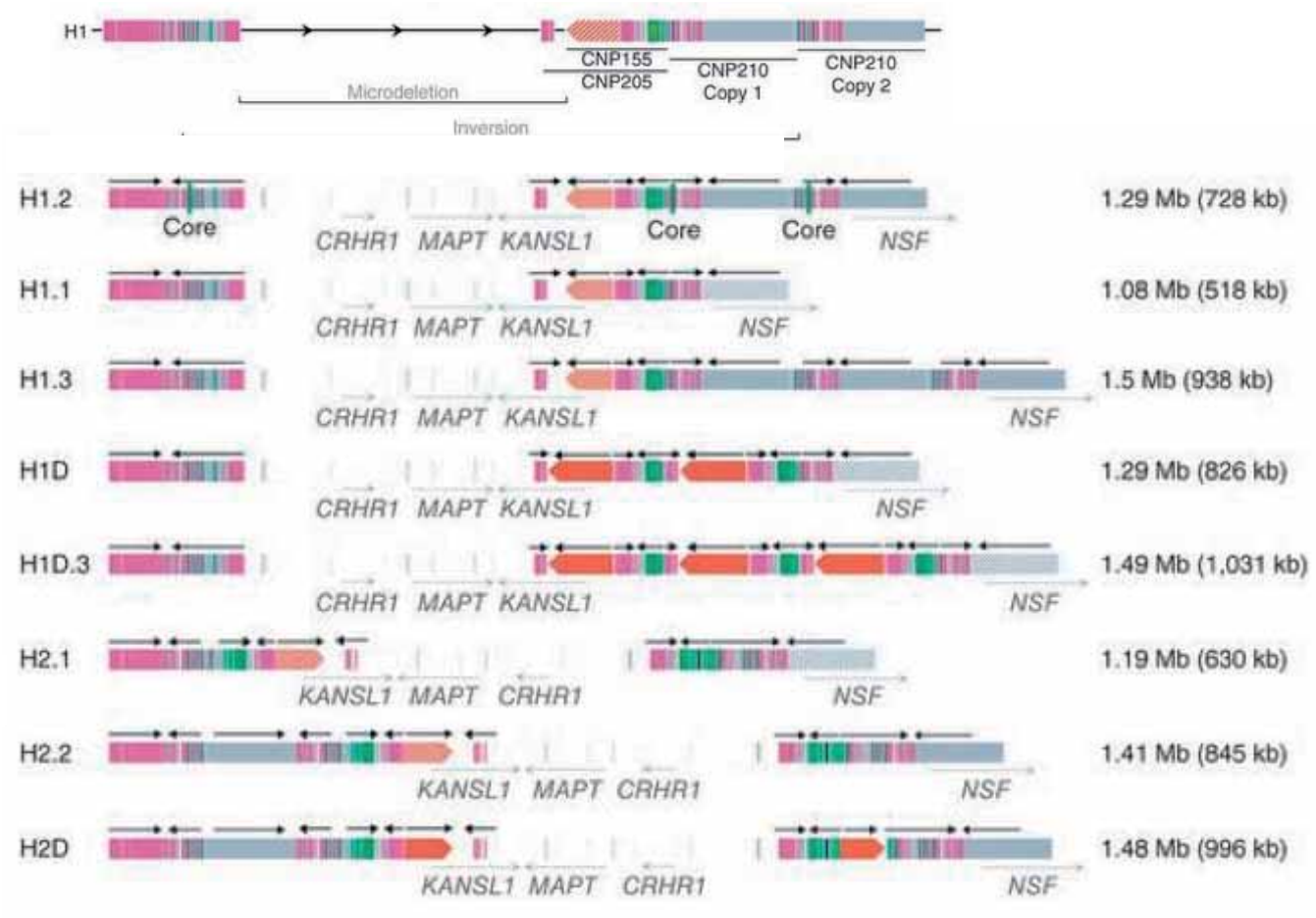


Figure 33 - 17q21.31 differentially expressed genes – Boxplots of a subset of differentially expressed genes with respect to 17q21.31 inversion. Results derived from the Geuvadis dataset.



Steinberg et al	Boettger et al
H1.1	H1.β1.γ1
H1.2	H1.β1.γ2
H1.3	H1.β1.γ3
	H1.β1.γ4
H1D	H1.β2.γ1
H1D.3	H1.β3.γ1
H2.1	
H2.2	H2.α1.γ2
	H2.α2.γ1
H2D	H2.α2.γ2

Figure 34 - 17q21.31 structural haplotypes - Eight distinct structural haplotypes, 5 H1 (Std) and 3 H2 (Inv), ranging in size from 1.08–1.49 Mb (Steinberg et al., 2012). Colored boxes indicate segmental duplications. Hashed boxes correspond to regions present in single copy in that specific haplotype but duplicated in others. Arrows indicate direction of repeat. The duplication content for each haplotype is indicated in parentheses. Equivalences in nomenclature of Boettger et al. (2012) and Steinberg et al. (2012) are shown (right). Four main haplotypes are defined on the basis of KANSL1 copy number and the length of the duplication: H1′ (standard) and H2′ (inverted) with one copy each of KANSL1, H1D H1.β2.γ1) with a long duplication of the gene and H2D (H2.α2.γ2) with a short duplication. H1′ configurations with one copy of NSF are defined as H1.1 (H1.β1.γ1), with two copies as H1.2 (H1.β1.γ2) and with three copies as H1.3 (H1.β1.γ3). H1D configurations with three copies of the long duplication are defined as H1D.3 (H1.β3.γ1). Similarly, H2′ configurations with one copy of NSF are defined as H2.1 and with two copies as H2.2 (H2.α1.γ2). Taken from Steinberg et al. (2012).

For inversion 8p23.1, Salm et al. (2012) examined the influence of 8p23.1 inversion on local gene expression in 5 datasets of different tissues (LCLs and liver) from individuals from European, Asian and African ancestries (see Supplementary Materials of Salm et al. (2012) for more details). The authors found 8p23.1 inversion robustly associated with *BLK*, *PPP1R3B*, *XKR6*, *FAM167A*, and *CTSB* mRNA levels, but only in the association between 8p23.1 inversion and *PPP1R3B* expression the trend is consistent across populations. In our results, we found 8p23.1 inverted allele associated with 8 genes, 3 of which are found by Salm et al. (2012) (**Table 14**).

Salm et al. (2012) found a robust association between inversion genotype and *BLK* transcript abundance accross all four European LCL-derived datasets. The number of *Inv* alleles correlated with decreased *BLK* expression level but the authors hypothesize that this relationship may be population-specific as it was not evident in the YRI or JPT samples. In agreement with the results of Salm et al., we find *BLK* negatively correlated with 8p23.1-inv in all of the three LCL-derived datasets, either in European population or the pooled set of various populations (see Materials and methods) and the same trend is present in GIH population (**Table 14**). The same results are found for *FAM167A*, but in this case the mRNA levels are significantly positively correlated with *Inv* allele dosage. The inversion was also significantly associated with *PPP1R3B* mRNA levels with *Inv* allele dosage positively correlating with transcript abundance. However, contrarily to results of the original work, we do not find that this trend is consistent across populations and we do not detect the association in the RNA-Seq derived dataset (Geuvadis). Finally, we do not find any statistically significant association for *XKR6* and *CTSB.* In the *CTSB* case, although not significant, expression levels present the same trend as in the original work. In *XKR6*, the authors found an association between 8p23.1-inv and the levels of *XKR6* transcripts (AB073660 and AJ305312) but only in an allele-specific expression (ASE)-derived expression dataset (Ge et al. 2009), so we cannot reproduce their results in our analysis.

Gene	LogFC	Avg. Expr	Expression Dataset	Adj. P-val *cis*	Populations	Trend other pops.
BLK	-0.33	6.59	Geuvadis Stranger 2007 Stranger 2012	3.82E-05	Pooled Europeans	GIH
FAM167A	0.44	3.93	Geuvadis Stranger 2007 Stranger 2012	3.02E-03	Pooled Europeans	GIH
PPP1R3B*	0.08	6.76	Stranger 2007 Stranger 2012	5.39E-03	Pooled Europeans	-
ANGPT2	0.33	-1.24	Geuvadis	9.80E-02	Europeans	-
AF131215.2 [NC]	0.20	6.99	Geuvadis	8.33E-05	Europeans	NI
RP11-148O21.2 [NC]	-0.25	3.23	Geuvadis	2.80E-02	Europeans	NI
RP11-10A14.3 [NC]	0.31	1.61	Geuvadis	2.80E-02	Europeans	NI
PRSS51 [NC]	0.48	1.20	Geuvadis	9.19E-02	Europeans	NI

Table 14 - 8p23.1-inv differentially expressed genes – Summary of the differentially expressed genes with respect to 8p23.1 genotype in LCLs. In bold, genes found to be differentially expressed in LCLs in Salm et al. (2012) study. The remaining genes have been only found to be associated to 8p23.1–inv in our study. P-values from Geuvadis study, European (CEU+TSI) samples are shown, adjusted for multiple testing of genes in *cis*. *Results derived from Stranger 2012 and pooled populations are shown [NC] non protein-coding. We consider an association to display the same trend in other populations if the correlation has the same signal and the original pval<0.1. In Stranger 2007, Stranger 2012 genes with normalized expression values below 6.4, have been filtered out. NI: not included in Stranger 2007 and Stranger 2012 analyses. LogFC: log2-fold-change. Sign of LogFC indicates direction of change. Transformation of expression data (Avg. exp.) can give place to negative expression values (Materials and Mathods).

2.4 Inversion-eQTL analyses

To investigate the effect of inversions in non-LCL derived tissues, we identified polymorphisms inherited together with inverted alleles (inversion tag-SNPs) and also reported to be associated to particular gene expression profiles in certain tissues. That is, we identified loci that are both inversion proxies and gene expression quantitative trait loci (eQTLs) and we assume that a certain inversion mirrors the effect of a linked eQTL as they are inherited together (**Figure 35**).

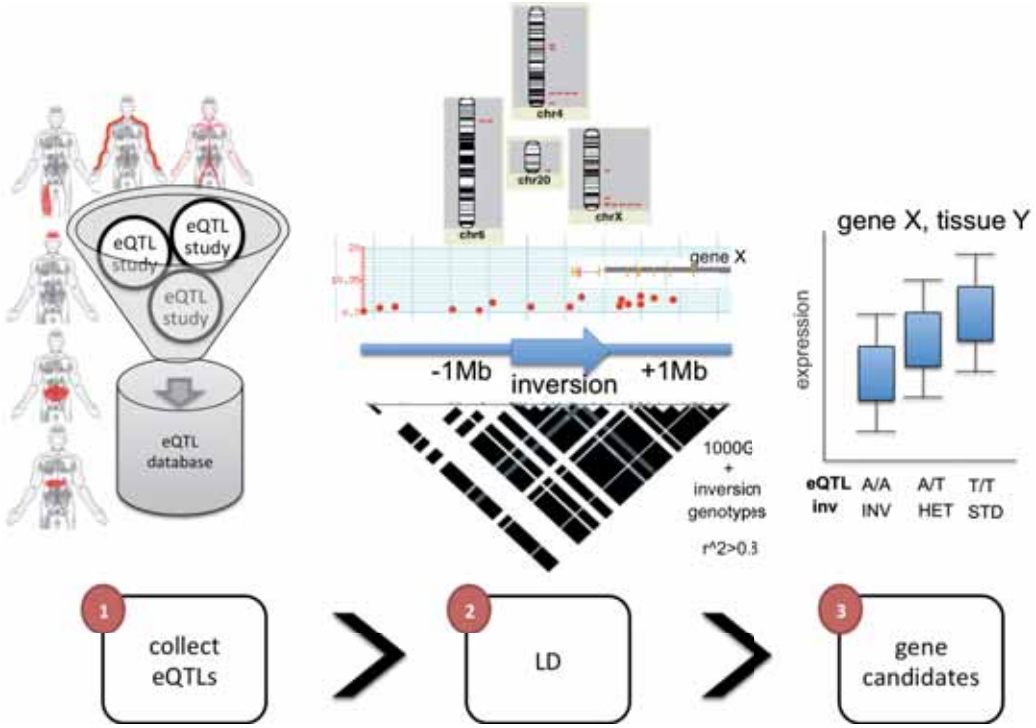


Figure 35 – Scheme of the pipeline to find inversion-eQTLs in non-LCL derived tissues – First step (box labelled 1) consists on collecting eQTL studies and record relevant information regarding polymorphism id., chromosomal position, associated gene, magnitude of the measured quantitative trait change and strength of the association. In the second step (box labelled 2), for each inversion, we calculate the squared correlation coefficient measure of LD (r2) between inversion allele and 1000GP SNPs in *cis*, and those with high LD (r2>0.8) with the inverted allele were recorded. In the last step (box labelled 3), information collected in steps 1 and 2 was cross-referenced and candidate inversions with functional effect (with eQTLs in high LD) were selected.

We first collected eQTL information from 15 eQTL studies covering 11 different tissues: liver, adipose, brain, muscle, skin, thyroid, heart, lung, artery, nerve and blood (**Table 15**). In some cases, eQTL data for different sub-tissues or cell types was available, such as cerebellum, frontal cortex, temporal cortex and brain pons eQTL data for brain; or lymphocyte and monocyte eQTL data for blood. Whereas the sample size of some studies is moderate and ranges from roughly fifty to a few hundreds of samples, in some other cases is close to a thousand (Grundberg et al. 2012) or even surpasses this figure (Zeller et al. 2010), which overall confers adequate statistical depth to the analysis. In addition, we aimed to compile eQTL information that is not population-specific but representative of human diversity, so studies derived from populations from different continents (Africa, Asia and Europe) have been included (**Table 15**). Furthermore, although the analysis is centered

mostly on genetic determinants of gene expression, a single study providing chromatin accessibility quantitative trait loci (DNAse I sensitivity QTLs: dsQTLs) has been included to assess the proportion of inversions that could influence transcription factor binding and having a functional effect. Finally, three studies identifying QTLs at a transcript ratio and exon level have also been included (Lappalainen et al. 2013; Montgomery et al. 2010; Pickrell et al. 2010). These datasets may allow us to identify: a) the effect of inversions altering exon orientation and potentially suppressing the exon expression without necessarily altering the overall expression of the gene; and b) inversions that change the transcript ratio between isoforms for genes with several splicing isoforms expressed in the same tissue.

Reference	Tissue	QTL type	Ancestry	N	Expression Platform
Degner et al. 2012	LCL	dsQTL	African	70	DNase I sequencing
Dimas et al. 2009	T-cell Fibroblast LCL	eQTL	European	75	Microarray
Gaffney et al. 2012 Stranger et al. 2007 Veyrieras et al. 2008	LCL	eQTL	European African Asian *	210	Microarray
Innocenti et al. 2011	Liver	eQTL	European African *	266	Microarray
Montgomery et al. 2010	LCL	eQTL (exon) eQTL (transcript)	European	60	RNA-Seq
Myers et al. 2007	Cortex	eQTL	European	279	Microarray
Pickrell et al. 2010	LCL	eQTL (gene) eQTL (transcript)	African	69	RNA-Seq
Schadt et al. 2008	Liver	eQTL	European	427	Microarray
Zeller et al. 2010	Monocyte	eQTL	European	1490	Microarray
Gibbs et al. 2010	Cerebellum Frontal cortex Temporal cortex Brain pons	eQTL	European African Asian *	143	Microarray
Grundberg et al. 2012	Adipose LCL Skin	eQTL	European	856	Microarray

Reference	Tissue	QTL type	Ancestry	N	Expression Platform
Lappalainen et al. 2013	LCL	eQTL (exon) eQTL (gene)	European African **	210	RNA-Seq
GTEx Consortium 2013	Adipose Artery Blood Heart Lung Muscle Nerve Skin	eQTL	European	>80	RNA-Seq

Table 15 - Characteristics of eQTL datasets -* Populations analysed separately. ** Populations analysed together. dsQTL stands for DNase I sensitivity QTL.

If we measure the number of eQTLs inside inverted rearrangements, we observe that the number of loci associated to expression changes differs greatly across the inversion set (**Figure 36 A**). However, there is a strong correlation of eQTL counts with inversion size (**Figure 36 B**), so the bias disappears when correcting for the number of nucleotides the inverted rearrangement spans. When measuring the number of eQTLs near inversions (+-1 Mb) we observe that there is a great heterogeneity, as there are inversions that totally lack eQTLs (e.g. HsInv0403) and inversions surrounded by thousands of eQTLs (e.g. HsInv0058 has ~4000 eQTLs), the average number for autosomal chromosome inversions being ~400. However, only a few eQTLs are found to be linked with inverted alleles (**Figure 36 C**). Overall, we observe 49 eQTLs associated to 13 inversions, ranging from 1 to 5 eQTLs per inversion, with HsInv0058 as a clear outlier again (18 eQTLs in high LD with the inversion).

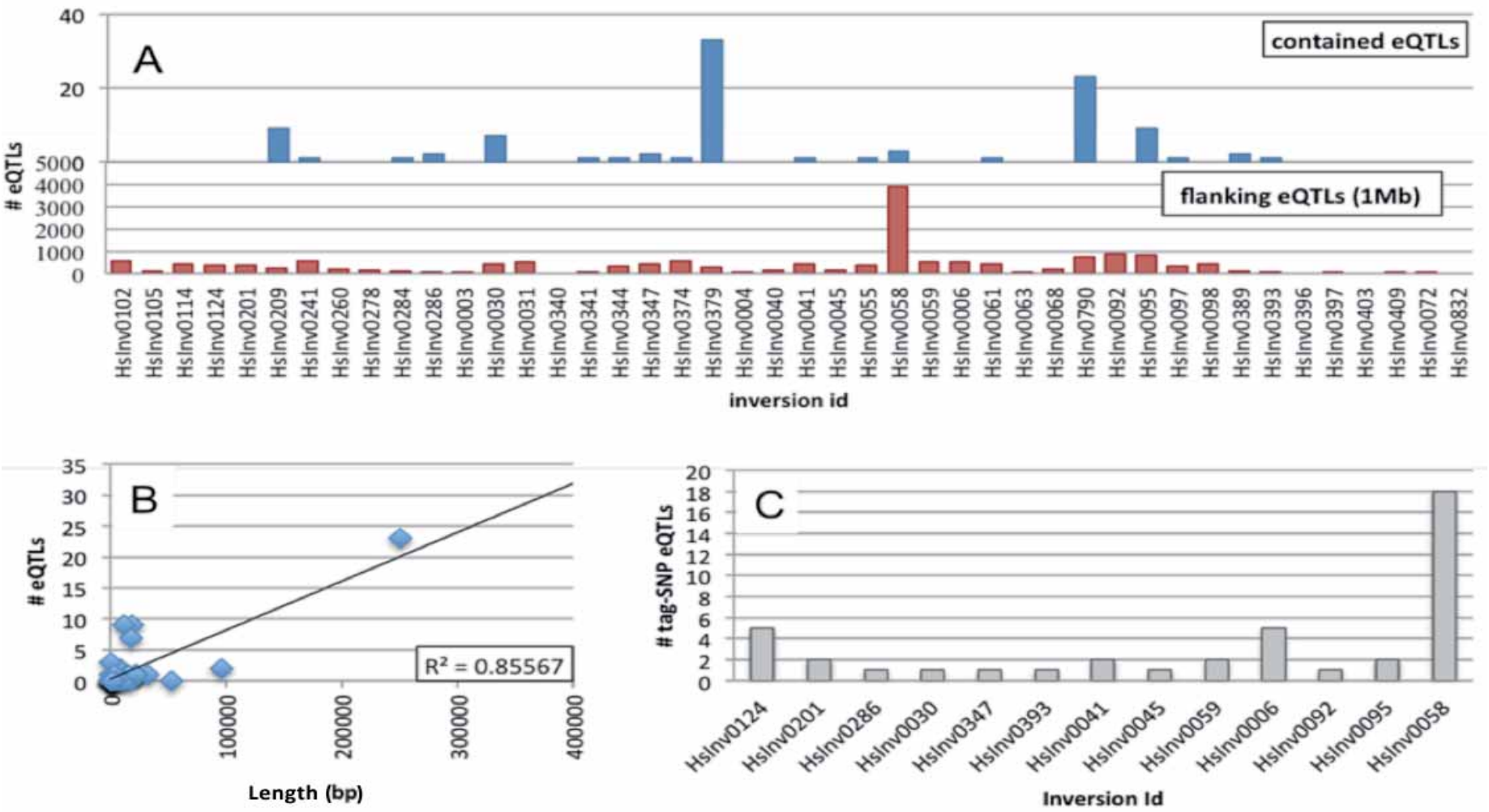


Figure 36 – Inversion eQTLs distribution – A) eQTL counts located in inverted regions (top panel) and flanking regions, spanning 1 Mb downstream and upstream the inversion boundaries (bottom panel). **B)** Number of eQTLs contained within inversion versus inversion size. R^2 = Pearson correlation. **C)** eQTL counts in high LD (r2>0.8) with corresponding inversion loci (contained or flanking the inversion). R2 = squared correlation coefficient.

The 49 candidate eQTLs associated with inversions potentially affect the expression of 36 genes in several tissues (**Table 16**). Interestingly, in some cases we observe consistency with previous LCL DE results as the same genes have been identified by both approaches; for instance HsInv0058-*HCG22* and HsInv0124-*IFITM2* associations (**Table 16**). In addition, some associations appear to be at an exon but not gene level (HsInv0124-*IFITM3*) which suggests that the inversion is affecting the gene at a transcript level, either changing the transcript ratio or suppressing the expression of an exon of a particular isoform.

Inv.	Tissue	Gene	Effect Gene/Exon	Studies	Ancestry
HsInv0124	Blood, LCLs	*IFITM2**	Gene (Blood) Exon (LCLs)	Geuvadis, GTEx	European
	LCLs	*IFITM3**	Exon	Geuvadis	European
	LCLs	*PKP3*	Gene	Montgomery 2010	European
	LCLs	*RP11-326C3.11*^{NA}	Gene	Geuvadis	European
	LCLs, nerve, thyroid, blood	*RP11-326C3.12**	Gene+Exon (LCLs) Gene (nerve, thyroid, blood)	Geuvadis	European
	Artery, heart, lung, nerve, blood	*RP11-326C3.7**	Gene	GTEx	European
HsInv0201	Lung, thyroid	*SPINK14*	Gene	GTEx	European
HsInv0286	Adipose	*SEC61G*	Gene	Grundberg 2012	European
HsInv0030	Skin	*MLKL*	Gene	Grundberg 2012	European
HsInv0347	Adipose	*SIX1*	Gene	Grundberg 2012	European
HsInv0393	LCL	*NXF2*	Gene	Stranger 2007	Asian
HsInv0041	Adipose, skin	*FAM124B*	Gene	Grundberg 2012	European
	Monocyte	*DOCK10*	Gene	Zeller 2010	European
HsInv0045	LCL	*JAM2*	Gene	Stranger 2007	Multiple
HsInv0059	Artery, skin	*GABRR1*	Gene	Grundberg 2012 GTEx	European
	Adipose	*LEMD1*	Gene	Grundberg 2012	European
	LCL	*DSTYK*	Gene	Grundberg 2012	European
	LCL	*TMEM81*	Exon+Gene	Geuvadis	European
	Skin	*CNTN2*	Gene	Grundberg 2012	European
HsInv0006	Skin	*NUAK2*	Gene	Grundberg 2012	European
HsInv0092	Skin	*SAMD3***	Gene	Grundberg 2012	European
HsInv0095	Adipose, LCL	*SPP1**	Gene	Grundberg 2012	European

Inv.	Tissue	Gene	Effect Gene/Exon	Studies	Ancestry
HsInv0058	LCL, adipose, artery, heart, nerve, skin	*HCG22**	Gene	Geuvadis GTEx Montgomery 2010 Stranger 2007 Pickrell 2010	Multiple
	Lung	*PSORS1C3***	Gene	GTEx	European
	Adipose	*SFTA2*	Gene	Grundberg 2012	European
	LCL	*C6orf27*	Gene	Stranger 2007	Asian
	LCL	*CDSN*	Gene	Stranger 2007	Asian
	LCL Monocyte	*HCG27**	Gene & Exon (LCLs) Gene (monocyte)	Geuvadis Grundberg 2012 Zeller 2010	European
	LCL, monocyte	*HLA-B**	Exon (LCLs) Gene (monocyte)	Geuvadis Zeller 2010	European
	LCL	*HLA-C**	Gene	Geuvadis Montgomery 2010 Stranger 2007	Multiple
	LCL	*MICB*	Gene	Grundberg 2012	European
	LCL, skin	*VARS2*	Exon (LCLs) Gene (skin)	Geuvadis Grundberg 2012	European
	Skin	*C6orf15*	Gene	Grundberg 2012	European
	Skin	*HCP5*	Gene	Grundberg 2012	European
	LCL	*POU5F1*	Exon	Geuvadis	European
	LCL	*TCF19*	Exon	Geuvadis	European

Table 16 – Inversion-eQTLs gene associations - * Candidate genes identified in LCL by inversion-eQTL analysis also reported in LCL DE analysis. ** Candidate gene identified in a non-LCL tissue by inversion-eQTL analysis also reported in LCL DE analysis of inversions.

Chapter 3

Characterization of candidate inversions with functional effects

SUMMARY - In this chapter, we characterize in detail the functional impact of a subset of inversions. These cases have been selected on the basis of results obtained by global functional exploratory analyses (LCL DE analyses and inversion-eQTL analyses) that aims to reveal associations of inversions with changes in gene expression in *cis* and *trans*. Therefore, the selected candidates consist of inversions for which there is evidence that they affect gene expression levels. First, we evaluate at which level the inversion is modulating gene expression (gene or transcript level) and in which tissue the association occurs. Second, we attempt to determine the causality of the associations between the inverted rearrangements and the changes of expression of nearby genes in *cis*. Specifically, we aim to distinguish causal associations mediated by positional effects, including cases of gene breakage and inversion of alternatively spliced exons, from indirect associations caused by the linkage of the inversion with a putative causal haplotype.

3.1 HsInv0058 inversion

HsInv0058 is a small (876 bp) inversion that is located in an intergenic region of the short arm of chromosome 6 (p21.33), precisely 11,569 bp upstream of gene *HCG22* and 6,479 bp downstream of gene *MUC22* (distances referred to inversion midpoint and gene TSS). This locus is part of the human major histocompatibility complex (MHC), one of the most gene-dense and polymorphic stretches of human DNA containing 260 genes in a 4-Mb span on chromosomal region 6p21.3 (**Figure 37**). MHC encodes proteins critical to immunity including several controlling antigen processing and presentation.

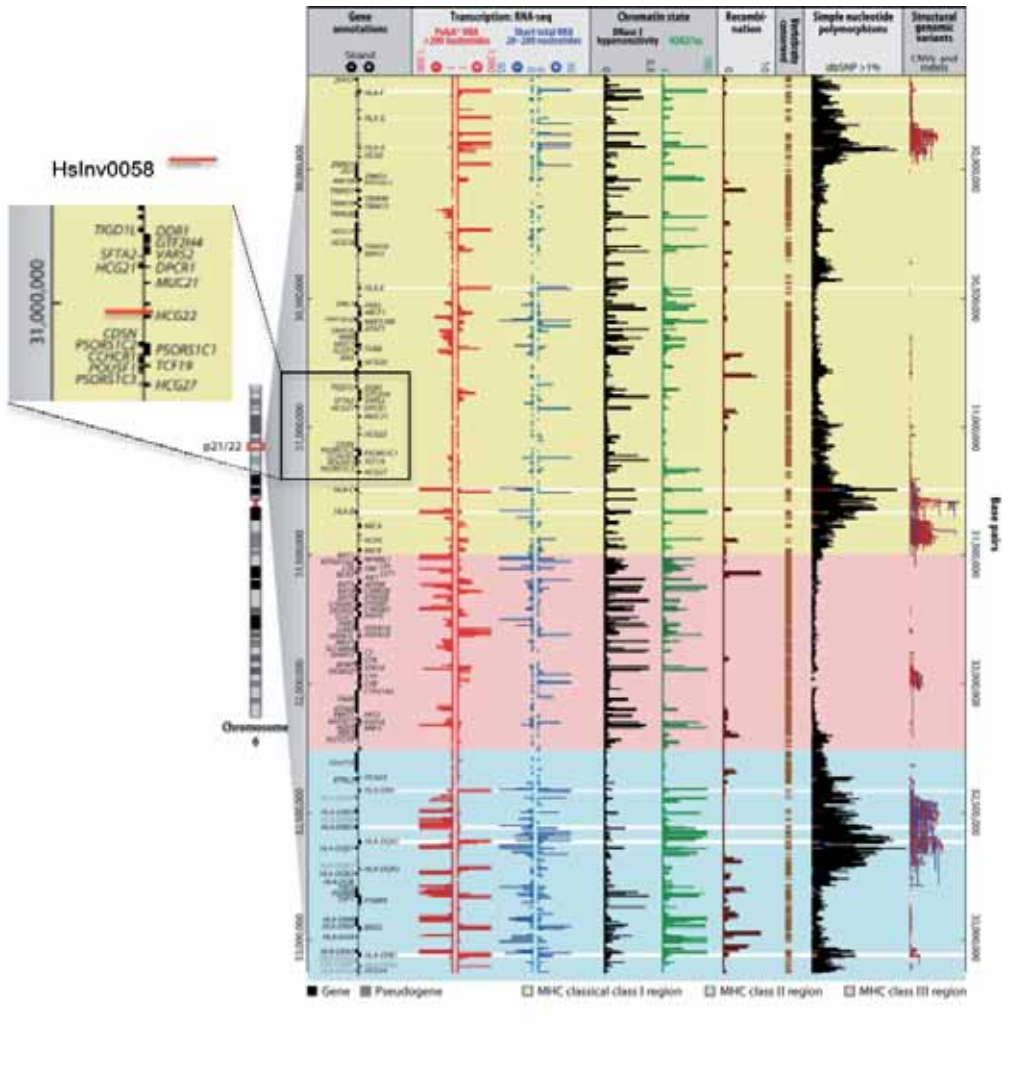


Figure 37 – Overview of MHC region - Scheme of genomic features of MHC region. HsInv0058 region is depicted by red line. Figure adapted from Trowsdale et al. (2013).

Korbel et al. (2007) first identified this structural variant in an African individual (NA18505) by means of PEM. However, inversion boundaries were subsequently refined by Levy et al. (2007) by comparing the genome sequence of a male of Caucasian ethnicity (HuRef) against the reference genome (HG18). Two additional studies identified this structural variant (Arlt et al. 2011; Kidd et al. 2008). Together, these studies provide evidence of the presence of HsInv0058 in individuals of both African and non-African ethnicity. However, Kidd et al. (2008) failed to correctly annotate this inversion, defining it as an insertion. This misannotation is not expected taking into account that the complete sequence of a 40 kb fosmid clone (AC207175.3) harbouring the variant was employed for refining the boundaries of the rearrangement. But this fact pinpoints the difficulty of correctly identifying structural variants even when the entire region is fully sequenced. The clone AC207175.3 is part of the Human Genome Structural Variation Project catalogue and belongs to an individual of Japanese ethnicity. Alignment of the clone against the reference genome provides evidence of the presence of the inversion and reveals the presence of additional structural variants (**Figure 38**). Alignment of the HsInv0058 genomic region with other primate species shows that the ancestral allele corresponds to the inverted conformation, present in chimp. Furthermore, the inversion was experimentally validated in our laboratory by PCR and its exact BPs were characterized and defined at the nucleotide level (D. Vicente and M. Cáceres, unpublished results). The standard rearrangement is associated with additional deletions adjacent to BP1 (2,187 bp) and BP2 (630 bp) (Vicente et al., unpublished results). In the ancestral conformation (*Inv*), the inverted region overlaps SINE element, two of which appear to be partially deleted and inverted in the derived allele (**Figure 38**). Moreover, the inversion also overlaps with regulatory regions as the inverted region contains the TF binding sequences RAD2 and CTCF and partially overlaps with FOSL2 sequence (Wang et al. 2012). As the mapping and annotation of these regulatory regions has been carried out in the reference human genome that comprehends HsInv0058 derived allele, it is not clear which of these elements are also present in HsInv0058 ancestral allele.

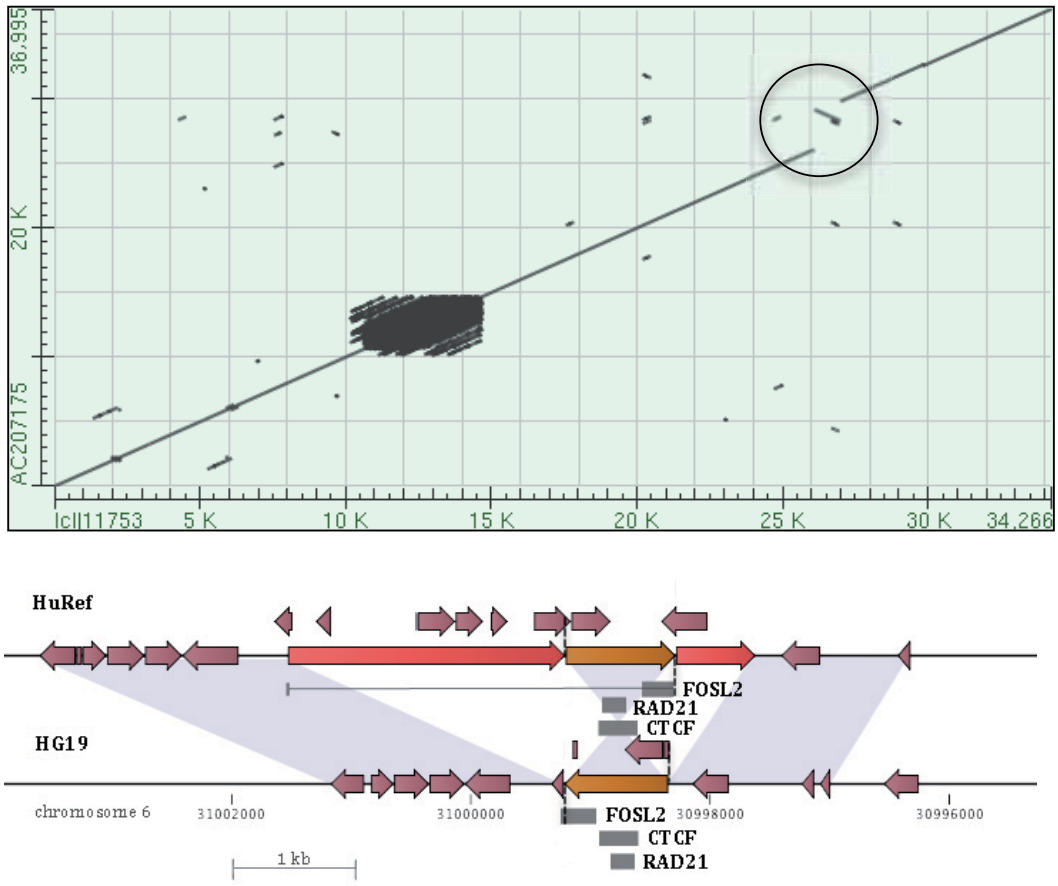


Figure 38 – Mapping of clone AC207175.3 against HG19 and HsInv0058 complex rearrangement – Top panel: HsInv0058 mapping signature is indicated by a circle. Notice the presence of additional deletions adjacent to inversion BP regions. Bottom panel: HsInv0058 ancestral allele (HuRef genome) presents two deletions (red arrows) compared to the derived allele (HG19 genome). HG19 coordinates (chromosome 6) indicated below genome scheme. HsInv0058 inverted region is indicated by an orange arrow. Repeat elements are indicated by purple arrows. Grey boxes represent transcription factor binding sites that overlap with HsInv0058 inverted region. Annotation of TF binding sites was obtained from the TF repository Factorbook (Wang et al. 2012). Notice that FOSL2 sequence is disrupted in HsInv0058 ancestral allele, present in HuRef genome.

.

3.1.1 Recurrence, population distribution and evolutionary history

Based on the available results within the group, it is seen that the derived allele is present and frequent (global freq. = 0.34) in all HapMap3 populations genotyped by MLPA (S. Villatoro and M. Cáceres, unpublished results) (**Table 17**). The variant is in Hardy-Weinberg equilibrium and computation of fixation index (Fst) scores indicates that HsInv0058 frequency is not significantly different across populations (M. Gayà, personal communication).

HsInv0058							
Population	**INV**	**HET**	**STD**	**Total**	**Inverted allele frequency**	**Derived allele frequency**	**OH**
CEU	35	44	12	91	0.63	0.37	0.48
TSI	37	43	10	90	0.65	0.35	0.48
GIH	52	30	8	90	0.74	0.26	0.33
CHB	14	21	10	45	0.54	0.46	0.47
JPT	12	25	8	45	0.54	0.46	0.56
LWK	55	29	6	90	0.77	0.23	0.32
YRI	38	48	13	99	0.63	0.37	0.48
TOTAL	**243**	**240**	**67**	**550**	**0.66**	**0.34**	**0.44**

Table 17 - HsInv0058 distribution in HapMap populations. - HsInv0058 inverted allele frequencies, derived allele frequencies (Derived allele frequency = 1 – Inverted allele frequency), observed heterozygoity levels (OH) and genotype counts for inverted homozygotes (INV), heterozygotes (HET) and standard homozygotes (STD) samples. Data obtained from MLPA genotyping (S. Villatoro and M. Cáceres, unpublished results).

3.1.2 Functional characterization

Results from both LCL DE analysis and inversion-eQTL analysis show strong evidence of association between HsInv0058 genotype and gene expression profile changes in different tissues. Associations occur both in *cis* and *trans*, across different populations and expression datasets. A summary of the top gene candidates is provided here (**Table 18**, **Table 19**).

Gene	Effect	Pop.	Avg. exp.	LogFC	p-val	p-val adj.	p-val adj. cis	Distance HsInv0058 (kb)	
C6orf136	cis	Pooled	-3.679	-0.395	6.28E-04	1.00E+00	8.29E-02	394	U
GTF2H4	cis	Pooled CEU+TSI	1.791	0.208	4.70E-07	1.02E-01	8.30E-04	116	U
HCG22	cis	Pooled CEU+TSI	4.373	-1.532	1.11E-20	8.75E-17	2.11E-19	11	D
HCG27	cis	Pooled CEU+TSI	1.491	0.323	1.99E-07	2.98E-03	1.55E-05	155	D
HLA-C	cis	Pooled	4.340	-0.276	3.65E-04	1.00E+00	5.87E-02	229	D
HLA-B	cis	Pooled CEU+TSI	6.126	-0.393	1.66E-07	2.53E-02	1.47E-04	314	D
DDX39B	cis	Pooled	-0.337	0.249	5.53E-04	1.00E+00	8.14E-02	489	D
MSH5	cis	Pooled	0.058	-0.285	1.14E-05	6.96E-01	1.01E-02	721	D
HLA-H	trans (MHC)	Pooled	6.823	0.364	9.59E-06	3.79E-02	-	1154	U
HLA-J	trans (MHC)	Pooled	5.553	0.478	4.22E-05	8.33E-02	-	1035	U
HLA-DQB1-AS1	trans (MHC)	Pooled	4.757	-0.382	3.50E-05	8.33E-02	-	1618	D
EFNA5	trans (chr5)	CEU+TSI	-2.201	-1.497	4.25E-09	1.29E-03	-	-	
CTC-308K20.1	trans (chr5)	Pooled	1.804	-0.384	6.09E-05	9.61E-02	-	-	

Table 18 - HsInv0058 top LCL DE analysis candidates. - Candidates have been selected according to the following criteria: original p-value < 1x10-3, logFC > 0.2, if candidates show association according to multiple models (additive, dominant, overdominant, recessive) the most significant p-value is chosen (additive). If candidates show association in multiple populations (Pop.), scores for pooled population are shown. If candidates show association for multiple exons, scores for most expressed exon are shown. If candidates show association at gene and exon level, gene scores shown. Distances refer to gene TSS with respect to inversion midpoint, upstream (U) and downstream(D). P-values corrected for multiple testing (see Materials and methods). Abbreviations: population (Pop.), average expression (Average exp.), adjusted (adj.).

For LCL DE analysis, only results derived from Geuvadis RNA-Seq expression dataset are considered as the extreme sequence variation in MHC region may confound associations based on microarray datasets due to probe hybridization mismatching. Nevertheless, we find that top candidates (*HCG22, HCG27, HLA-B, HLA-C*) have also been detected using microarray data and by means of inversion-eQTL analyses, which shows consistency with findings based on RNA-Seq expression data. We observe 8 genes associated to HsInv0058 in *cis* (*HCG22, HCG27, HLA-B* [exon 31323944-31324219], *HLA-C* [exon 31238216-31238262], *DDX39B, MSH5, C6orf136, GTF2H4*), with p-values ranging from 6.28E-04 to 1.11E-20 and logFC ranging from 0.21 to 1.5. Among all *cis* gene candidates, *HCG22* (the closest one to HsInv0058) clearly stands out from the rest, with a much stronger association (p-value = 1.11E-20, logFC = 1.5) compared to the second best candidate *HLA-B* (p-value = 1.66E-07, logFC = 0.4).

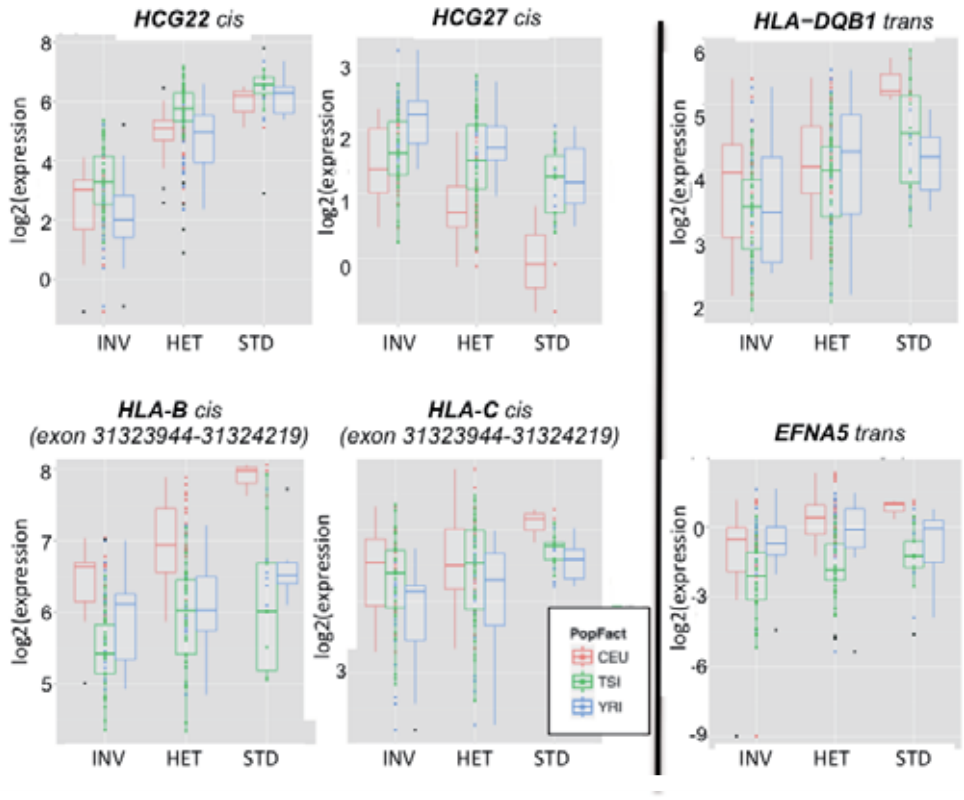


Figure 39 - Top LCL DE analysis candidates of expression association with HsInv0058 genotype – Covariation of expression of candidate genes affected in *cis* (left panel) and *trans* (right panel) with HsInv0058 genotype. Gene expression values (Y axis) are log2 transformed (see LCL DE methodology). Number of samples per genotype: 19 (STD), 87 (HET), 68 (INV).

Top 4 candidates (*HCG22, HCG27, HLA-B, HLA-C*) have also been detected by inversion-eQTL analysis in LCL cell lines. While associations have been obtained by analysing all available samples together (pooled), 4/8 *cis* candidates appear in European populations only, but none of them appear in African populations (YRI). Regarding effects in *trans* (distance gene-inversion > 1 Mb), we observe 5 candidate genes (*HLA-DQB1-AS1, EFNA5, CTC-308K20.1, HLA-J, HLA-H*). Remarkably, 3 of them (*HLA-DQB1-AS1, HLA-J, HLA-H*) are also part of the MHC complex. We also observe that the direction of the changes is not the same for all candidates. HsInv0058 inverted allele is associated with a lower expression of *HCG22, HLA-B* and *HLA-C* but with a higher expression of *HCG27* (**Figure 39**). Remarkably, although these trends are consistent for all analysed populations (CEU, TSI, YRI), CEU shows the strongest and least disperse association pattern in all cases. HsInv0058 also associates to gene expression changes in other non-lymphocyte white blood cells (e.g. monocytes) and in non-blood tissues (**Table 19**). *HCG22* is strongly associated to HsInv0058 inverted allele tag-SNPs in adipose subcutaneous tissue (rs115345573, p-value = 8.70E-11) and displays a moderate association in artery, heart, nerve and skin; whereas *HCG27* and *HLA-B* are differentially expressed in monocytes. The eQTL-inversion analysis provided 4 additional *cis* candidates (*SFTA2*, *VARS2*, *C6orf15* and *HCP5*) in skin and adipose tissue. Importantly, we need to take into account that although eQTLs association scores provide evidence of the strength of the association, these scores are not comparable as they were inferred by several studies that use different statistic models and multiple testing correction methods.

Gene	Tissue	p-val	eQTL
PSORS1C3	Lung	1.40E-07	rs150278551
SFTA2	Adipose	8.99E-01	rs2844665
HCG27	Monocyte	2.64[1]	rs2844669
HLA-B	Monocyte	6.34[1]	rs2844669
VARS2	Skin	2.85E-04	rs2844645
C6orf15	Skin	9.31E-04	rs2517550
HCP5	Skin	1.88E-04	rs2523872
HCG22	Adipose	8.70E-11	rs115345573
	Artery	4.30E-05	rs114448082
	Heart	2.90E-07	rs115345573
	Nerve	4.10E-05	rs116195588
	Skin	1.40E-08	rs114086521

Table 19 – HsInv0058 top inversion-eQTL candidates - Genes with eQTLs in non-LCL tissues and in high LD (r2> 0.8) with HsInv0058 in at least one population (Europeans, Asians, Africans). [1] log10(q-value).

To scan for inversion associations with gene expression changes in non-LCL tissues, we have performed our association analysis (inversion-eQTL analysis) on available eQTL data at the moment. Therefore, this approach is limited to tissues for which an eQTL study has previously been performed. However, there are on-going projects such as GTEx that are aiming at to characterize gene expression and perform eQTL analyses on additional tissues (GTEx Consortium 2013), where some HsInv0058 candidate genes are highly expressed (**Figure 40**). These genes could also be interrogated in GTEx available tissue expression datasets to find associations with HsInv0058 genotype. For instance, we observe a trend between HsInv0058 inverted allele and *HCG22* low expression in breast mammary tissue (rs2517545, p-value = 7.0E-5) (**Figure 41**), but this tissue has not been yet interrogated for eQTLs by GTEx consortium as the number of available samples at the moment (N = 57) is considered not to be enough to perform eQTL analysis (only tissues with N > 60 were selected). Interestingly, we do not observe any *HCG22*-HsInv0058 association in thyroid, where the gene is highly expressed and the number of analysed samples by GTEx is high (N = 112), suggesting that the associations we find could be tissue-specific (**Figure 41**). Additionally, we observe that although *HCG22* is highly expressed in LCLs, it does not seem to be expressed in whole blood (contrary of the other gene candidates that are expressed in both). Indeed, we observe that HsInv0058 is associated to *HCG22* in LCL but not in whole blood (**Figure 41**). This fact could be explained by an artifactual expression of *HCG22* in lymphocytes caused by the EBV–mediated transformation or cell culture conditions but a characterization of *HCG22* expression on the entire spectrum of blood cell types should be performed to validate this hypothesis.

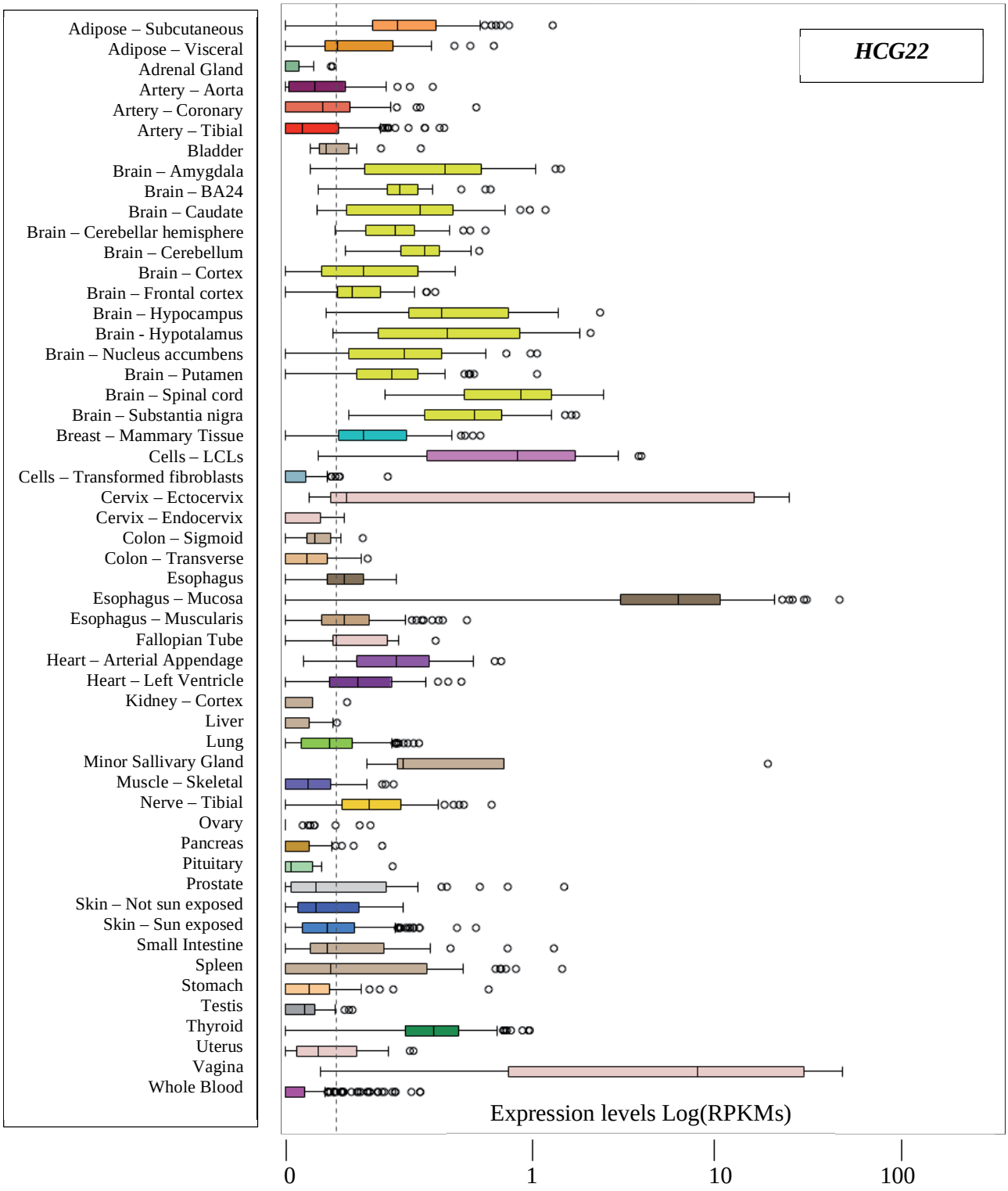


Figure 40 - Expression of HsInv0058 gene candidates in multiple tissues – Expression values are transformed (see GTEx project documentation). Plots obtained from GTEx database; see URLs.

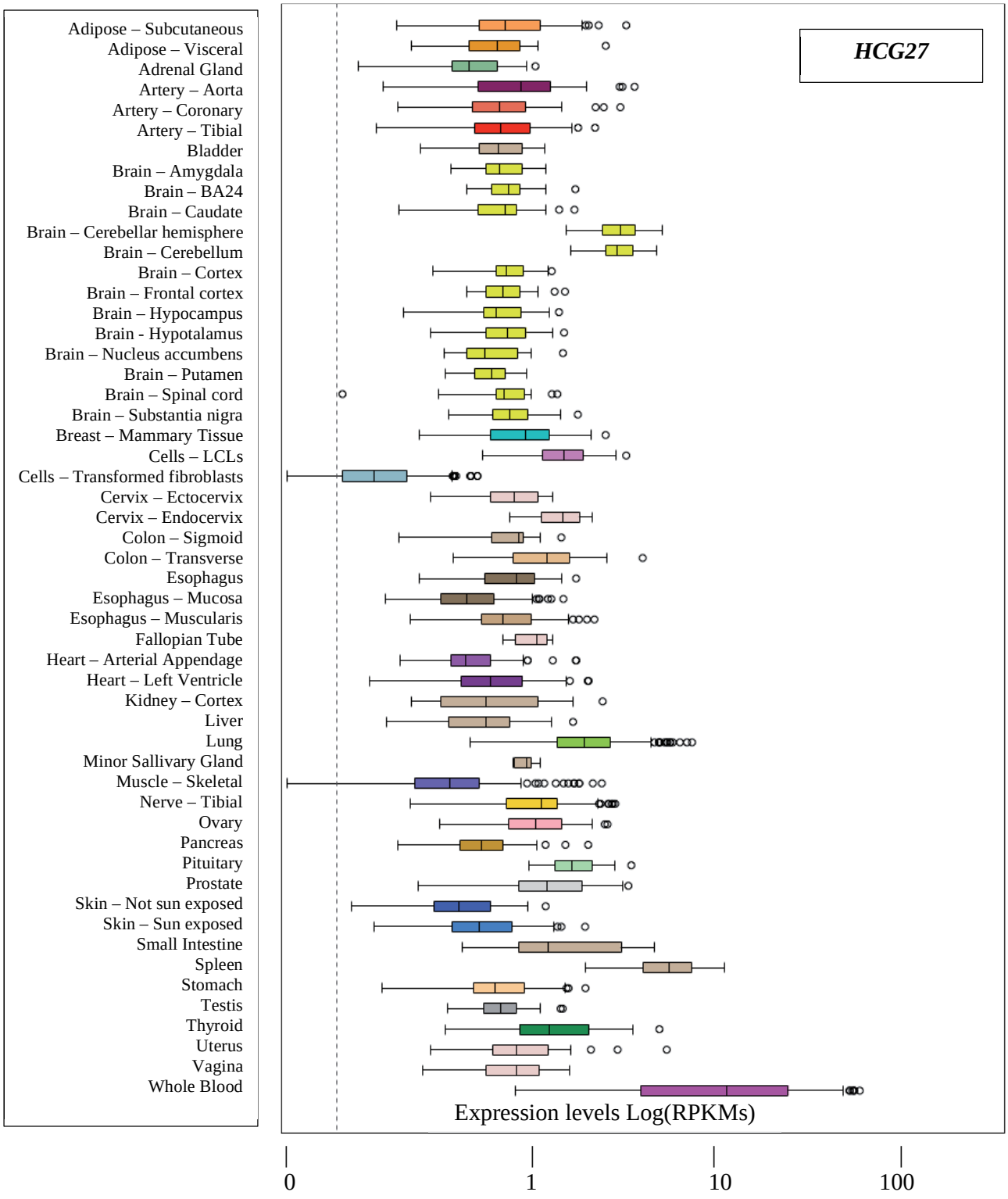


Figure 40 - Expression of HsInv0058 affected gene candidates in multiple tissues – Expression values are transformed (see GTEx project documentation). Plots obtained from GTEx database (see URLs).

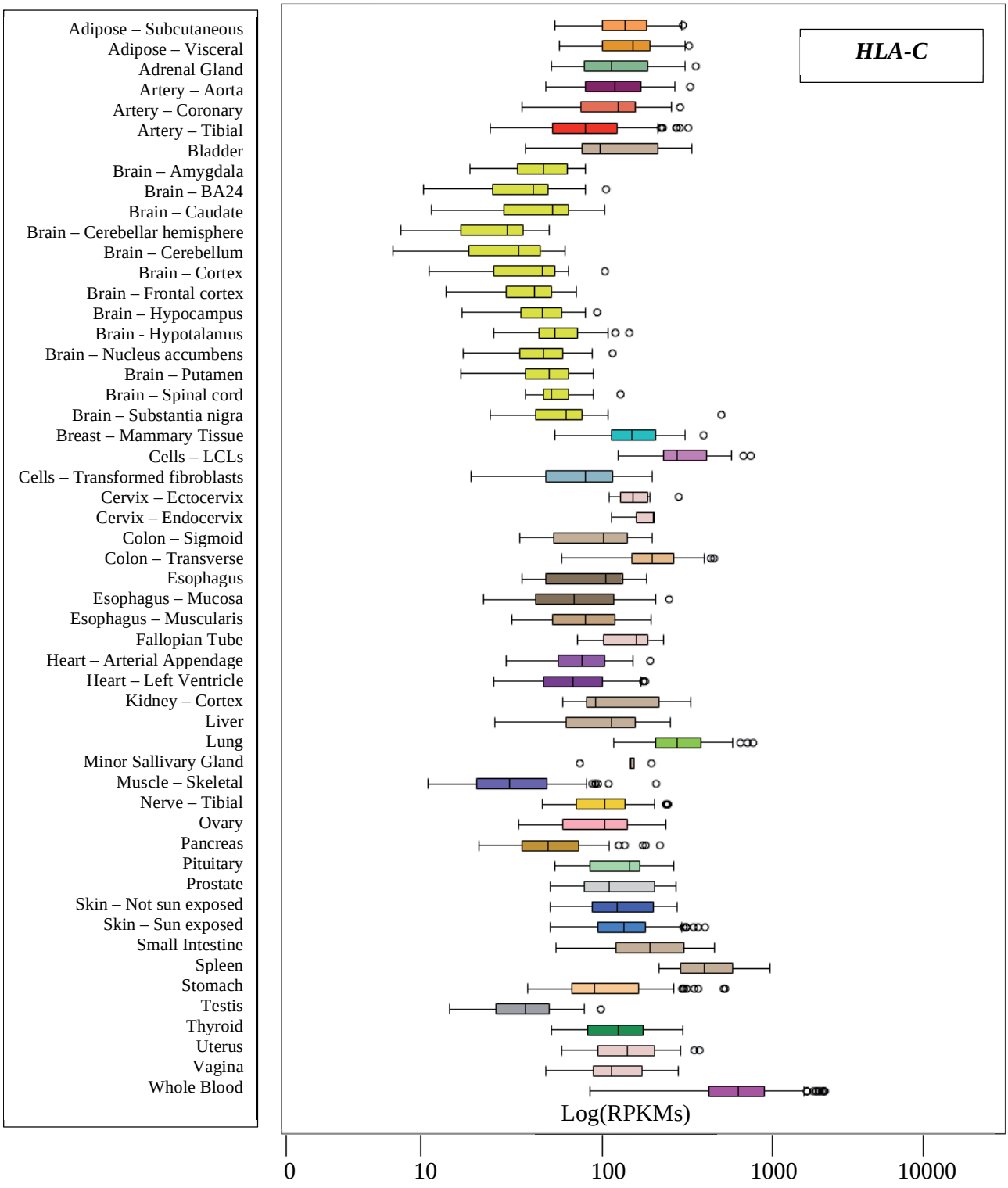


Figure 40- Expression of HsInv0058 gene candidates in multiple tissues – Expression values are transformed (see GTEx project documentation). Plots obtained from GTEx database; see URLs.

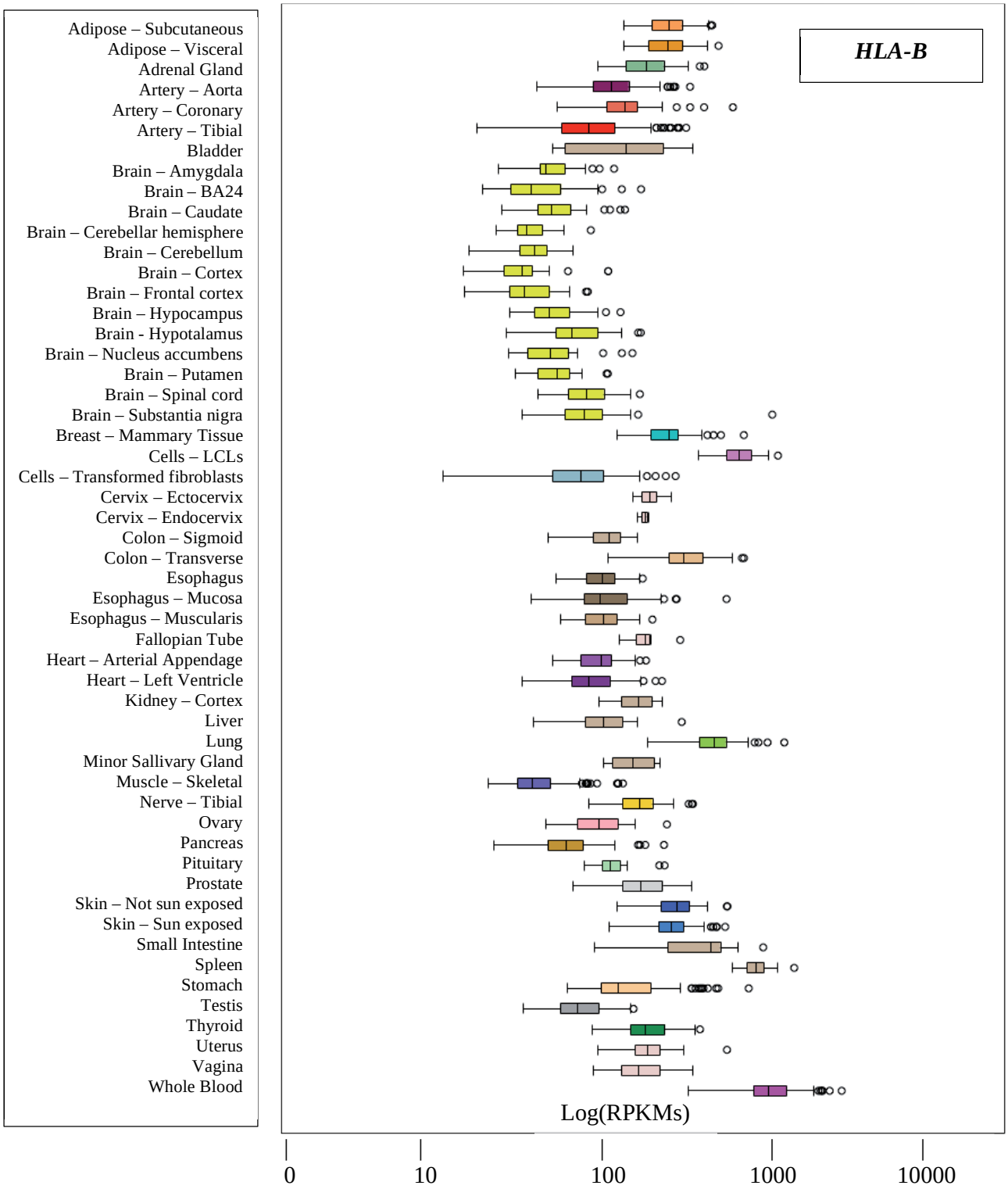


Figure 40- Expression of HsInv0058 gene candidates in multiple tissues – Expression values are transformed (see GTEx project documentation). Plots obtained from GTEx database; see URLs.

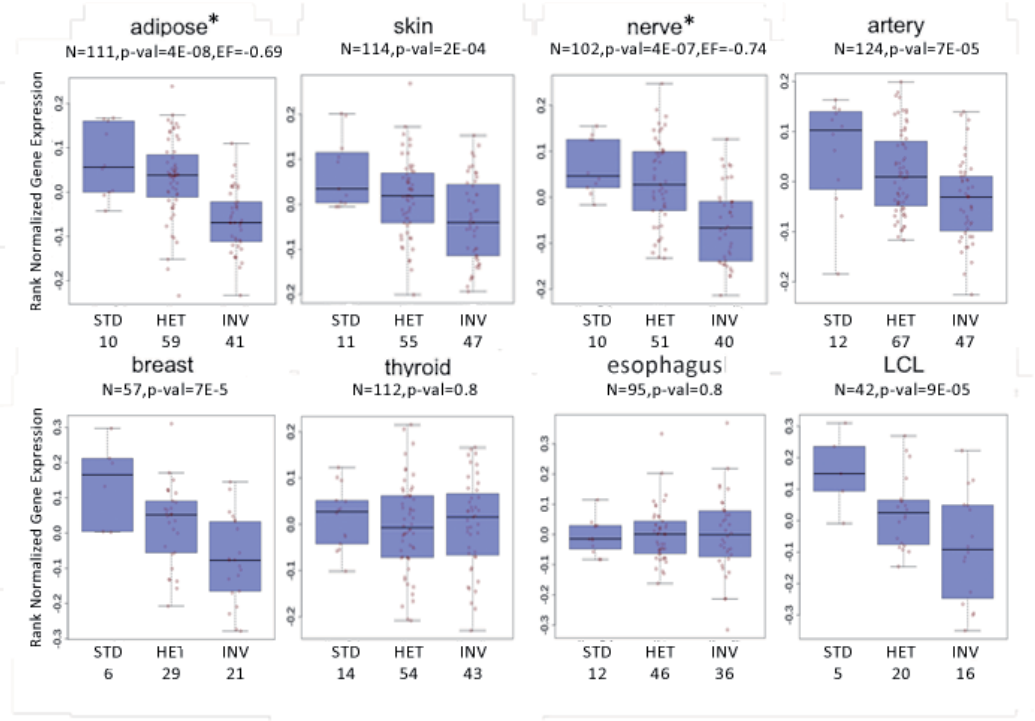


Figure 41 - *HCG22* expression association with HsInv0058 genotype in multiple tissues - Associations of HsInv0058 global tag SNP rs2517545 with HCG22 expression in adipose, skin, nerve, artery, breast, thyroid, and esophagus tissues and LCL cell lines. Number of samples used for the computation of the associations in every tissue is shown (N) as well as the p-value. Tissues with significant eQTL-gene associations (results obtained from GTEx Portal (see URLs), Sep 2014) are marked by an asterisk, and effect size (EF) of the variant is shown. Number of samples per genotype is shown below genotype label. Plots obtained from GTEx Portal (see URLs).

In a parsimonious scenario, if several polymorphisms associate to gene expression changes of a gene in *cis*, the causal variant is the one showing the strongest association signal; and the association of the remaining variants directly correlates with the degree of linkage between them and the causal variant (unless they also have an independent effect). As LD values between HsInv0058 and *cis* regulatory variants was calculated for the inversion-eQTL analysis, we used this data to address the issue. We observe that the inversion is in high LD with both adjacent and far away polymorphisms (data not shown).

To identify the putative causal variant associated to the top candidate gene (*HGC22*) expression changes, we first looked for the variant with the strongest association (the top eQTL) in CEU and YRI populations (**Table 20**). In CEU, the

HCG22 top eQTL corresponds to the single nucleotide polymorphism rs149351084 located ~11,500 bp downstream the inversion and 98 bp downstream the TSS of gene *HCG22*, in an intronic region (Table 14). This variant correlates strongly [-log10(p-value) = 40.96] and negatively (ρ = -0.62) with *HCG22* expression (data from Geuvadis CEU eQTL dataset). In YRI, this variant is also strongly correlated to *HCG22* expression [-log10(p-value) = 11.37, ρ = -0.65], but top eQTL that displays the strongest association [-log10(p-value) = 12.68, ρ = -0.68] in this population is rs114086521. This variant is located close (504 bp) to rs149351084 but is out of the *HCG22* gene body (406 bp upstream) and is also strongly associated to *HCG22* expression in CEU [-log10(p-value) = 38.9, r ρ = -0.60].

The variant rs149351084 is in high LD with the inversion in Europeans (r^2 = 0.8) but not in Asians (r = 0.31) nor Africans (r^2 = 0.57) (**Table 20**). The fact that rs149351084 is not linked with the inversion, but despite that displays strong association with *HCG22* expression changes in Africans could be explained by its putative causal role. The variant rs114086521 is in LD with the inversion in Africans (r^2 = 0.89) and in Europeans (r^2 = 0.71) but not in Asians (r^2 = 0.27). In this case, because this variant is in moderate-high LD with HsInv0058 both in Europeans and Africans, it is difficult to tease apart the effects of the inversion and the SNP on the modulation of *HCG22* expression. The two SNPs are in moderate-high LD in all populations (r^2 = 0.63, 0.81, 0.65 for African, Asian and Europeans, respectively).

	-Log10 (p-val)			
Gene	**Top eQTL**	**Top eQTL r2 > 0.8 HsInv0058**	**rs149351084**	**rs114086521**
HCG22	CEU 40.96 (rs149351084) YRI 12.68 (rs114086521)	CEU 40.96 (rs149351084) YRI 12.68 (rs114086521)	CEU 40.96 YRI 11.37	CEU 38.9 YRI 12.68
HCG27	CEU 22.77 (rs144084123)	CEU 7.66 (rs115545454)	CEU 6.61	CEU 7.6
HLA-C exon (31170149-31171745)	CEU 34.33 (rs115899777)	CEU 5.71 (rs114086521)	-	-
HLA-B exon (31323944-31324219)	CEU 69.3 (rs2523605)	CEU 21.78 (rs116448331)	CEU 19.92	CEU 20.87

		LD SNP ~ HsInv0058			Distance (kb)		
Gene	**Top eQTL r^2 > 0.8 HsInv0058**	**Asians**	**Europeans**	**Africans**	**HsInv0058 Gene**	**SNP Gene**	**SNP HsInv0058**
HCG22	rs149351084	0.31	0.8	0.57	11.6 U	98 bp D, intron	11.6 D
HCG27	rs115545454	1	1	0.98	15.6 U	160.4 U	4.5 U
HLA-C	rs114086521	0.27	0.72	0.89	230.2 U	219 U	11.1 D
HLA-B	rs116448331	0.95	0.94	0.96	315.3 U	312 U	3.3 D

Table 20 – HsInv0058 candidate genes eQTL associations and eQTLs linked with HsInv0058 – Top table: *HLA-B* and *HLA-C* eQTL association values obtained at exon level: *HLA-B* [exon 31323944-31324219], *HLA-C* [exon 31238216-31238262]. eQTL scores obtained from Geuvadis eQTL study in CEU and YRI. Bottom table: Distances refer to SNP position or inversion midpoint with respect to to gene TSS, upstream (U) and downstream.

Then, we explored the correlation between *cis* eQTL top candidates association scores (p-values) for *HCG22*, *HCG27*, *HLA-B*, *HLA-C* with the LD values (r^2) of these loci with HsInv0058, rs149351084 and rs114086521 (**Table 21**, **Figure 42**, **Figure 43**). We observe that *HCG22* association scores are moderately but consistently correlated more with rs149351084 (Spearman's correlation coefficient ρ = 0.87) than with HsInv0058 in CEU (ρ = 0.80) and show a lower correlation with rs114086521 (ρ = 0.77). In YRI both variants rs114086521 and rs149351084 show an equally strong correlation (ρ = 0.83, ρ = 0.82, respectively) and a milder correlation with HsInv0058 (ρ = 0.77). This fact supports the putative causal role of rs149351084 in *HCG22* expression regulation. However, the other candidate genes do not show the same pattern in CEU: *HLA-C* eQTL associations remarkably correlate more strongly with HsInv0058 than with rs149351084 or rs114086521 (ρ = 0.84, ρ = 0.56, ρ = 0.68, respectively). *HLA-B* shows a moderate correlation ($\rho \approx 0.5$) with all variants, whereas *HCG27* shows a moderate correlation with rs114086521 and HsInv0058 (ρ = 0.54) and a poor correlation with rs149351084 and HsInv0058 (ρ = 0.34, ρ = 0.4).

	CEU			YRI		
Gene	**HsInv0058**	**rs149351084**	**rs114086521**	**HsInv0058**	**rs149351084**	**rs114086521**
HCG22	0.80	0.87	0.77	0.77	0.82	0.83
HCG27	-0.41	-0.34	-0.54	-	-	-
HLA-C	-0.84	-0.56	-0.68	-	-	-
HLA-B	-0.50	-0.48	-0.50	-	-	-

Table 21 – Correlations of candidate genes eQTL scores with putative causal variants - Spearman correlation coefficients (ρ) of candidate genes eQTL association values with eQTLs LD values (r^2) with HsInv0058, rs149351084 and rs114086521 loci in European and African populations. eQTL associations were obtained from Geuvadis study in CEU and YRI populations. HLA-B and HLA-C eQTL association values were obtained at exon level: *HLA-B* [exon 31323944-31324219], *HLA-C* [exon 31238216-31238262]. No eQTLs found for *HCG27*, *HLA-B*, *HLA-C* genes in YRI eQTL Geuvadis study.

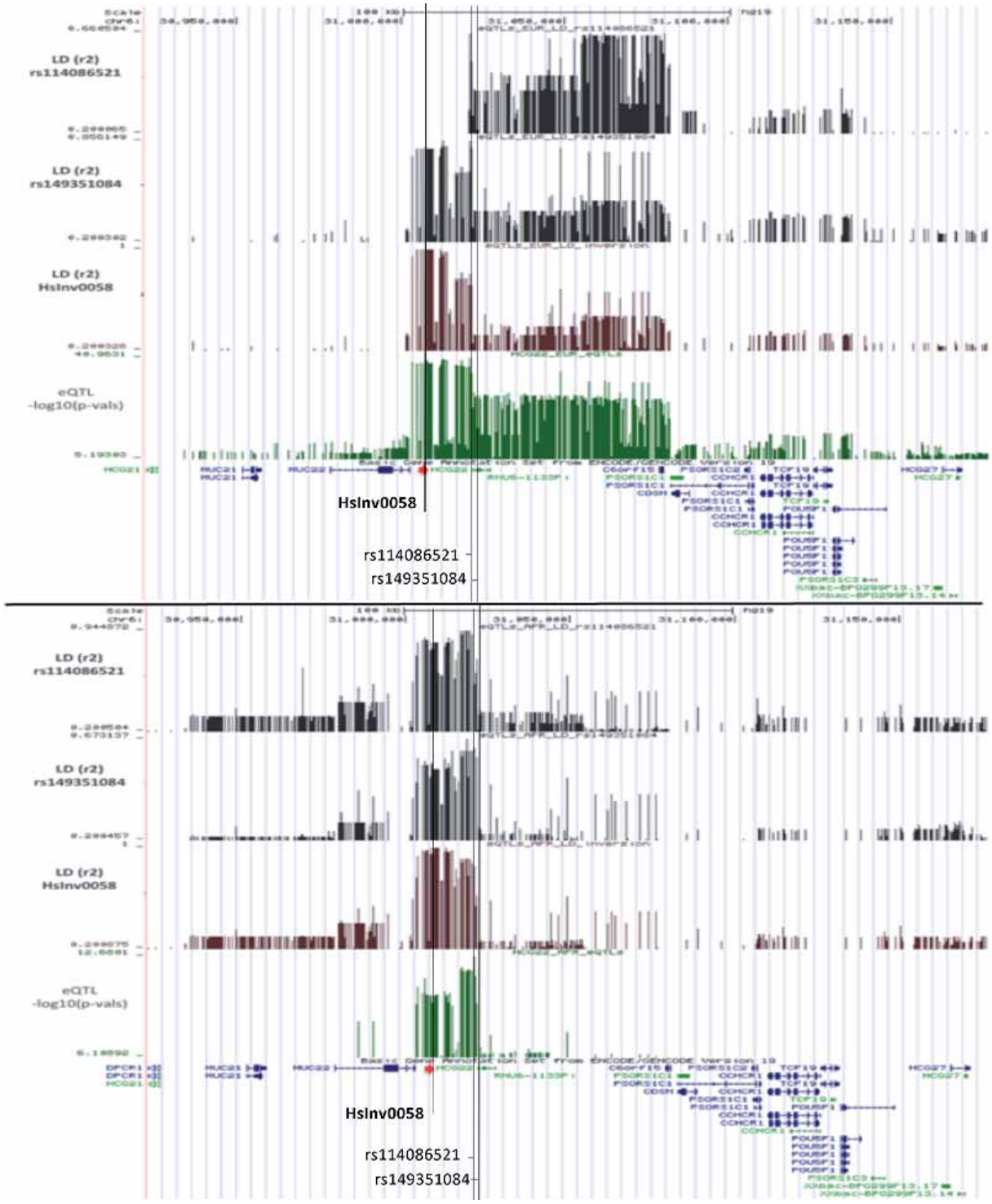


Figure 42– Linkage disequilibrium of *HCG22* eQTLs with putative causal variants - Profiles of LD values between *HCG22* eQTL positions with HsInv0058 (in brown), rs149351084 and rs114086521 loci (in black) in 373 Europeans (top panel) and 89 Africans (bottom panel). *HCG22* eQTLs association scores (in green) are displayed at the bottom of each panel and were obtained from Geuvadis eQTL study in CEU and YRI populations. *HCG22* location is indicated by red arrow. HsInv0058 is located between *HCG22* and *MUC22* genes, the position of HsInv0058, rs149351084 and rs114086521 is indicated by vertical lines.

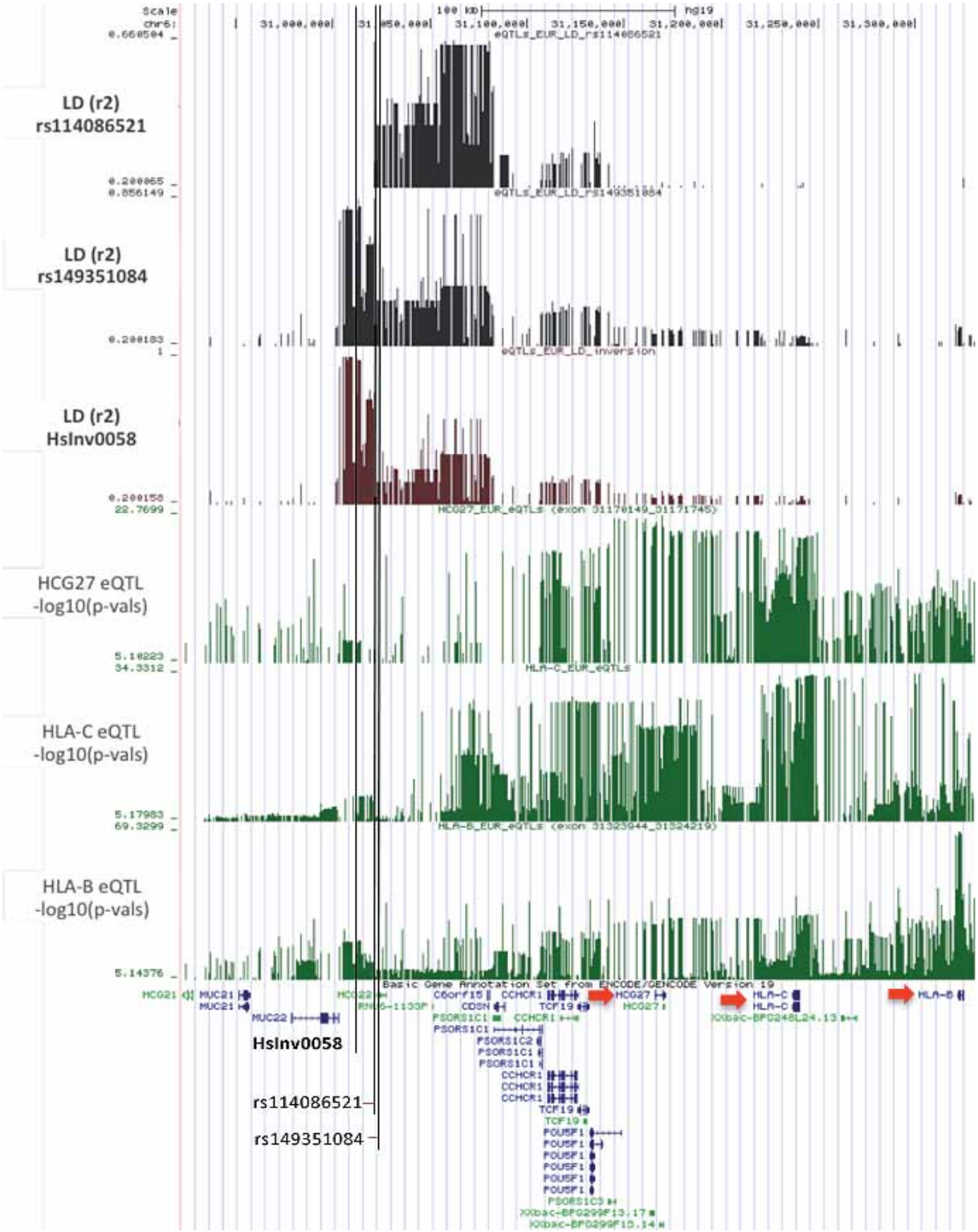


Figure 43 - Linkage disequilibrium of *HCG27, HLA-B, HLA-C* eQTLs with putative causal variants - Profiles of LD values between eQTL positions with HsInv0058 (in brown), rs149351084 and rs114086521 loci (in black) in Europeans. *HCG27*, *HLA-B* and *HLA-C* eQTLs association scores (in green) are displayed at the bottom of the panel and were obtained from Geuvadis eQTL study in CEU. *HLA-B* and *HLA-C* eQTL association values obtained at exon level: *HLA-B* [exon 31323944-31324219], *HLA-C* [exon 31238216-31238262]. *HCG27*, *HLA-B* and *HLA-C* locations are indicated by red arrow. HsInv0058 is located between *HCG22* and *MUC22* genes, and the position of HsInv0058, rs149351084 and rs114086521 is indicated by vertical lines.

This lack of evidence for rs149351084 directly regulating the other gene candidates could be explained by the causal role of HsInv0058 on the regulation of expression of this set of genes. However, this possibility seems unlikely as the inversion is located > 155 kb away of the genes. Moderate linkage between HsInv0058 and a causal haplotype not linked to rs149351084 seems to be a more likely scenario to explain the gene expression changes for this set of genes. To explore that hypothesis, for each candidate gene, we looked at the corresponding top eQTLs in linkage with HsInv0058 (r^2 > 0.8 in at least one population) (**Table 20**). We observe that, contrary to the *HCG22* case, these variants are not the top eQTLs for these genes as they do not show a strong association with the corresponding gene expression changes; in other words, they are not the polymorphisms that seem to be responsible for the observed changes on candidate genes expression profile. For instance, in the *HLA-C* case, its top eQTL rs115899777 (not in LD with HsInv0058) seems to be much more strongly associated to *HLA-C* expression [-log10(p-value) = 34.33] than rs114086521 [-log10(p-value) = 34.33], the top *HLA-C* eQTL among HsInv0058 tag-SNPs. Moreover, the tag-SNPs locate far from the affected genes (> 160 kb) and close to the inversion (~3.3 kb-11.1 kb, **Table 20**).

Together, this evidence suggests that rs149351084 can be the causal variant for *HCG22* differential expression and that the association we found between HsInv0058 genotypes and *HCG22* expression is only due to the strong linkage between the inversion and the putative causal variant. However, an explanation for HsInv0058 association with changes in the expression of the remaining candidates (*HCG27, HLA-B, HLA-C*) has not been found. A possible hypothesis is that lincRNA *HCG22* acts as a master regulator of the expression of this set of genes and the observed association between HsInv0058 and *HCG27, HLA-B, HLA-C* expression is spurious due to partial LD between the inversion locus and *HCG22* regulatory variants. To explore this scenario, functional analyses followed by DE tests comparing *HCG22* high expression versus *HCG22* low expression genotypes could be performed. Another possible scenario is that other uncharacterized variants (SNPs or structural variants) linked to HsInv0058, rs149351084 and rs114086521 exert a causal role in the regulation of gene candidates. For all this we suggest that a refinement of the extended HsInv0058 genomic region should be performed, characterizing and genotyping all unknown variants, previously to invest further effort on trying to determine the causal one. Furthermore, a visual inspection of Geuvadis RNA-Seq split reads mapping to *HCG22* region suggests the presence of non-annotated transcripts (data not shown), so the annotation of this region also needs to be revised and improved (and regulatory regions correctly identified).

3.2 HsInv0124 inversion

HsInv0124 is a ~ 7.7 kb inversion located in the short arm of chromosome 11 (p15.5) (**Figure 44**). Korbel et al. (2007) first identified this structural variant in an African individual (NA18505) by means of PEM. Four additional studies also detected the variant (Arlt et al. 2011; Kidd et al. 2008, 2010; McKernan et al. 2009). However, none of the studies coincided in the determination of the precise boundaries of the inversion. Kidd et al. (2008) predicted the inversion region in individuals from African, European and Asian ancestry by means of PEM, providing evidence for the presence of the rearrangement in the 3 continents. The region was fully sequenced by Kidd et al. (2010) as part of the Human Genome Structural Variation Project; the fosmid sequence AC210760 harbours the inversion and belongs to an individual of African ethnicity. Aguado et al. (2014) genotyped this inversion in 77 individuals mostly with European ancestry and described its recurrence patterns, effects on nucleotide variation and evolutionary history. In addition, global distribution of the inversion has been further investigated by genotyping in 7 human populations (S. Villatoro and M. Cáceres, unpublished results). However, to our knowledge, HsInv0124 functional implications have not been yet assessed.

3.2.1 Recurrence, population distribution and evolutionary history

To determine the uniqueness or recurrence of inversion origin, Median-Joining networks and Neighbor-Joining trees based on CEU HapMap and 1000GP variation data were previously carried out by Aguado et al. (2014). Results show many potential recombination events, including possible gene conversion between arrangements, possibly indicating recurrence of the inversion rearrangement event, which has been subsequently confirmed (M. Gayà and M. Cáceres, unpublished results). In the same article, the authors analysed the distribution of HsInv0124 inverted allele in CEU and observed a frequency of 0.39 and an observed heterozigosity of 0.43. However, this frequency is not global: genotyping results of HsInv0124 in HapMap3 African (YRI, LWK), European (CEU, TSI) and Asian (JPT, CHB, GIH) populations (S. Villatoro and M. Cáceres, unpublished results) show that the inversion frequency differs significantly across populations (Fst = 0.24, p-value < 0.0001) (M. Gayà, personal communication).

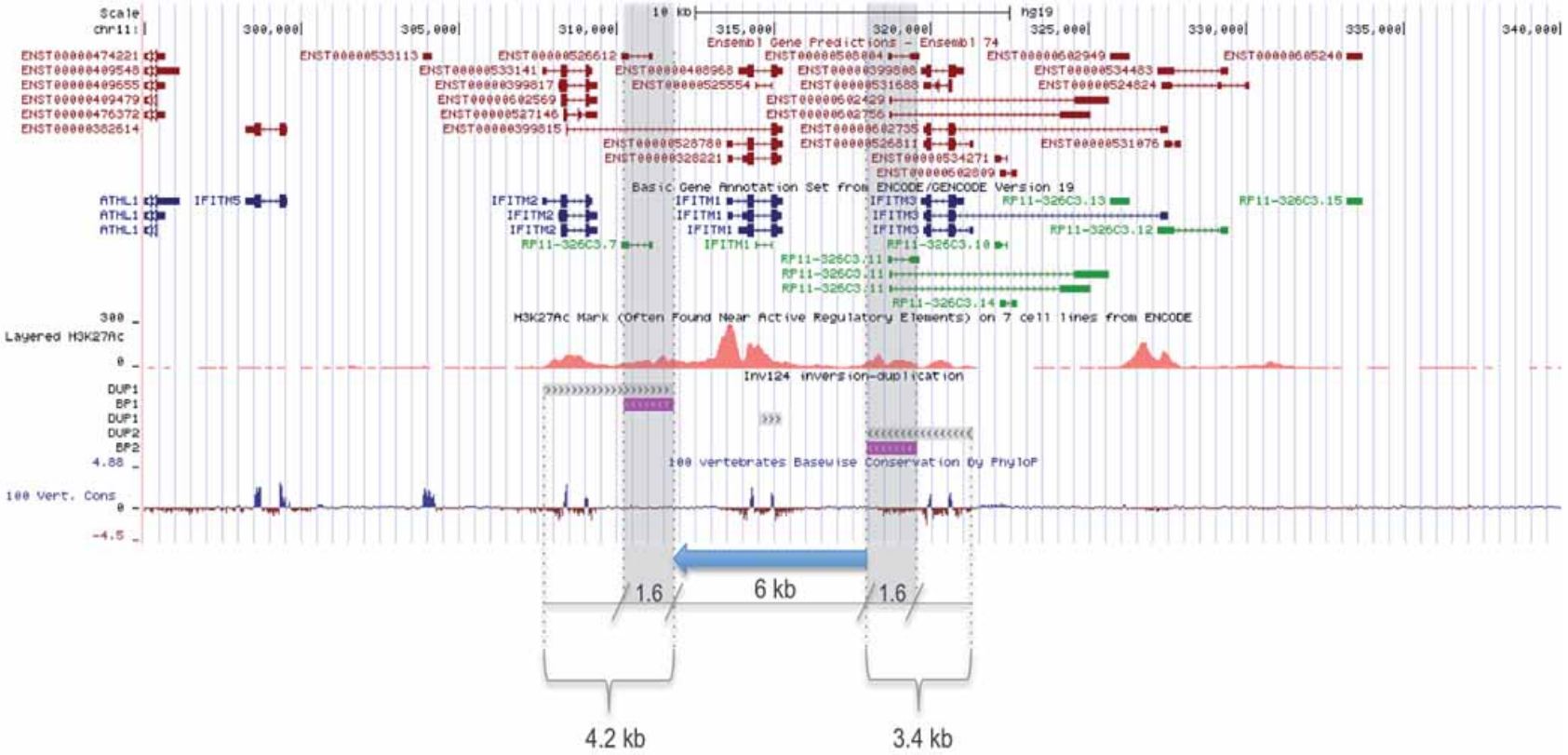


Figure 44 - HsInv0124 genomic region - Scheme of HG19 chr11:294999-340001 region, adapted from UCSC browser. A blue arrow depicts the inverted region minimum interval. BP predicted regions span ~ 1.6 kb and are indicated by pink boxes; duplicated regions span 4.2 kb (Dup1) and 3.4 kb (Dup2) and are indicated by grey boxes. Gene *IFITM1* is located inside the inverted area and *IFITM2* and *IFITM3* are located inside the duplicons that contain inversion breakpoints, but outside of the predicted inverted region. However, *IFITM2* alternative isoform ENST00000399815 is altered by the rearrangement. *RP11-326C3.11, RP11-326C3.7* antisense RNA genes may be broken by the inversion, as inversion predicted BPs overlap their sequence and are not identical. Inverted region also overlaps regulatory regions defined by ENCODE project. Bottom track depicts sequence conservation levels across 100 vertebrate species: high conservation is coloured in blue and poor conservation in red. We observe that *IFITM* genes coding sequences are highly conserved across species.

HsInv0124 inverted allele is more common than standard conformation in Asians (freq. = 0.94-0.98) and Africans (freq. = 0.86-0.88), being less frequent in Europeans (freq. = 0.40-0.45) and in Indian populations (freq. = 0.43) (**Table 22**). Noticeably, although the inversion is in Hardy-Weinberg equilibrium, no evidence for standard homozygotes has been found in CHB and JPT populations, while Europeans show high levels of heterozygosity. Comparison of the region with chimp and gorilla genomes revealed that standard conformation is the ancestral one and genotyping of the inversion in samples of 5 primates (2 gorillas and 4 chimpanzees) showed no evidence of HsInv0124 polymorphism in these species (Aguado et al. 2014).

	HsInv0124					
Population	**INV**	**HET**	**STD**	**Total**	**Inverted freq.**	**OH**
CEU	16	41	34	91	0.40	0.45
TSI	21	38	30	89	0.45	0.43
GIH	45	39	6	90	0.72	0.43
JPT	43	2	0	45	0.98	0.04
CHB	40	5	0	45	0.94	0.11
LWK	64	22	1	87	0.86	0.25
YRI	77	22	1	100	0.88	0.22
TOTAL	**306**	**169**	**72**	**547**	**0.71**	**0.31**

Table 22 – HsInv0124 distribution in HapMap populations - HsInv0124 inverted allele frequencies, observed heterozygoity levels (OH) and genotype counts for inverted homozygotes (INV), heterozygotes (HET) and standard homozygotes (STD). Data obtained from high-throughput genotyping (S. Villatoro and M. Cáceres, unpublished results).

3.2.2 Functional characterization

HsInv0124 contains large, highly identical (~ 97%) inverted duplications in the BP regions spanning 4.16 kb (BP1) and 3.39 kb (BP2). Most probably, these duplicated regions mediated the generation of the inverted rearrangement by a NAHR mechanism. The duplicons contain *IFITM2* and *IFITM3*, 2 members of IFN-induced transmembrane protein (*IFITM*) genes. A third shorter, less identical duplicated region, with the same orientation as BP1 duplicon is located inside the inversion and codes for *IFITM1*, another *IFITM* paralog. *IFITM* gene family plays a crucial role in innate immune system by restricting the replication of multiple pathogenic viruses such as influenza A virus, SARS coronavirus (SARS-CoV), Marburg virus (MARV), Ebola virus (EBOV), Dengue virus (DNV), West Nile virus

(WNV), human immunodeficiency virus type 1 (HIV-1) and vesicular stomatitis virus (VSV) (Everitt et al. 2012; Mudhasani et al. 2013, Jiang et al. 2010, Huang et al. 2011). The inversion also contains 2 antisense RNA genes (*RP11-326C3.11, RP11-326C3.7*) in the BP duplicons and several genes flanking the inverted region; both upstream (pseudogene RP11-326C3.4) and downstream (antisense *RP11-326C3.10* and lincRNAs *RP11-326C3.14, RP11-326C3.13, RP11-326C3.12, RP11-326C3.15*) (**Figure 44**). The inversion does not directly affect the gene structure of the main isoforms of *IFITM2* and *IFITM3* as these protein-coding genes locate outside of the predicted inversion BP region (Aguado et al. 2014). Therefore their coding sequence would remain unchanged in inverted chromosomes. However, *IFITM2/IFITM1* alternative diexonic transcript ENST00000399815 is predicted to be broken by HsInv0124 BP1 in an intronic region, causing the inversion of ENST00000399815 3' coding exon and potentially giving place to the formation of an *IFITM3-IFITM1*fusion gene. The 2 antisense RNA genes (*RP11-326C3.11, RP11-326C3.7*) contain exonic sequences that are not 100% identical in the 2 SDs implicated in the inversion: the RNA transcribed sequence of the isoform ENST00000526612 of the antisense gene *RP11-326C3.7* is 99.7% identical to its duplicated counterpart *RP11-326C3.11* gene isoform ENST00000508004.2. Transcript sequences differ in 4/976 nucleotides, 2 of which locate inside the 5' RNA-coding exon (2/366 differences in total transcript sequence; 99.5% identity). We observe that HsInv0124 BP regions overlap *RP11-326C3.11, RP11-326C3.7* intronic sequence. Therefore, if there is an exchange of genomic fragments between the SDs caused by the inversion rearrangement, the sequence of the mentioned RNA genes could be slightly altered.

Results from both LCL DE analysis and inversion-eQTL analysis show strong evidence of association between HsInv0124 genotype and gene expression profile changes several tissues but mainly in blood. Associations occur in *cis* and *trans* across different populations and expression datasets. A summary of the top gene candidates is provided in **Table 23** and **Table 24**.

Gene	Effect	Pop.	Exonic region	Avg. exp.	LogFC	p-val	p-val adj.	p-val adj. cis	Distance to HsInv0124 (kb)	
IFITM2	*cis*	Pooled CEU+TSI YRI	chr11 308408-308438	1.51	-0.46	6.77E-09	2.06E-03	7.83E-06	6.5	U

Gene	Effect	Pop.	Exonic region	Avg. exp.	LogFC	p-val	p-val adj.	p-val adj. cis	Distance to HsInv0124 (kb)	
IFITM2	*cis*	Pooled CEU+TSI	chr11 308320-308407	3.64	-0.30	2.14E-05	7.88E-01	6.19E-03	6.5	U
IFITM2	*cis*		chr11 308231-308319	3.89	-0.29	3.39E-05	7.88E-01	7.83E-03	6.6	U
IFITM3	*cis*		chr11 320773-321050	-1.57	0.81	5.77E-06	5.85E-01	3.34E-03	6.0	D
IFITM3	*cis*		chr11 320565-320683	-0.77	0.67	1.62E-05	7.88E-01	6.19E-03	5.7	D
IFITM3	*cis*		chr11 320684-320772	-0.96	0.66	5.16E-05	7.88E-01	9.95E-03	5.8	D
IFITM3	*cis*		chr11 319669-319745	-1.35	0.70	1.15E-04	8.14E-01	1.82E-02	5.9	D
IFITM3	*cis*		chr11 319746-319772	-0.87	0.74	1.26E-04	8.14E-01	1.82E-02	4.8	D
IFITM3	*cis*		chr11 319773-319990	0.09	0.65	4.41E-04	8.93E-01	5.67E-02	4.9	D
IFITM3	*cis*		Whole gene	1.49	0.57	8.21E-04	5.15E-01	3.34E-02	4.8	D
RP11-326C3.12	*cis*	Pooled	Whole gene	3.19	-0.66	6.98E-04	9.98E-01	8.37E-03	12.3	D
RP11-326C3.11	*cis*	Pooled	Whole gene	1.29	-0.71	4.37E-05	1.15E-01	5.25E-04	4.8	U
DENND1B	*trans chr1*	Pooled	Whole gene	2.79	-0.79	8.49E-06	4.23E-02	-	-	
FABP3	*trans chr1*	Pooled	Whole gene	3.00	-0.96	9.76E-08	1.46E-03	-	-	
GIMAP6	*trans chr7*	Pooled	Whole gene	6.15	0.52	1.39E-05	5.11E-02	-	-	
IGHV4-34	*trans chr14*	Pooled	Whole gene	3.18	1.07	2.15E-05	8.49E-02	-	-	
LINC00649	*trans chr21*	YRI	Whole gene	5.94	-1.10	7.15E-06	2.82E-02	-	-	

Gene	Effect	Pop.	Exonic region	Avg. exp.	LogFC	p-val	p-val adj.	p-val adj. cis	Distance to HsInv0124 (kb)
RP11-416N13.1	*trans chr7*	Pooled	Whole gene	1.86	0.35	1.39E-05	8.49E-02	-	-
SDC1	*trans chr2*	Pooled CEU+TSI	Whole gene	4.13	1.01	8.17E-06	4.23E-02	-	-
TTC21A	*trans chr3*	Pooled	Whole gene	1.31	0.36	1.71E-05	5.11E-02	-	-

Table 23 – HsInv0124 top LCL DE analysis candidates. Candidates have been selected according to following criteria: original p-value < 1E-03, LogFC > 0.2, if candidates show association according to multiple models (additive, dominant, overdominant, recessive) the most significant p-value is chosen. If candidates show association in multiple populations, scores for pooled population are shown. If candidates show association at gene and exon level, both values are shown. Distances refer to gene TSS with respect to inversion midpoint, upstream (U) and downstream (D). Only the results of the RNA-Seq expression dataset (Geuvadis) are shown, because most of the associations exist only for RNA genes and for protein coding genes alternative isoforms for which no microarray probe was available. For *IFITM* genes, p-value for each exonic region is provided, as a synthetic exon dataset was generated for the analysis (Lappalainen et al., 2013). P-values corrected for multiple testing (see Materials and methods). Abbreviations: population (Pop.), average expression (Avg. exp.), adjusted (adj.).

In LCL DE analysis, we observe 4 genes associated to HsInv0124 in *cis* (*IFITM2, IFITM3, RP11-326C3.11* and *RP11-326C3.12*), with p-values ranging from 8.2E-04 to 6.77E-09 and logFCs ranging from 0.30 to 0.81 (**Table 23**). However, *IFITM2* and *IFITM3* do not show any signal at gene but at exon level, suggesting that the observed associations are isoform specific. Precisely, for *IFITM2* only the 5' exon of the main isoform is differentially expressed, whereas for *IFITM3* the DE applies to both exons of the main isoform. The associations found for *IFITM* genes are found in all populations analysed altogether (pooled) and also in Europeans analysed separately (CEU+TSI), but not in Africans (YRI), with the exception of the 5' exon of *IFITM2* alternative isoform ENST00000399815 for which the association is found in both populations analysed separately. Regarding effects in *trans* (distance gene-inversion > 1 Mb), we observe 8 candidate genes, none of them located in the same chromosome as HsInv0124 (chr11). Among *trans* candidates we find a gene that is also related (like *IFITM* genes) to HIV infection: *SDC1*, encoding a protein that is a transmembrane (type I) heparan sulfate proteoglycan and is part of the syndecan proteoglycan family. The syndecan receptors are required for internalization of the HIV-1 tat protein.

P-values of eQTL associations

eQTL	*IFITM2* (5' exon)	*IFITM3* (5'exon)	*RP11-326C3.11*	*RP11-326C3.7*	*RP11-326C3.12*	Tissue
rs75117940	-	-	9.73E-13	4.40E-05	-	LCL (Geuvadis) lung (GTEx)
rs3809111	-	-	2.47E-13	4.00E-05	-	LCL (Geuvadis) lung (GTEx)
rs909097	-	-	1.25E-12	3.50E-05	-	LCL (Geuvadis) lung (GTEx)
rs909098	-	-	7.93E-13	3.60E-05	-	LCL (Geuvadis) lung (GTEx)
rs72867737	9.57E-07	2.29E-08	2.03E-16	1.70E-06 4.50E-05 1.20E-06 8.10E-10	-	LCL (Geuvadis) blood (GTEx) heart (GTEx) lung (GTEx) nerve (GTEx)
rs77612739	2.76E-07	1.01E-08	1.64E-15	2.00E-06 3.50E-07 2.60E-09	-	LCL (Geuvadis) blood (GTEx) lung (GTEx) nerve (GTEx)
rs12421894	4.16E-07	2.44-07	6.20E-14	5.10E-06 1.10E-05 6.60E-08 9.50E-05	-	LCL (Geuvadis) blood (GTEx) lung (GTEx) nerve (GTEx) artery (GTEx)
rs9666295	-	-	-	-	2.10E-09 1.40E-05 6.90E-05 2.60E-06 7.40E-09	LCL (Geuvadis) blood (GTEx) thyroid (GTEx) nerve (GTEx) lung (GTEx)

Table 24 – HsInv0124 top inversion-eQTL candidates - Genes with eQTLs in high linkage disequilibrium (LD > 0.8) with HsInv0124 in at least one population (Europeans, Asians, Africans). eQTLs contained by the inversion are shown in bold. The study in which the eQTL association has been found is indicated in parenthesis.

Inversion-eQTL analyses corroborate LCL DE analysis findings, providing evidence for *IFITM2, IFITM3, RP11-326C3.11* and *RP11-326C3.12* being differentially expressed in LCL according to HsInv0124 genotypes, with *IFITM2*, *IFITM3* showing the a transcript specific effect pattern (**Table 24**). We also observe that the direction of expression changes is not the same for all affected genes. HsInv0124 inverted allele is negative correlated with *IFITM2* and *RP11-326C3.7* expression and positively correlated with *IFITM3* and *RP11-326C3.12* (**Figure 45**). Although most of HsInv0124 associations have only been observed in LCL, the inversion linked SNPs also correlate to gene expression changes in whole blood and in non-blood tissues. We observe that *IFITM2, RP11-326C3.7* and *RP11-326C3.12* are differentially expressed in whole blood, but the expression of RNA genes is also associated to HsInv0124 in other tissues such as nerve, thyroid, heart, artery and lung. However, all the associated eQTL loci are not linked to HsInv0124 in Africans

and Europeans (r^2 < 0.2) even though they were included in the inversion-eQTL analysis because they are in perfect LD disequilibrium (r^2 = 1) with the inversion in Asians. In HsInv0124 case, all eQTL associations have been found only in individuals with European origin (Geuvadis and GTEx study), so the lack of linkage of the inversion with the aforementioned eQTLs in this population casts doubt on the validity of the associations found by inversion-eQTL analysis for this inversion and we limit our findings to the results obtained by LCL DE analyses. With respect to the expression profile of HsInv0124 candidate genes, we observe that the expression profile of *IFITM2* and *IFITM3* is quite similar and both genes are more expressed in whole blood than in LCL (ratio ~ 1.8 fold) (**Figure 46**) although there is evidence for high *IFITM* genes expression in several non-blood tissues.

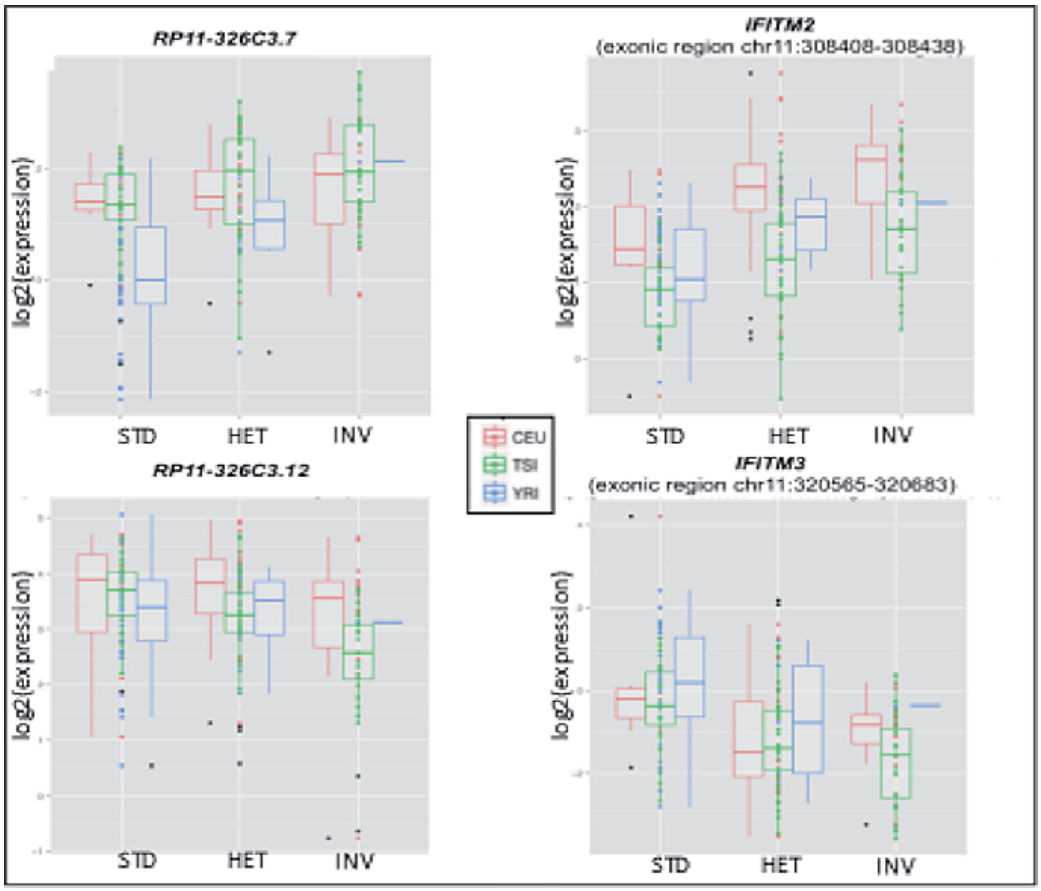


Figure 45 – LCL expression of top DE analysis candidates for HsInv0124 – Covariation of **LCL** expression of candidate genes affected in *cis* with HsInv0124 genotype. Gene expression values (log2(expression)) are log2 transformed (see LCL DE methodology). Number of samples per genotype: 124 (STD), 65 (HET), 43 (INV).

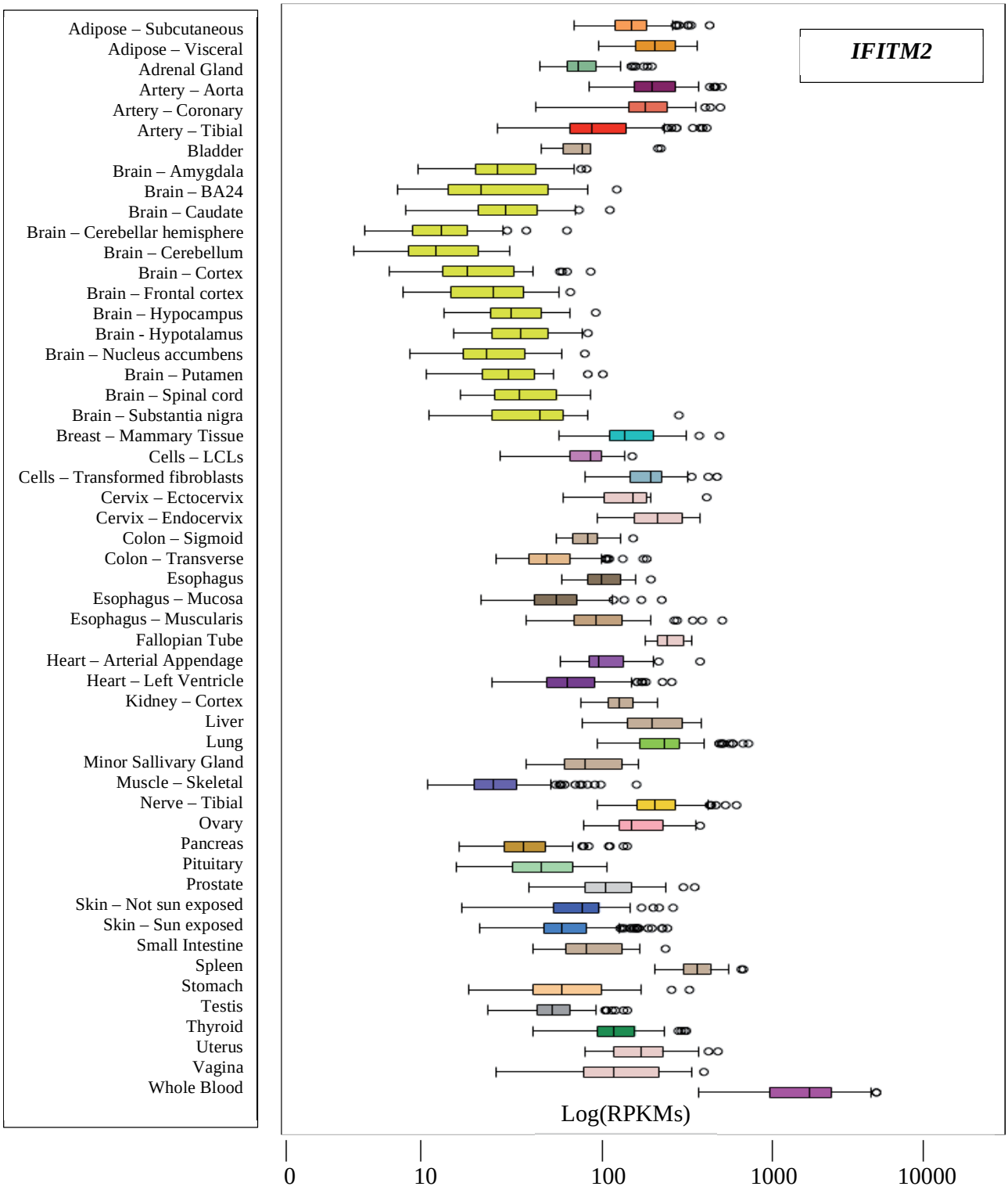


Figure 46 - *IFITM2* and *IFITM3* expression profile in multiple tissues - Image adapted from GTEx database (see URLs). Expression values are transformed (see GTEx project documentation).

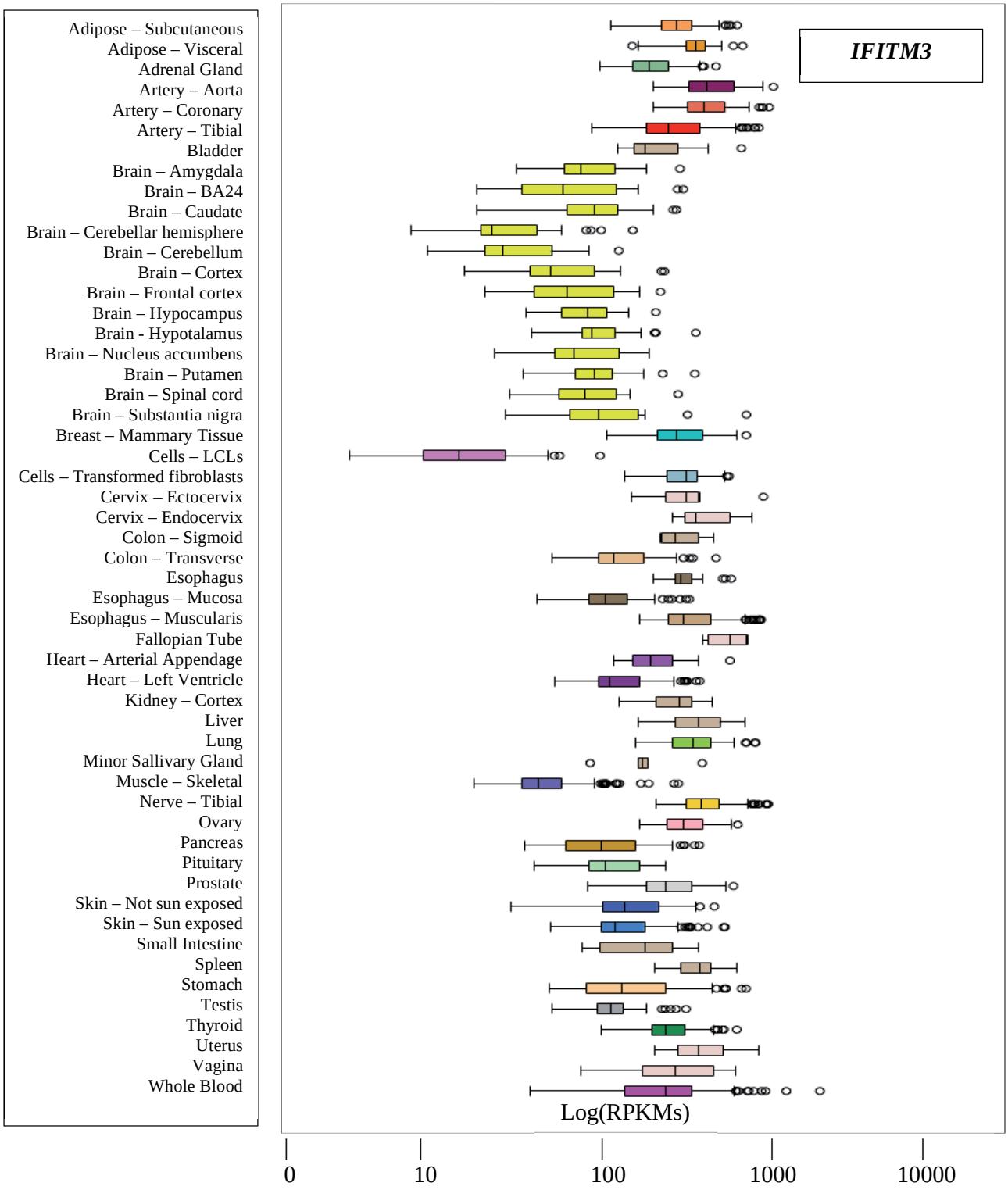


Figure 46 - *IFITM2* and *IFITM3* expression profile in multiple tissues - Image adapted from GTEx database (see URLs). Expression values are transformed (see GTEx project documentation).

3.3 Other inversion candidates

HsInv0102 is an inversion located in the short arm of chromosome 4 (p14) (**Figure 47**), spans 3022 bp and contains no inverted repeats at the BPs, although microhomology is observed at the inversion boundaries and it possibly mediated the inversion rearrangement via NHEJ (Martínez-Fundichely et al., 2014). HsInv0102 affects an intronic region of gene *RHOH* by inverting an untranslated alternatively spliced exon of transcript ENST00000508513 (**Figure 47**). Korbel et al. (2007) and Arlt et al. (2011) identified this structural variant in European individuals by means of PEM, but they did not coincide in the determination of the precise boundaries of the inversion. The INVFEST database (Martinez-Fundichely et al. 2014) also reported that the inversion has been identified in a YRI HapMap individual (Martinez-Fundichely et al., in preparation). Thus, these studies provide evidence of the presence of HsInv0102 in individuals of European and African ethnicity, but HsInv0102 frequency, worldwide distribution, evolutionary history and functional implications have not been determined and are analysed elsewhere (Villatoro et al. in preparation).

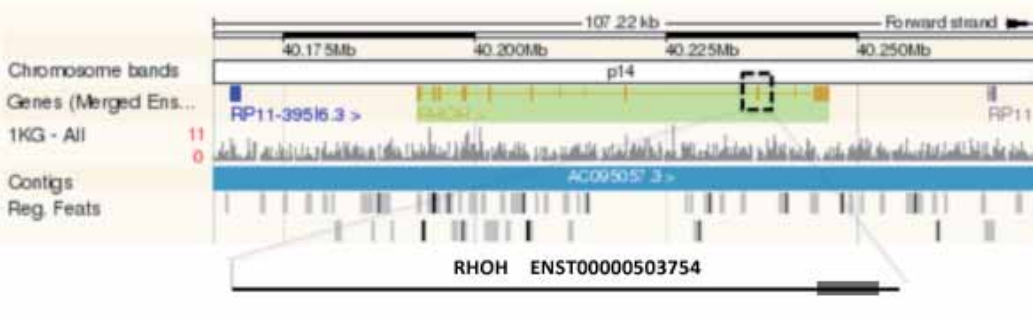


Figure 47 – HsInv0102 locus – Inversion region is depicted by a black dashed box. We observe that the inversion affects an exon of RHOH alternative transcript ENST00000508513, present only in a single isoform (section of the transcript containing the exon depicted in grey). Image adapted from Ensembl v.75 genome browser.

Results of MLPA and PCR genotyping of HsInv0102 in HapMap3 individuals (YRI, LWK, CEU, TSI, JPT, CHB, GIH) show that the inversion is present in all 7 populations with genotype frequencies in Hardy-Weinberg equilibrium, although no evidence for inversion homozygotes has been found in

CHB and JPT populations (Villatoro et al., unpublished results). Computation of fixation index (Fst) scores indicates that HsInv0102 frequency differs significantly across populations and continents (Fst = 0.067, $P < 0.0001$) (M. Gayà, personal communication), having a low frequency (MAF<1%) in Asians (average freq. = 0.035) and being abundant in Africans (average freq. = 0.31), with Europeans showing a low-intermediate frequency (average freq. = 0.15).

Alignment of the HsInv0102 genomic region with other primate species shows that the ancestral allele corresponds to the standard conformation, shared by chimpanzee, gorilla and orangutan. Results from LCL DE analysis at exon level confirm that HsInv0102 inverted allele negatively correlates with the expression of *RHOH* alternatively spliced exon in both CEU+TSI and YRI populations (**Figure 48**). This association has been experimentally validated (Villatoro et al. in preparation) and confirms that the exon is not expressed in HsInv0102 inverted homozygotes. Additionally, 36 genes are affected in *trans*, with logFCs ranging from 0.23 to 2.74.

gene	model	Population	AvgExpr	logFC	p-val	p-val.adj
RHOH	additive	mixed, YRI, CEU+TSI	-2.40	-1.35	9.46E-09	2.88E-03
LHX2	INV-OTHER	mixed	1.29	2.46	4.43E-11	6.62E-07
MAPK13	INV-OTHER	mixed	3.25	1.12	1.86E-08	1.39E-04
KIAA0391	INV-OTHER	mixed	-4.52	1.74	4.37E-08	1.86E-03
SLC23A2	additive	mixed	-2.64	0.78	1.37E-07	2.09E-02
IL13RA1	INV-OTHER	mixed	2.44	1.23	4.07E-07	2.03E-03
DNAJB5	additive	mixed	0.94	0.53	5.35E-07	8.73E-03
SNX25	INV-OTHER	mixed	3.42	-0.47	8.16E-07	1.91E-02
CD44	INV-OTHER	mixed, YRI	8.74	0.84	1.07E-06	3.54E-03
SOX30	INV-OTHER	mixed	-2.78	1.87	1.19E-06	3.54E-03
BLM	additive	mixed	0.19	-0.23	1.52E-06	7.09E-02
CYB5A	additive	mixed	0.63	0.38	1.63E-06	7.09E-02
SERPINB6	INV-OTHER	mixed	0.59	1.56	1.92E-06	4.79E-03
KCNN2	INV-OTHER	mixed	-1.86	2.27	2.64E-06	5.05E-03
FRAS1	INV-OTHER	mixed	-2.42	2.74	2.70E-06	5.05E-03
INTS8	additive	mixed	-0.07	-0.24	2.99E-06	9.10E-02
ZNF595	INV-OTHER	mixed	-2.24	3.70	3.23E-06	3.54E-02
VDR	additive	mixed	-0.90	0.83	3.43E-06	3.17E-02
SLAIN1	additive	mixed	2.09	-0.42	5.36E-06	4.37E-02
KAL1	INV-OTHER	mixed	-3.18	2.11	5.74E-06	9.54E-03
LRRC32	INV-OTHER	mixed	1.96	1.60	8.15E-06	1.22E-02
CIRBP	INV-OTHER	mixed	4.18	0.42	1.07E-05	5.61E-02
CKB	additive	mixed	0.94	0.86	1.58E-05	7.77E-02
CLEC4A	additive	mixed	0.54	-0.44	2.84E-05	6.81E-02
MMP7	additive	mixed	2.52	0.69	3.27E-05	6.81E-02
GALNT10	INV-OTHER	mixed	7.76	-0.62	4.30E-05	4.69E-02
UBASH3B	additive	mixed	1.87	-0.53	4.42E-05	6.81E-02
RIMS3	INV-OTHER	mixed	4.10	-1.02	4.62E-05	4.69E-02
PLCL1	additive	mixed	-0.82	-0.66	4.63E-05	6.81E-02
CXCL12	INV-OTHER	mixed	0.93	1.52	4.91E-05	4.69E-02
SLC35F3	INV-OTHER	mixed	0.80	1.40	5.02E-05	4.69E-02
GPM6A	INV-OTHER	mixed	0.11	1.96	6.06E-05	5.16E-02
ANXA3	additive	mixed	-0.59	0.75	6.08E-05	6.81E-02
PERP	INV-OTHER	mixed	2.46	1.05	6.56E-05	5.16E-02
MYO1F	additive	mixed	2.25	-0.58	8.61E-05	6.81E-02
SEMA5A	INV-OTHER	mixed	0.47	1.81	8.89E-05	5.93E-02
IL12B	additive	mixed	1.23	0.52	1.01E-04	7.21E-02

Table 25 - Top LCL DE analysis candidates - Candidates have been selected according to following criteria: logFC > 0.2. If candidates show association according to multiple models (additive, STD-OTHER (dominant), HET-OTHER (overdominant), INV-OTHER (recessive)) the most statistically significant p-value is chosen, (additive). If candidates show association in multiple populations, values for the pooled population are shown. If candidates show association for multiple exons, scores for most expressed exon are shown. If candidates show association at gene and exon level, gene scores are shown. *RHOH* scores correspond to the inverted exon. All candidates are affected in *trans* with *RHOH* exception in *cis.*

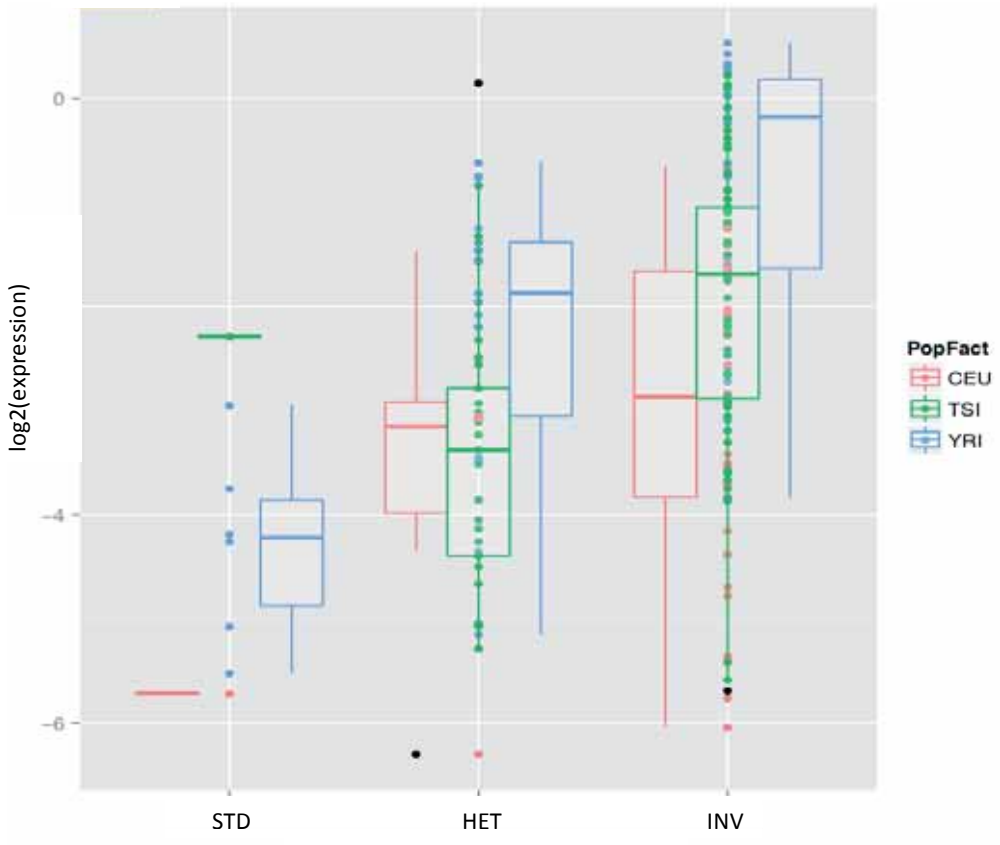


Figure 48 – LCL expression of *RHOH* inverted exon association with HsInv0102 genotype – Number of samples per genotype: 8 (INV), 50 (HET), 116 (STD).

HsInv0059 inversion spans 322 bp and is located in the long arm of chromosome 6 (q15), inside an intronic region of gene *GABRR1*. The inversion was first discovered by comparing the genome of an European individual to the human reference genome (Levy et al. 2007) and was subsequently identified in Asian (Ahn et al. 2009) and African (McKernan et al. 2009) individuals by means of PEM. These two studies also identified HsInv0201, an inversion spanning ~ 400 bp and located in the long arm of chromosome 5 (q32). A deletion associated to the inversion affects gene coding sequence as it contains an exon of the gene *SPINK14* (Martínez-Fundichely et al., 2014).

Both inversions have been characterized and validated as part of the work of the group (D. Vicente and M. Cáceres, unpublished results; Martínez-Fundichely et al., in preparation) and are complex rearrangements as they associate to additional

structural variants. For instance, HsInv0059 appears together with a deletion of 617 bp at BP2 that overlaps a LINE element. In addition, the presence of an additional 7-bp inverted repeat sequence at HsInv0059 BP regions suggest that the rearrangement could have been mediated by mechanisms such as FoSTeS or NHEJ. The ancestral allele corresponds to the inverted rearrangement, which does not contain the deletion (**Figure 49**) (D. Vicente and M. Cáceres, unpublished results).

Figure 49 – HsInv0059 complex rearrangement – Scheme of HsInv0059 region in chimpanzee and HG19 human reference genomes. Inverted region (orange arrow) is flanked by two inverted sequences in HG19 (green arrows) and a deletion at BP2 with respect to chimpanzee genome (red arrow).

HsInv0201 inverted region is flanked by two deletions of 1.2 kb (BP1) and 0.2 kb (BP2) and was probably generated by NHEJ or FoSTeS (Martínez-Fundichely et al., 2014). The ancestral allele corresponds to the standard rearrangement. HsInv0201 is present in all HapMap analysed populations in high frequency (average Inverted Allele Freq. (IAF) = 0.62), albeit less frequently in some populations (IAF JPT freq. = 0.51) than others (IAF CEU freq. = 0.68). On the other hand, HsInv0059 frequency differs significantly across populations and continents. It is low in frequency (derived allele freq. = 0.09) in Africans and abundant in Asians (average derived allele freq. = 0.70) with the exception of Indians (GIH), in agreement with Europeans showing a low-intermediate frequency (average derived allele freq. = 0.18) (S. Villatoro and M. Cáceres, unpublished data). Both inversions are in Hardy-Weinberg equilibrium, although no inverted homozygotes have been found in Africans for HsInv0059. Moreover, HsInv0059 presents strong signals of positive selection in Chinese and Japanese populations (D. Castellano and, M. Cáceres, unpublished data). Both inversions affect the sequence of genes that do not appear to be expressed in blood tissue (**Figure 50**, **Figure 51**), which is in agreement with the lack of differentially expressed genes obtained by the LCL DE analyses for these inverted rearrangements. However, according to GTEx pre-computed results currently (Sep 2014) available in GTEx Portal (see URLs), there is evidence of *SPINK14* and *GABRR1* gene expression changes associated with the genotype of SNPs in high LD with the inversions in non-blood tissues (**Figure 52**).

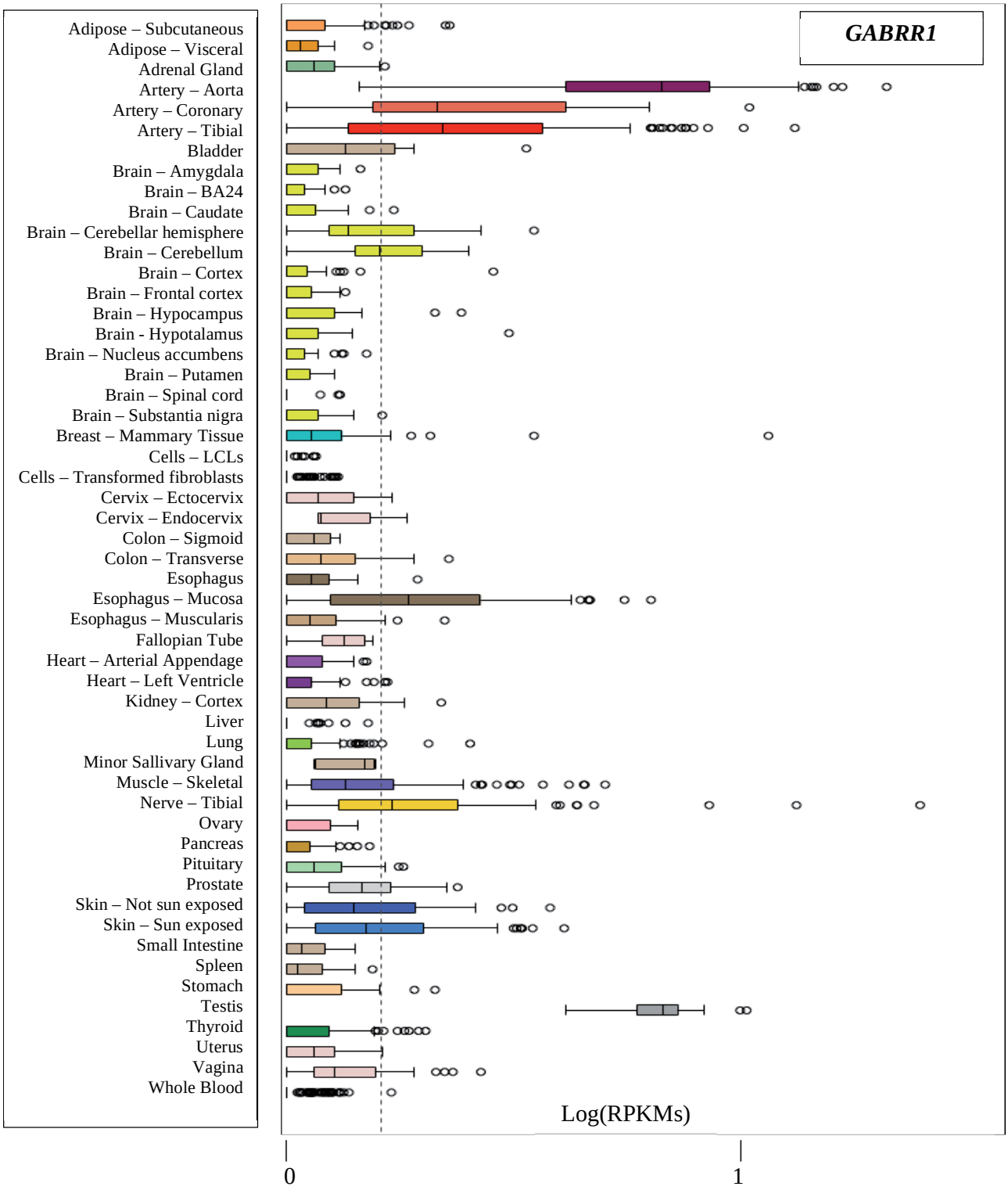


Figure 50 – Expression of *GABRR1* in multiple tissues – Expression values are transformed (see GTEx project documentation). Plots obtained from GTEx database; see URLs)

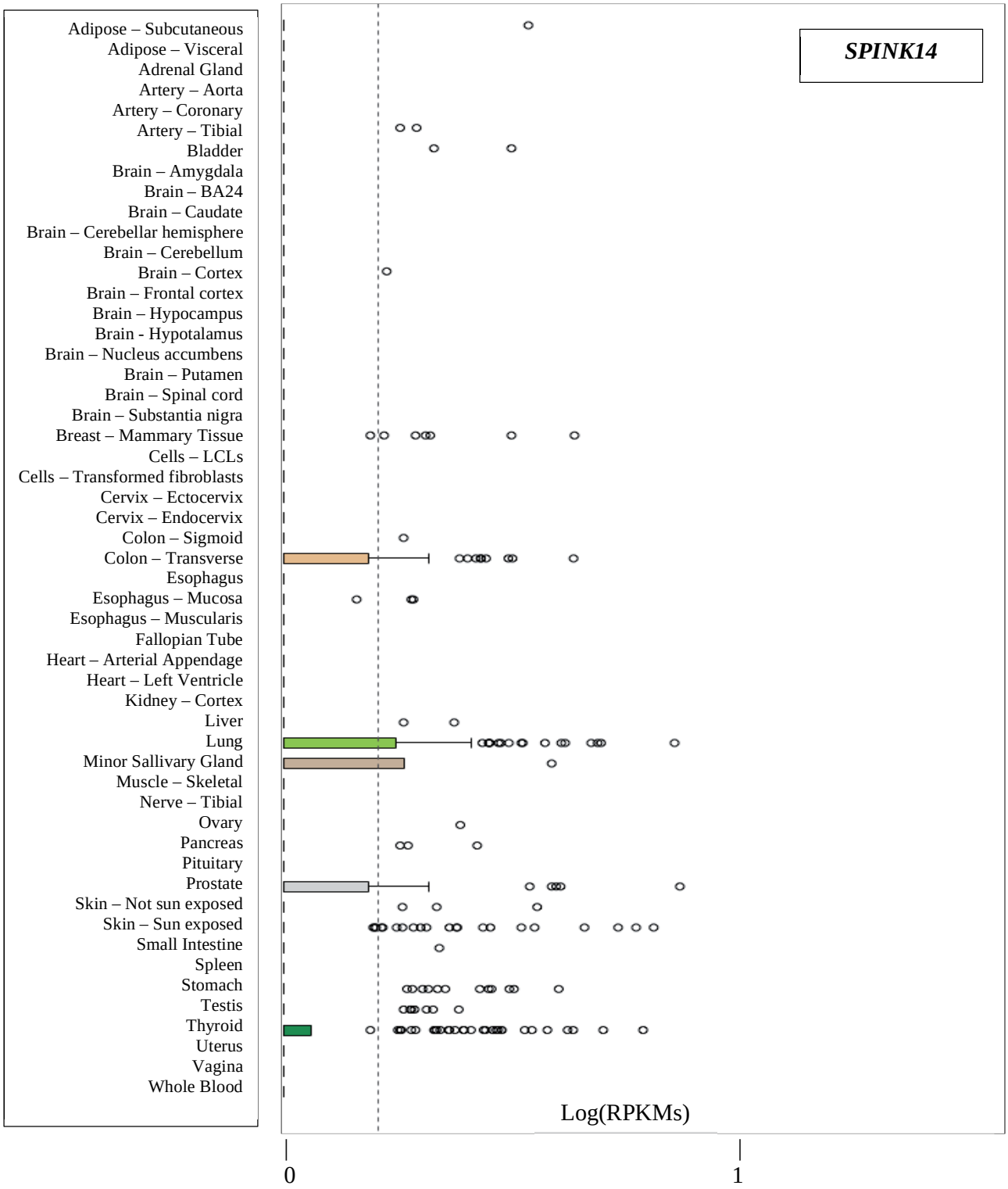


Figure 51 - Expression of *SPINK14* in multiple tissues – Expression values are transformed (see GTEx project documentation). Plots obtained from GTEx database; see URLs).

Based on GTEx data, *SPINK14* is poorly expressed in colon, lung, skin and thyroid. HsInv0201 inverted allele (tagged by SNP rs6864124) seems to correlate with *SPINK14* low expression in thyroid and lung, although the trends are not significant according to GTEx pre-computed results (**Figure 52**).

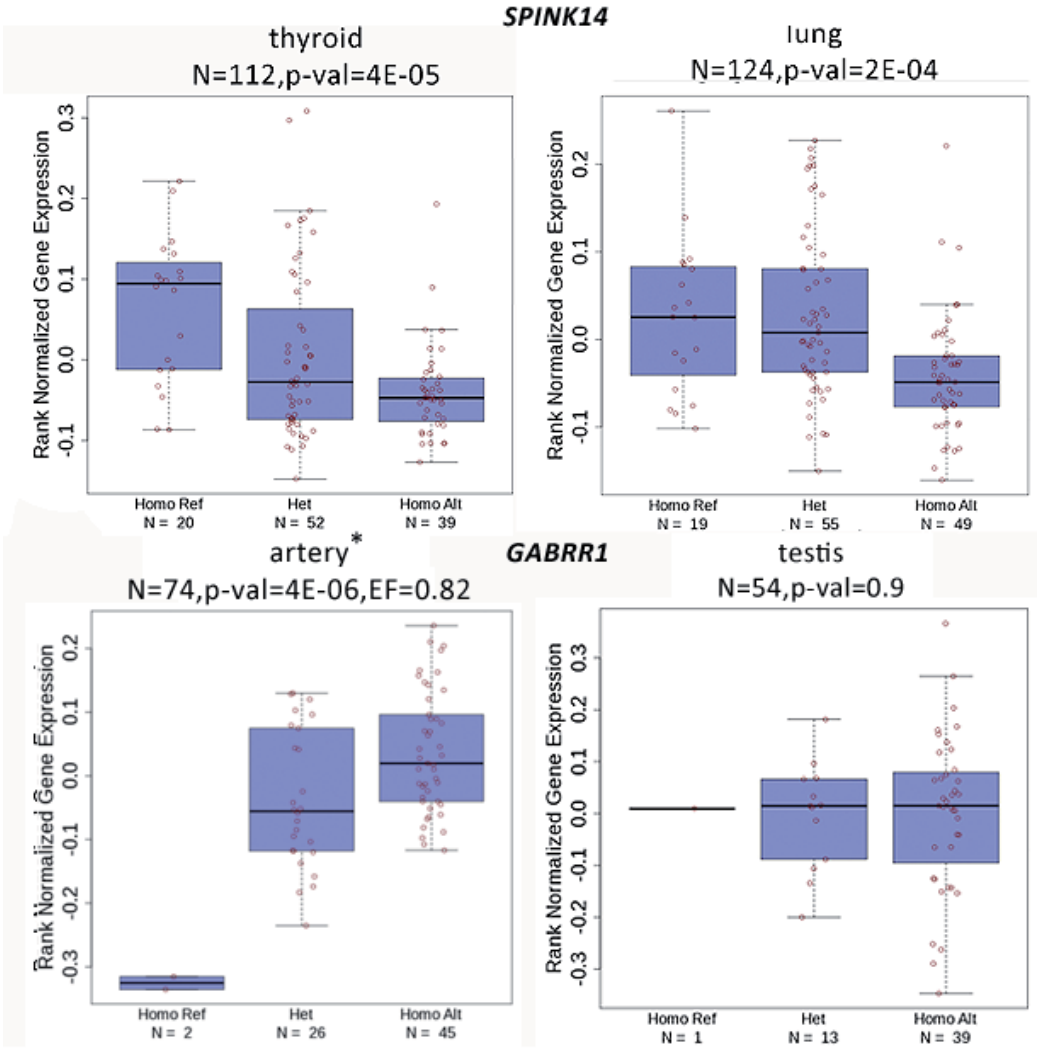


Figure 52 – *SPINK14* ~ HsInv0201 and *GABRR1* ~ HsInv0059 associations in multiple tissues - Associations of *SPINK14* expression with HsInv0201 genotype in thyroid and lung (top panel) derived from tag-SNP rs6864124 and associations of GABRR1 expression with HsInv0059 genotype in artery (tibial) and testis (bottom panel) derived from tag-SNP rs4707529. Number of samples used for the computation of the associations in every tissue (N) and p-value of the association are shown. Number of samples per genotype is shown. Inversion genotypes: "Homo Ref" refers to standard homozygotes, "Het" to heterozygotes and "Homo Alt" to inversion homozygotes Plots obtained from GTEx database (see URLs). Tissues with significant eQTL-gene associations (results obtained from GTEx Portal (see URLs), Sep 2014) are marked by an asterisk, and effect size (EF) of the variant is shown.

Regarding HsInv0059 inversion, *GABRR1* expression is negatively correlated with the inversion derived allele (tagged by SNP rs4707529) in skin, artery and esophagus (**Figure 52**). These associations are significant according to GTEx pre-computed results. However, GTEx analyses are

unpublished to date and can be subject to further modifications and therefore the described results here are not conclusive and need to be interpreted with caution.

Chapter 4

Inversions and disease

SUMMARY - In this chapter, we ascertain the possible associations between inversions and diseases, given that little is known about the role of inverted rearrangements in pathologies or complex phenotypes. To address this issue, we have surveyed a comprehensive compendium of complex human traits and disease phenotypes (67 million association tests for over 1600 GWAS studies) in order to identify variants associated to a certain phenotype that are also inversion tag-SNPs. We have also assessed, when possible, which variants are known to regulate gene expression and in which direction. Finally, the possible causal role and impact of each candidate inversion in the related disease or phenotype is discussed.

4.1 Meta-analyses of disease and complex phenotype GWAS variants associated to inversions

In this chapter, analogously to the inversion-eQTL analysis, we have tried to infer relations between inversions and phenotypes indirectly, by compiling and analyzing associations of inversion tag-SNPs with a given phenotype. Conceptually, both analyses are similar, but they differ in the nature and complexity of studied phenotypes. In the inversion-eQTL case, we examine only a single phenotype (gene expression) that is measured in a quantitative manner at a molecular level. On the other hand, in the analysis based on GWAS data explained herein, the tested phenotypes are multiple, most of the times complex and studied at an organism level. Therefore, the latter analysis constitutes in principle a more challenging scenario than the former. For this and other reasons, we consider the employed strategy described in this chapter as exploratory and the analysis results as preliminary data. Nevertheless, the information generated may be suitable to provide a first glance at the relationships between inversions and disease.

The first step of the analysis consisted on providing the inversion tag-SNP dataset built for the inversion-eQTL analysis (Materials and Methods) as input to `GWAS Central database` (Beck et al. 2014) to retrieve associations of each SNP with diseases or phenotypes. An association was considered significant if the original p-value in the corresponding GWAS study was below a defined threshold. As mentioned, the analysis was meant to be exploratory; therefore we selected the GWAS variants using a non-strict threshold (p-value<0.001) in order to maximize finding possible associations. In the following step, the SNPs of the candidate associations were queried for being genetic determinants of gene expression by crossing the information with the resulting eQTL database build for the inversion-eQTL analysis, and the results were interpreted.

Globally, we find 5 inversions associated (p-value < 0.001) to 15 different phenotypes through 14 SNP variants in LD with the inversion ($r^2 > 0.8$), 11 of which are eQTLs for 12 genes in different populations (**Table 27**, **Table 26**).

Inversion	GWAS Disease / phenotype	SNP	p-val	OR	Ancestry Cases / controls	Reference
HsInv0041	Asthma	rs12694641	3.40E-04	NP	Different ancestries 10,365/ 16,110	Moffatt et al. 2010
HsInv0058	Psoriasis	rs2844645	7.29E-19	NP	NP	NP

Inversion	GWAS Disease / phenotype	SNP	p-val	OR	Ancestry Cases / controls	Reference
HsInv0058	Type I diabetes mellitus	rs2523864	2.54E-14	0.68	European 1,146/563	Hakonarson et al. 2007; Johnson and O'Donnell 2009
	Hypothyroidism	rs2517532	1.3E-08	0.86	European 3,736/35,546	Eriksson et al. 2012
	Stevens-Johnson syndrome and toxic epidermal necrolysis	rs2844665	2.69E-07	1.54	European 1,881/424	Génin et al. 2011
	Drug-induced liver injury due to flucloxacillin	rs3131927	4.82E-07	NP	European 51//282	Daly et al. 2009
	Multiple sclerosis	rs2517538	1.80E-06	NP	European 931/2431	International Multiple Sclerosis Genetics Consortium et al. 2007
	Narcolepsy in a Japanese population	rs2523865	7.62E-06	NP	Japanese 389/222	Koike et al. 2009; Miyagawa et al. 2008
	Ulcerative colitis	rs2517532	5.95E-05	NP	European 6687/19,718	Anderson et al. 2011
	Height	rs2523856	1.09E-04	NP	European 183,727 individuals	Lango Allen et al. 2010
	Pulmonary function	rs2844645	1.69E-04	NP	European 48,201 individuals	Artigas et al. 2011
	Crohn's disease	rs2523857	4.30E-04	NP	European 6,333 /15,056	Franke et al. 2010
	Glycated hemoglobin levels	rs2523857	7.37E-04	NP	European 46,368 individuals	Soranzo et al. 2010
	asthma	rs2523864	7.98E-04	NP	Different ancestries	Moffatt et al. 2010
HsInv0098	Movement-related adverse antipsychotic effects	rs1487569	2.86E-04	NP	Different ancestries 738	Aberg et al. 2010
HsInv0347	Height	rs10148202	1.36E-05	NP	European 183,727 individuals	Lango Allen et al. 2010
HsInv0409	Amyotrophic lateral erosis	rs5916341	1.15E-04	NP	European 276/271	Schymick et al. 2007
		rs1882409	2.47E-04	NP	European 461/450	Schymick et al. 2007

Table 26 - Inversions with tag-SNPs associated to diseases and complex phenotypes - Nominal p-values, as obtained from the corresponding study (GWAS central repository) are shown. If several variants associate to disease, top one is shown. NP: not provided.

					LD SNP ~ inversion (r^2)					
Inversion	**SNP**	**Distance SNP ~ inversion (bp)**	**Distance SNP ~ genes (bp)**	**eQTL gene**	**CEU**	**TSI**	**JPT**	**CHB**	**YRI**	**LWK**
HsInv0041	rs12694641	11166 D	*FAM124B* (37474) *CUL3* (30684)	*FAM124B*	0.89	0.76	1	0.9	0.44	0.6
HsInv0058	rs2844645	5524 D	*PBMUCL1* (12003) *HCG22* (6802)	*CDSN HCG22 HCG27 HCP5 HLA-B TCF19 VARS2*	0.69	0.52	0.9	0.82	0.76	0.88
	rs2517532	8749 D	*PBMUCL1* (15228) *HCG22*	*C6orf27 HCG22 HCG27*	0.83	0.76	0.41	0.35	0.87	0.91
	rs2844665	2803 U	*PBMUCL1* (3676) *HCG22* (15129)	*HCG22 HCG27 HCP5 HLA-B HLA-C PSORS1C3 SFTA2 VARS2*	1	1	1	1	0.96	1
	rs3131927	3338 D	*PBMUCL1* (9817) *HCG22* (8988)		1	0.93	0.9	1	0.92	1
	rs2517538	3883 D	*PBMUCL1* (10362) *HCG22* (8443)	*HCG22 HCG27 HCP5 HLA-B HLA-C PSORS1C3 SFTA2*	1	1	1	1	0.96	1
	rs2523856	11970 D	upstream HCG22	*HCG22 HCG27 HCP5 HLA-B HLA-C PSORS1C3 VARS2*	0.94	0.68	0.42	0.36	0.59	0.66
	rs2844645	5524 D	*PBMUCL1* (12003) *HCG22* (6802		0.69	0.52	0.9	0.82	0.76	0.88
	rs2523857	11846 D	upstream HCG22		0.94	0.67	0.36	0.42	0.4	0.32
	rs2523864	8888 D	*PBMUCL1* (15367) *HCG22* (3438		0.58	0.41	0.51	0.5	0.69	0.81
	rs2523865	8790 D	*PBMUCL1* (15269) *HCG22* (3536)		0.58	0.41	0.51	0.5	0.69	0.81
HsInv0098	rs1487569	5454 D	intronic (ULK4)	-	1	0.83	0.51	0.6	0.74	0.65
HsInv0347	rs10148202	19667 U	*SIX6* (77415)	*SIX1 (adipose)*	0.84	0.84	0.47	0.39	0.41	0.46
HsInv0409	rs5916341	1736 U	intronic (*NLGN4X*)	-	0.85	1	1	1	1	1
	rs1882409	3904 U	intronic (*NLGN4X*)	-	0.58	0.53	1	1	0.22	0.26

Table 27 - Inversion tag-SNPs associated to diseases and complex phenotypes and effects on gene expression – Distances of SNPs with inversions and closest genes reported. Linkage disequilibrium scores (r^2) calculated for a subset of 1000GP samples (90 TSI, 35 CEU, 45 YRI, 78 LWK, 37 JPT and 40 CHB).

4.2 Inversion candidates

HsInv0409 displays a global tag-SNP (rs5916341) for CEU, CHB, JPT, YRI and LWK and a SNP in high LD with the inversion in Asians (rs1882409) that present association (p-value < 0.001) with amyotrophic lateral sclerosis (ALS) in two different GWAS studies (van Es et al. 2007; Schymick et al. 2007). The former study analysed 317,000 unique SNPs in 461 patients with ALS and 450 controls from Netherlands. The latter analysed 555,352 unique SNPs in 276 American non-hispanic patients with sporadic ALS and in 271 neurologically normal controls. Although none of the mentioned variants (rs1882409, rs5916341) were found to be significantly associated to ALS after multiple testing correction in any of the two studies, rs1882409 is the 39th most associated variant, falling in the 0.0068 percentile of tested variants p-value distribution (Schymick et al. 2007) and rs5916341 ranks 53rd, falling in the 0.0167 percentile (van Es et al. 2007). These variants are upstream the inversion (3904 and 1736 bp respectively, with regard to inversion midpoint) and both the inversion and SNPs are located in the same intronic region of *NLGN4X*, a gene that encodes a member of the type-B carboxylesterase/lipase protein family. To date, more than 20 different genetic mutations have been associated to ALS according to OMIM database (see URLs.) but none of them is related to *NLGN4X*. However, *NLGN4X* has been found to be associated with a wide spectrum of neuropsychiatric disorders such as X-linked autism-2 syndrome and X-linked mental retardation (Lawson-Yuen et al. 2008).

We find that HsInv0058 associates to 13 different diseases and complex phenotypes. However, not all variants in high LD with HsInv0058 associate equally to the phenotypes (**Table 28**).

HsInv0058	rs2844665	rs3131927	rs2517538	rs2523865	rs2517532	rs2523856	rs2844645	rs2523857	rs2523864
LD (Europeans)	**1**	**0.94**	**0.98**	**0.40**	**0.77**	**0.69**	**0.52**	**0.69**	**0.4**
Type I diabetes mellitus					**1.33E-10**				**2.54E-14**
Psoriasis					0.398		**7.29E-19**	**2.95E-10**	
Hypothyroidism					**1.3E-08**				

HsInv0058	rs2844665	rs3131927	rs2517538	rs2523865	rs2517532	rs2523856	rs2844645	rs2523857	rs2523864
Stevens-Johnson syndrome and toxic epidermal necrolysis	**2.69E-07**								
Drug-induced liver injury due to flucloxacillin	1.14E-06	4.82E-07			0.039		6.34E-7		0.075
Multiple esclerosis			1.80E-6						
Narcolepsy*	7.62E-06								
Ulcerative colitis	0.384		0.042	0.3	5.95E-05		4.02E-02	9.56E-02	0.327
Height	3.22E-04	1.76E-04	1.92E-04	0.171	1.09E-03	1.08E-04	0.320	1.99E-04	0.139
Pulmonary function	0.052	0.029	0.025	6.8E-04	0.077	0.173	1.68E-04	0.128	5.7E-04
Chron's disease	1.2E-03	7.1E-04	0.044	0.011	0.18		6.2E-02	4.3E-4	65E-02
Asthma	0.42				0.373				7.98E-04

Table 28 – HsInv0058 linked variants associated to disease and complex traits – Associations of variants in high LD with HsInv0058 to complex traits or diseases are shown. In bold, p-values for variants reported as significant in corresponding study. LD scores (r^2) calculated for a subset of 1000GP European samples (90 TSI and 35 CEU individuals), as all the studies are derived from samples with European ancestry, with the exception of narcolepsy GWAS study (carried out in Japanese population). *P-value corresponds to GWAS Allelic test (HGVRS176), out of the 5 different association tests carried out in this GWAS study.

A hyperthyroidism GWAS (Eriksson et al. 2012) reports 5 significantly associated variants (OR = 1.36~0.78, p-value = 1.3E-8~2.4E-19), in 4 different chromosomes. Among these 5 variants, the least significant one (rs2517532, p-value = 1.3E-8, OR = 0.86) locates in HLA region as part of MHC complex and is in moderate LD (r^2 = 0.77) with HsInv0058 in Europeans. The study also reports a second HLA variant (rs2516049) showing only suggestive evidence of association (OR = 1.15, p-value = 6.0E-7) independent of rs2517532 effect. However, rs2516049 is not in LD with HsInv0058 and falls out from the inversion *cis* region (> 1,668 kb).

For Stevens-Johnson syndrome and toxic epidermal necrolysis (TEN) study (Génin et al. 2011) all the 6 reported top SNPs locate in the HLA region and are significantly associated to the disease (top variant = rs9469003, OR = 1.73, p-value = 1.6E-9). However, the top 5 variants are not in LD with HsInv0058 in any studied population. The 6th most associated variant (rs2844665, OR = 1.54, p-value = 2.69E-

7) is a global tag-SNP for HsInv0058 in non-African populations and locates 2,803 bp upstream of the inversion midpoint.

Regarding drug-induced liver injury due to flucloxacillin phenotype, Daly et al. (2009) report a SNP from the MHC class I region, rs2395029, being strongly associated with the phenotype (OR = 45, p-value < 1E-30). The SNP is in almost complete LD with *HLA-B*5701* and is not in LD with HsInv0058 in any analysed population. Although previous studies report the existence of other HsInv0058-linked MHC alleles with suggestive evidence of association with the phenotype, a conditional analysis indicated that no other SNP in MHC region presents a genome-wide significant association independent of rs2395029 effect. Therefore, despite the significant associations with several SNPs linked to the inversion, the causal role of HsInv0058 on this disease can be rejected.

Multiple sclerosis GWAS study by the International Multiple Sclerosis Genetics Consortium (2007) reports several non-MHC variants associated to the disease, such as rs12722489, located in chromosome 10p15, precisely in intron 1 of the *IL2RA* gene (OR = 1.25, p-value = 2.96E-8). Regarding MHC variants, HLA-DR locus tag-SNP rs3135388 is unequivocally associated with the disease (OR = 1.99, p-value = 8.94E-81). All other SNPs falling in MHC region (positions between 29 and 34 Mb on chromosome 6) were analysed conditional on HLA-DR haplotype linked to rs3135388 (HLA-DRB1*1501) and showed a residual association signal peaking at rs9270986 (p-value = 1.83E-17), which lies close to DRB1 locus and is unlinked to HsInv0058. The only reported MHC-variant with potential association with the disease and in LD with HsInv0058 is rs2517538 (p-value = 1.8E-6). However, this variant is not reported as significant and is not even part of the list of the top 100 reported variants of the study. Therefore, HsInv0058 seems to be unrelated to multiple sclerosis.

A narcolepsy GWAS study (Miyagawa et al. 2008) reports many (> 100) MHC variants associated to the disease and also some associated non-MHC variants, such as rs5770917, a SNP located between *CPT1B* and *CHKB*, which is associated with narcolepsy in Japanese (OR = 1.79, p–val = 4.4E-7) and other population groups such as Koreans, Europeans and African Americans (OR = 1.40, p-value = 0.02). The only reported MHC-variant with potential association with the disease and in LD with HsInv0058 is rs2523865 (p-value = 7.62E-06). This variant is not reported as significant and is not even part of the list of the top 100 reported variants of the study. In addition, the SNP is in high LD (r2 = 0.81) with HsInv0058 only in LWK population, not in Asians (r2 ~ 0.5). All together, these evidences suggest that HsInv0058 is not associated with narcolepsy.

A compilation of ulcerative colitis GWAS studies (Anderson et al. 2011) identified 47 loci associated to the disease, but none of them belongs to MHC complex. In fact, rs2517532 (p-value = 5.95E-5) is the only MHC variant reported in the study that passes our association threshold (p-value < 0.001) and is in moderate LD with HsInv0058 in Europeans. Therefore, a putative association of HsInv0058 with the disease is very unlikely.

A massive height GWAS study (183,727 individuals, 2,834,208 markers) (Lango Allen et al. 2010) reported at least 180 loci that influence adult height. Among them, 3 (rs3129109, rs2256183, rs6457620) are part of the MHC complex but none of them is in high LD with HsInv0058, and the closest variant to HsInv0058 (rs6457620, p-value = 3.60E-8) is located ~ 1,762 Mb downstream of the inversion midpoint. Moreover, rs2523856, which is in moderate LD with HsInv0058 in CEU (r^2 = 0.94) but not in TSI (r^2 = 0.68) is 4 orders of magnitude (p-value = 1.08E-4) less associated than the mentioned HLA variant rs6457620. Therefore, HsInv0058 seems to not play a role in modulating height phenotype.

Pulmonary function measures like tested forced expiratory volume (FEV) reflect respiratory health and are used in the diagnosis of chronic obstructive pulmonary disease. Soler Artigas et al. (2011) measured FEV in 48,201 individuals of European ancestry with follow up of the top associations in up to an additional 46,411 individuals. Among all variants linked to inversions, rs2844645, in moderate LD with HsInv0058 in Europeans (r^2 = 0.52) is the most associated one to FEV capacity, but the association is mild (p-value = 1.68E-4) and not significant. In fact, there is only one MHC variant (rs2857595, p-value = 2.28E-10) reported to be significant in this study, but it is not linked to HsInv0058. Another variant, rs9325087, is in high linkage with HsInv0201 (r^2 = 0.98 in Europeans) but the association is poor (p-value = 9.99E-03) and does not pass our association threshold. Therefore, neither HsInv0201 nor HsInv0058 seem to play a role in modulating pulmonary function FEV phenotype.

A Crohn's disease GWAS study (Franke et al. 2010) reports 71 associated variants, although none of them is located in MHC. Variant rs2523857 (p-value = 4.3E-4) is in moderate LD (r^2 = 0.69) with HsInv0058, although lays far from the study significance threshold (p-value < 5E-8). Therefore, HsInv0058 seems to not play a role in Crohn's disease.

An asthma study (Moffatt et al. 2010) reports *HLA-DQ* variant rs9273349, unlinked to HsInv0058, as the top MHC associated variant (p-value = 7E-14). Variant rs2523864, in low LD with HsInv0058 in Europeans (r^2 = 0.4), presents a mild, not

significant association (p-value = 7.9E-4) with the disease, excluding HsInv0058 causal role.

Finally, a study of movement-related adverse antipsychotic effects (Aberg et al. 2010) reported a mild association (p-value = 2.86E-4) for phenotype of tardive dyskinesia (TD) with variant rs1487569, in high LD with HsInv0098 in Europeans. However, overall the study reports no variants associated to TD with genome-wide significance.

V DISCUSSION

Refining the catalogue of human polymorphic inversions

Inversions constitute one of the most difficult types of structural variants to be predicted due to their particular genomic features such as balanced rearrangement status and frequent presence of big inverted repeats at the breakpoint regions. The need to develop approaches particularly suited to predict inversions guided the development of GRIAL, a novel method in our lab (Martínez-Fundichely et al., in preparation) that predicts inversions by interpreting the specific signatures that this type of SV presents in PEM data. Compared to similar PEM based methods, GRIAL has proved to be an accurate algorithm particularly suitable for inversion prediction.It has shown good performance in: a) obtaining low false positive rates; b) detecting inversion regions efficiently by minimizing the compatible PEMs that are clustered together and the number of reliable predictions per locus; and c) resolving BPs with high accuracy, where it clearly outperforms other methods.

The good results produced by GRIAL can be partially attributed to the particular characteristics of the PEM dataset employed, as big insert size of the template and length of sequenced paired-end reads, combined with filtering out of low quality reads constitute factors that lower the false positive rate in inversion prediction (Lucas-Lledó and Cáceres 2013). For this reason, the performance of GRIAL should be tested in a more challenging scenario such as a real next generation sequencing based PEM dataset with smaller insert size and shorter reads than the fosmid PEM dataset (Kidd et al., 2007). Moreover, the benchmarking of PEM methods could be complemented with an additional analysis providing as input PEM datasets with multiple mappings per paired-end, as some probabilistic methods have been designed to accommodate these features and therefore could perform better (e.g. VH, GASVPro). Despite all that, a benchmarking in a simulated data set shows that GRIAL also performs better than other methods (Lucas-Lledó and Cáceres 2013).

It is important to note here that although relevant, the method employed in predicting the inversion is only one of the many factors that generate false positive predictions. Certain genomic features such as inverted duplications in tandem, evolutionary events such as gene conversion between paralogous IRs or technical artifacts in the assembly of the reference genome like mixing of two haplotypes, gaps

or other errors, produce spurious PEM signals that can be misattributed to inverted regions (Aguado et al. 2014; Antonacci et al. 2010; Kidd et al. 2008; Lucas Lledó and Cáceres 2013; Martínez-Fundichely et al., in preparation, Vicente et al., unpublished results). In this work, additional sources of errors have been detected, some of them for the first time. In a few cases, we observe that false positive predictions arise from events that occur during the paired-end library generation, such as fosmid chimeras and amplification of artefactual products due to primer misspriming before the amplification step (see section IV1.2). The misspriming artifact is an example of a typical problem of high-throughput techniques, in which a small error in a single step of the procedure generates a significant number of false positives. The identification of these artifacts allowed us to increase the accuracy of the generated inversion catalogue by filtering 14.6% (93/636) of `GRIAL` inversion predictions. In addition, the veracity of a considerable number of inversion predictions is dubious due to supporting paired-ends mapping multiply to several locations with similar identity. Thus, our strategy based on remapping and scoring the inversion-supporting paired-read set allowed us to categorize inversion predictions according to their level of reliability and therefore filter 50% (318/636) ambiguous predictions to further increase the reliability of our inversion catalogue. However, less than a dozen cases identified by the remapping strategy to lose support on HG19 can be explained by the alternative inverted allele represented in this more recent assembly compared to HG18. Therefore, in these cases the prediction may be correct and it may have been wrongly categorized as unreliable.

Nevertheless, in some cases (principally when inversions are located in highly complex genomic regions), even the usage of an accurate inversion prediction method combined with appropriate filtering and scoring strategies may fail to produce correct predictions,. We previously described a case for this scenario (Aguado et al. 2014), where manual inspection of PEM signatures in a re-assembled region in the human genome coupled with customized paired-read filtering was necessary to assess the veracity of an inversion prediction with BPs overlapping not only big IRs but also an additional polymorphic indel.

Finally, we want to emphasize the importance of having cataloges of curated human structural variants in order to make reliable inferences on the variants. In this work, we dedicated effort to refine an inversion catalogue to accurately predict the functional effects of inverted alleles. For that, it is of special importance to analyse real inversion predictions that overlap or locate at the vicinity of genes. However, there exists a high false positive rate of inversion predictions within this class. Out of

the 306 inversion predictions overlapping genes (either exonic or intronic regions) available in InvFEST (Martinez-Fundichely et al. 2014), 7.5% turn to be false predictions, 31% are categorized as unreliable according to the GRIAL scoring system (Martínez-Fundichely et al, in preparation) and only 7.5% correspond to real inversions (**Figure 53**). If we consider the 119/306 predictions that break genes (one inversion BP overlaps 5' or 3' genic sequence, including exons), we observe similar numbers: only 7.5% have been validated and 54% are considered unreliable predictions. If we consider only the curated inversion set (discarding predictions without further validation), the percentage of validated predictions is 16% (**Figure 53**).

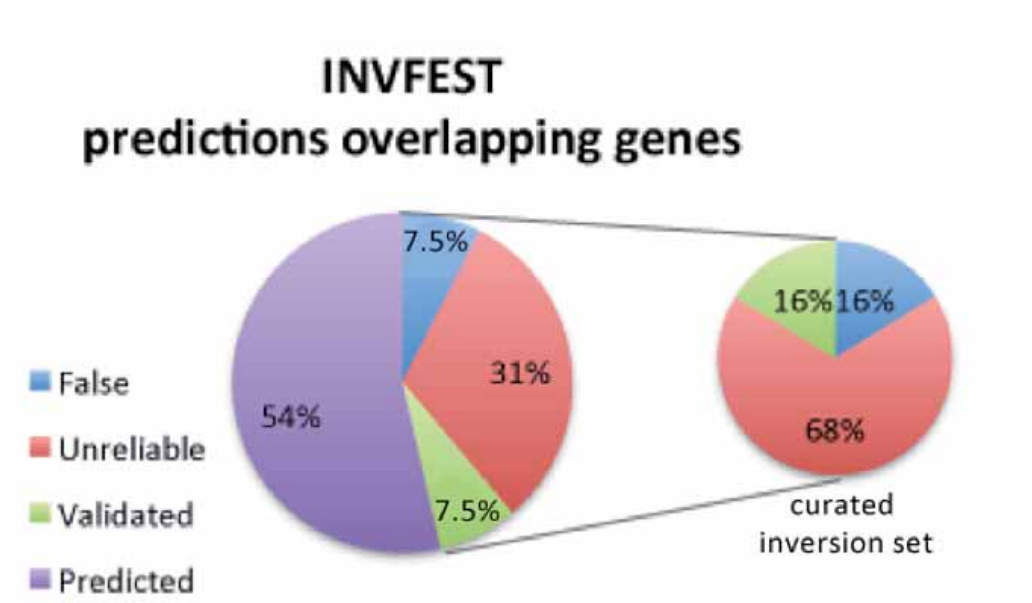


Figure 53 - INVFEST predictions overlapping genes – Classification of INVFEST gene-overlapping inversion predictions.

Methodology

The different approaches (LCL DE analysis and inversion-eQTL) employed here to investigate the effect of inversions on gene expression have shared specific advantages and disadvantages. In both strategies, we have analyzed publicly available expression datasets, which included microarrays and RNA-Seq expression

data. Microarrays are a mature technology and therefore the limitations of this technique are known and taken into account in a large catalogue of available algorithms to analyze the produced data, particularly for differential expression analysis (e.g. `limma`). On the other hand, RNA-Seq is a more recent technique and the most appropriate methods to perform DE analysis, even at the level of the selection of the most suitable distribution to fit the counts, is nowadays a topic of discussion in the field. However, microarrays have major limitations compared to RNA-Seq. First, the probes may contain SNPs that can result in false positive DE associations, so genes in highly divergent genomic regions such as MHC cannot be analyzed. Second, the threshold for detecting low expressed genes is higher than RNA-Seq. Third, the expression can only be measured at a gene level, not at a transcript level (with few exceptions of exon microarrays). And fourth, the tested gene set is biased for protein-coding genes compared to non-coding. For all these reasons, RNA-Seq expression data is preferred over microarray expression data to carry out any kind of gene DE analysis.

We conducted a DE analysis on HapMap LCLs for various reasons. First, the convenient logistics of acquiring and maintaining these commercial cell lines and the ease of obtaining enough genomic material for the DNA intensive- inversion genotyping techniques employed (Aguado et al. 2014) facilitated obtaining the inversion genotypes in these individuals. Second, because three expression datasets derived from these cell lines are publicly available. As we genotype each inversion *in situ* in the same samples, this DE analysis has the advantage of not depending on surrogate SNPs to investigate inversion effects on gene expression and so in theory one could study any candidate inversion. In practice however, we encounter technical problems in genotyping inversions with big (>20 kb) IRs at the BPs. In addition, the main limitation of the approach (as it has been conducted here) is that only gene expression changes in blood, specifically in lymphocytes, can be analyzed. Moreover, the EBV-transformation that lymphocytes undergo may affect their gene expression patterns and confound the analysis.

On the other hand, the other strategy employed (identification of inversion-eQTL associations) is a powerful, cheap and fast approach to scan for inversion effects in non-blood tissues. One of its main advantages is that it avoids some laborious, logistically challenging and expensive steps that a direct analysis (e.g. LCL DE analysis) requires: the collection of samples of tissues of interest and the genotyping of inversions *in situ*. However, important limitations are also present as only unique inversions with SNPs in high LD can be analysed: recurrent inversions

often lack tag-SNPs because polymorphisms located inside and close to inversion BPs are shared by both inverted and standard alleles. In addition, there is a dependence on the available eQTL studies, which may not explore a particular tissue of interest. We observe that the majority of eQTL studies suffer from several flaws and biases, like sex chromosomes being consistently overlooked for eQTLs and a predominant search for effects in *cis* compared to *trans*, with notable exceptions (Fairfax et al. 2014; M. N. Lee et al. 2014). The analysis of sex chromosome eQTLs lagging behind autosome chromosomes eQTLs creates a problem for the study of inversion effects on gene expression in an unbiased manner, because a considerable percent (8/44, 18% in our dataset) of human inversions are located in chromosome X (due to its enrichment in duplicated regions that include IRs).

Similarly, we have tried to investigate the relationship of inversions with diseases by identifying significant inversion surrogate SNPs in publicly available GWAS studies. This approach has several advantages. Importantly, it is costless and fast, as there is no need to carry out any GWAS study. In addition, as we do not have any *a priori* hypothesis about the possible association of an inversion with a particular disease, the exploratory nature of the analysis fits the purpose of testing as a wide spectrum of phenotypes. However, just as in inversion-eQTL analysis, this method has many pitfalls. As previously mentioned, only inversions with tag-SNPs can be tested. It is known that inversions mediated by NAHR are recurrent (Aguado et al. 2014) and therefore not linked to a particular haplotype. It is also known that most of the inversions originate by NAHR (Pang et al. 2013). Therefore, the approach can only be used for the non-recurrent, IR-free inversions with tag-SNPs, which are a minority. In addition, not all IR-free inversions may have tag SNPs, which reduces even more the candidates. Moreover, population stratification may be also a problem. A GWAS study is usually carried out on individuals from one population, with particular haplotypic structures. Therefore, a surrogate SNP for an inversion in a particular population may be monomorphic or low in frequency in the studied GWAS population, and therefore filtered out, which precludes the possibility of studying the inversion effect. In some other cases, the putative disease causal/associated variant is not a tag-SNP for the inversion in the population of interest. In these occasions the causal role of the inversion in the disease is not plausible. HsInv0058 association with narcolepsy (see section IV4.2) exemplifies this scenario, as the disease variant (rs2523865, p-value = 9.35E-06) has been found to be associated to narcolepsy in Japanese, but is poorly linked (r2? ~ 0.5) with HsInv0058 in this population, although is in perfect LD in some African populations (LWK).

An additional problem of the approach is the selection of the threshold to consider a GWAS SNP association significant, as different studies use different methods and statistics, which makes the strength of associations with disease for different inversions poorly comparable. Here, we have selected a non-stringent threshold (original p-value < 0.001) to obtain preliminary results, favoring sensitivity but compromising specificity and expecting false positive associations. In fact, we observe that all but one of the putative inversion-disease associations based on our selected significance threshold are considered to be non-significant in the original GWAS study (see section IV4.1).

Functional impact of polymorphic inversions on gene expression

The objective of the different DE analysis performed in this work is to identify inversions affecting the expression of genes far and in the vicinity of the inverted genomic region. This analysis presents both technical and biological challenges. From a biological point of view, given the exploratory nature of the analysis the main concern is the uncertainty regarding the expectation of findings. That is, the probability of each one of the studied inversions to have an effect on the measured molecular phenotype (gene expression) is a priori unknown. However, according to a parsimonious scenario, it is reasonable to hypothesize that inverted rearrangements including, overlapping or at the vicinity of genes have more chances to affect gene expression than the ones located in intergenic regions. As discussed in the introductory chapter, it is known that inversions can have positional effects on genes by translocation of regulatory elements such as promoters, enhancers or insulators; they can also alter the gene coding sequence; and, in the most severe case, an inversion can disrupt the gene exerting a detrimental effect and ultimately causing disease.

Therefore, from the results of the analysis of the gene content of the inverted regions, one can expect: a) a paucity of inversions overlapping or adjacent to genic regions; b) a paucity of inversions altering coding compared to non-coding sequence; and c) more effects in *cis* than in *trans*. Indeed, this is what we observe: most of the inversions of the analyzed set (23/44, 52.27%) locate outside gene domains. In addition, among the set of inversions overlapping genes in different manners, only 6

inversions affect gene exons, which is just 13.6% (6/44) of the total set. In addition, this figure may be on the higher end, because a) we favored the selection of inversions adjacent or breaking genic sequences compared to others located in intergenic regions, because of their potential functional impact and b) we considered overlaps between inversions and the entire Ensembl v.75 gene annotation set (we did not filter for unreliable or putative, non-validated transcripts). A list all transcripts considered to overlap with the 44 inversions is provided (**Table 29**, Appendix), including information about the reliability level of the isoforms (putative, known) and other additional features.

Inversions have characteristic features that may challenge the view of an effect of the rearrangement on genes when breaking them. For instance, in the case of NAHR-mediated inversions, even if it is clear that the inversion BP overlaps the coding sequence of the gene, this event may have no effect at the gene sequence and expression level. This is illustrated in 11 instances of the 44 inversion set, in which inversion BPs overlap genes situated partially or totally within IRs, with identical or very similar sequence, and only in one case (HsInv0124) we detected DE of the potentially broken gene associated with the inversion. In 4/11 cases (HsInv0393, HsInv0344, HsInv0278, HsInv0374) the entire sequence of the paralogous genes or pseudogenes is identical and contained in the IRs (see section IV2.1). In these cases, even if the predicted inversion BP regions overlap with described genes, the effect of the inversion is expected to be null, because the complete exonic sequences of the genes are completely identical in the two IRs implicated in the inversion. Therefore, if there is an exchange between the IRs, the sequence of the mRNA will not be affected. In agreement with this hypothesis, we have not found differential expression of genes associated to the inversion genotype in any of these cases. In 3 other cases (HsInv0241, HsInv0396, HsInv0209) the inversion overlaps with genes contained totally in the IRs but differing in their sequence. For instance, in HsInv0241, the inversion breaks two aquaporine genes: *AQP12A* and *AQP12B*. These genes are 99% identical in their CDS, and the inversion exchanges the 5' exon, that differs in 3 positions in the UTR and two positions in the CDS. In humans, both genes seem to be expressed mainly in pancreas and retina, and they have been associated to pancreas and retina-related diseases (MalaCards, see URLs). Although expression data of these genes in pancreas is available (GTEx, (Fagerberg et al. 2013)), we have not been able to analyze any expression change that could be attributable to the inversion, as it lacks tag-SNPs and therefore was excluded from inversion-eQTL analysis. In this case, a conservative approach that can be carried out to study the association of HsInv0241 with *AQP12A* and *AQP12B* expression is to

directly genotype the inversion in the pancreatic cells of a set samples for which expression data is available, and perform a differential expression analysis.

In the HsInv0396 case, the inversion BPs overlap the *PABPC1L2A* and *PABPC1L2B* polyadenylate-binding protein genes, which are 100% identical in their CDS, 98% at the mRNA level, and are expressed mainly in brain. However, as the inversion BPs have not been refined due to lack of inverted orientation sequence, the inversion putative functional effect is not clear. Similarly to HsInv0241, the inversion effect on the putatively broken genes could not be investigated by means of the LCL DE analysis because *PABPC1L2A* and *PABPC1L2B* are not expressed in lymphocytes, and the inversions was not included in inversion-eQTL analysis because it lacks tag-SNPs.

In the case of HsInv0209, the inversion is located within two inverted SDs of 7 kb and 94% average identity that include completely the *KRTAP5-10* and *KRTAP5-11* genes (85% identity). Similar to the HsInv0241 case, the BPs of the inversion have not been refined, which makes the prediction of the inversion functional effects difficult. Moreover, the expression of both genes in the 3 LCL expression datasets analyzed (Stranger 2007, Stranger 2012, or Geuvadis) is null and none of the tissues with expression data available (GTEx or Illumina Body Map 2.0, (Fagerberg et al. 2013)) showed median gene expression higher than 0.3 RPKMs. However, if we look at skin tissue expression data (GTEx), we observe bigger variation of the expression values range compared to the rest of the ~50 tissues. This is not surprising, as *KRTAP5-10* and *KRTAP5-11* interact with hair keratins intermediate filaments, so they are expected to be expressed in hair follicle stem cells. This kind of stem cells may constitute a unique type and a small percent of all skin tissue cells collected in GTEx project. In addition, *KRTAP5-10* and *KRTAP5-11* may express at different levels during the hair cycle, a process whereby the hair follicle goes from dormant to having its stem cells activated, making the hair grow. Therefore, all these factors may explain the variability of expression in this tissue and make it difficult to study the effect of a genetic variant. Additionally, HsInv0209 does not have a tag-SNP, so one cannot use any surrogate polymorphism to indirectly genotype the inversion. We suggest that to perform a good differential expression analysis of HsInv0209 on *KRTAP5-10* and *KRTAP5-11* expression, the inversion BPs should be refined, the inversion should be directly genotyped in the samples, and the gene expression should be measured only in hair follicle stem cells, isolated from other skin cells.

In 4 other cases (HsInv0124, HsInv1051, HsInv0030 HsInv0389) inversion BPs overlap with genes contained partially in the IRs but with part of their CDS or transcript sequence located outside the IRs, either for one extreme of the gene (5' or 3') or both. If the IRs contain, in part, two highly similar genes of the same family, with a unique extreme (either 5' or 3') of genic sequence located outside of the repeats at the inversion BPs, the inversion is just expected to alter the relative position of the genes without altering the mRNA sequence. However, the inversion could still affect gene expression levels by altering the relative positions of the genes with respect to distal regulatory elements located outside or inside the inversion locus.

Another possible scenario is an inversion altering two paralogous genes overlapping inversion BP IRs, but with unique sequence located both inside and outside the inverted region. For instance, HsInv1051's leftmost BP (BP1) overlaps the gene *CCDC144B* that has CDS at both sides of the BP IRs, so this case could potentially be n example of a gene disruption caused by an inversion. *CCDC144B* gene spreads over 87.8 kb, with several coding exons at both sides of the SD implicated in the inversion and the inversion moves the 2 first exons 200 kb away from the rest. In addition, the sequence of 2 internal exons of different alternatively spliced products of *CCDC144B* could also be modified by the inversion as these exonic regions are not identical between the two IRs. *CCDC144B* is part of a family with two other members, *CCDC144A* and *CCDC144CP*, with >99.1% identity and very similar exon-intron structure. Although there is evidence of *CCDC144A* at a protein level (Kim et al. 2014), its possible function is unclear and *CCDC144B* and *CCDC144CP* are considered to be pseudogenes. Nevertheless, in GTEx, *CCDC144B* is expressed in multiple tissues such as testis, ovary, bone marrow and predominantly brain cerebellum (Fagerberg et al., 2013), and *CCDC144B* isoforms are supported by split-reads from different sources (e.g. Illumina Body Map 2.0 or Geuvadis projects). Moreover, there exist alternatively spliced products lacking the first two exons, in agreement with the inversion effect (e.g. EST sequence BX371424), although they may correspond to incomplete transcript sequences. We do not report any association of the inversion with *CCDC144B* expression, but as in the previous cases, there is no tag-SNP for the inversion and therefore this case has not been studied by means of the inversion-eQTLs association. HsInv1051 BP2 overlaps with two additional genes, *PRPSAP2* and *AC107982.4*. However, in these cases the inversion is not expected to cause any functional impact: the exonic *PRPSAP2* sequences overlapping the BP2 region are identical between the two IRs; and *AC107982.4* is

considered to be a pseudogene, although there is some evidence that the affected isoform is expressed in esophagus, colon and testis (GTEx expression data).

In the case of HsInv0030, the inversion inverts the promoter and the 5' exon of genes *CTRB1* and *CTRB2*, members of the chymotrypsinogen B precursor of the digestive enzyme chymotrypsin (whose function is to cleave aromatic amino acids such as phenylalanine, tyrosine, and tryptophan). Considering the canonical isoforms, the genes are very similar (97% sequence identity) but differ considerably in the 5' (82% identity) and 3' exons. In addition, the genes code for 6 additional alternatively spliced transcripts, so the inversion can produce novel isoforms by combining exons from both genes. This hypothesis was confirmed by Pang et al. (2013), where the authors characterized the inversion rearrangement and found an alternative inverted haplotype associated with a frame-shift 585-bp deletion that causes the excision of *CTRB2* exon-6 and disrupts the trypsin-like serine protease domain, therefore probably deriving in a *CTRB2* non-functional protein. *CTRB1-CTRB2* gene fusions (including the exon-6 deletion) caused by HsInv0030 are supported by Genbank RNA databases sequences and ESTs. Furthermore, Pang et al. (2013), studied the inversion and inversion-deletion allele frequency in 71 individuals from 57 populations from the HGDP-CEPH Human Genome Diversity Panel, and concluded that there was evidence of significant population differentiation in haplotype frequency, in agreement with our results of HsInv0030 frequencies genotyped by MLPA in 550 HapMap samples. Pang et al. (2013) also investigated the ancestry of HsInv0030 inversion in primates and found that the ancestral state corresponds to the inverted conformation (*Inv*).

According to ENCODE data, there is a denser cluster of transcription factor binding sites and promoter-associated histone marks upstream of the exon-1 of *CTRB1* compared to *CTRB2*, therefore HsInv0030 inversion may alter the gene expression patterns in addition to the creation of hybrid protein structures. Both genes are highly expressed in pancreatic islet cells and several orders of magnitude less in other organs (**Figure 60**, **Appendix**). Despite all efforts, the inversion effect at modifying the expression levels of *CTRB1* and *CTRB2* has not been characterized yet. To investigate that, we have very recently performed an analysis making use of GTEx expression data and tools: we have tested the association between HsInv0030 tag-SNP rs1808427 (r^2 = 0.86 in CEU) and expression of *CTRB1* and *CTRB2* in pancreas. GTEx consortium has not yet made eQTL data for pancreas publicly available (the sample size, N = 58, is below the chosen threshold), and thus we could not investigate the inversion effect on gene expression in pancreas by means of the

inversion-eQTL study. However, in GTEx portal (see URLs) it is possible to calculate the gene SNP association for any selected tissue using linear regression. Results (**Figure 54**) show that there is a strong correlation between HsInv0030 inverted allele and low expression of *CTRB2*, and also a milder correlation between HsInv0030 inverted allele and low expression of *CTRB1*. According to expression data from Fagerberg et al. (2013), *CTRB1* expresses 1.5 times more than *CTRB2*. Given the increase of *CTRB2* expression coupled with *CTRB1* decrease in expression attributable to the presence of HsInv0030 derived allele (standard conformation, labeled as Homo Ref. in **Figure 54**), we hypothesize that these changes in expression are caused by the exchange of *CTRB1/CTRB2* respective promoters caused by the inverted rearrangement, in an illustrative example of a positional effect.

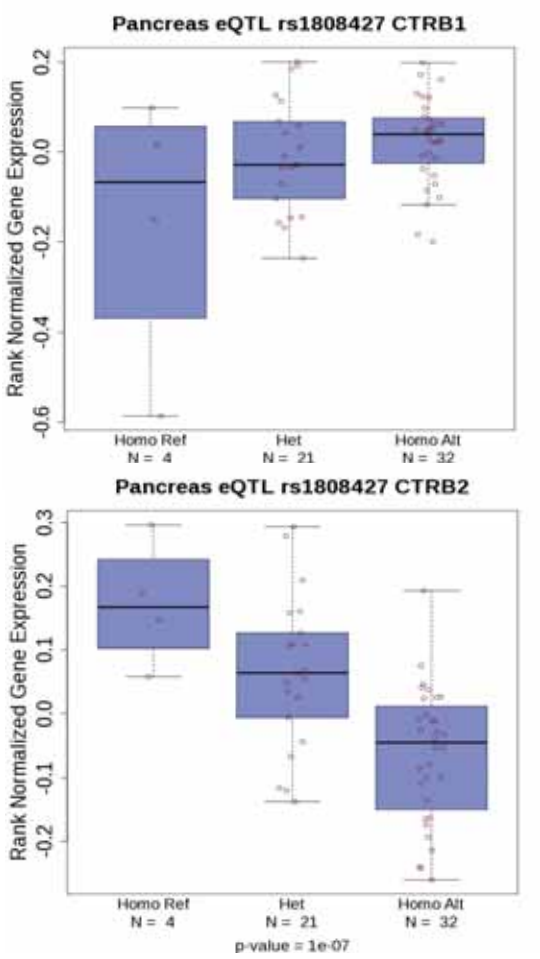


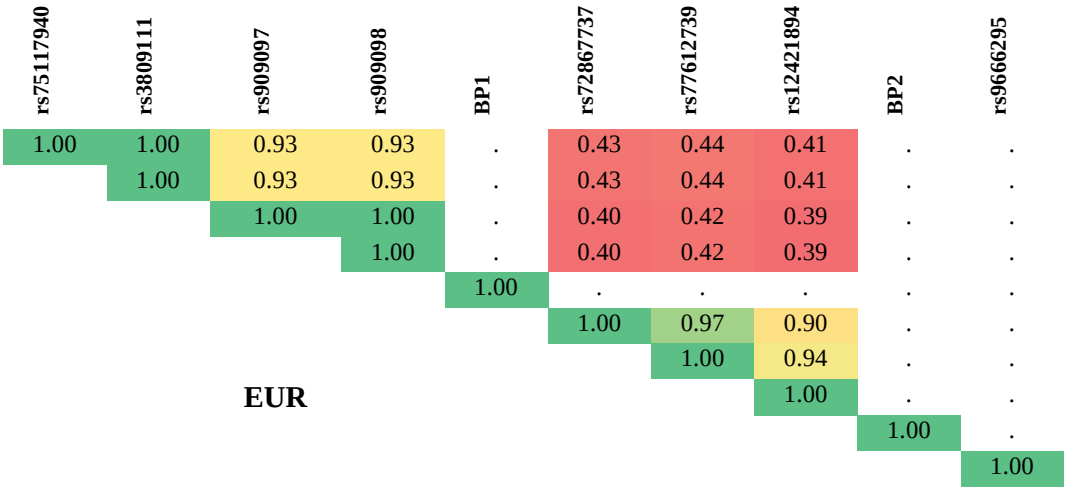
Figure 54 – *CTRB1* and *CTRB2* eQTL data – Top panel: association between *CTRB1* expression and rs1808427. Bottom panel: association between *CTRB2* expression and rs1808427. Homo Ref corresponds to HsInv0030 *Std* homozygotes, Het corresponds to HsInv0030 heterozygotes and Homo Alt corresponds to HsInv0030 homozygotes for *Inv* allele. Sample size (N) is indicated below each genotype category. Data obtained from GTEx portal (see URLs).

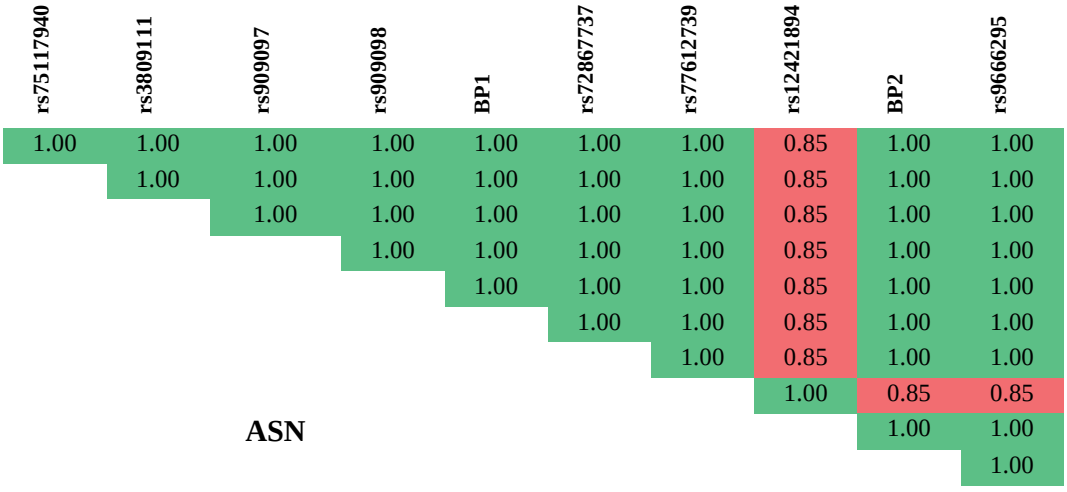
Recently, the *CTRB1/CTRB2* locus has been found to be involved in diabetes susceptibility and treatment via the incretin pathway (’t Hart et al. 2013). We conclude that further studies will be required to confirm this hypothesis, elucidate the functional impact of the inversion and the associated deletion, as well as their role in diabetes.

Among all the 11 inversions that overlap with genes in IRs, HsInv0124 is the only one for which we have strong evidences that the inversion affects the expression of nearby genes. In HsInv0124 inverted genomic region, two paralogous genes (*IFITM2, IFITM3*) are located in the SDs at the BP regions and an additional gene of the same family (*IFITM1*) is contained inside the inversion. The homologous coding sequence of these genes is only 83.48% similar. The inversion would cause a change in orientation of the entire *IFITM1* gene, including its 3’ exon that is also part of an alternative isoform of *IFITM2*. That event could be causing the formation of an *IFITM3-IFITM1* gene fusion and therefore a potential novel protein product. Although we have not found any evidence at a transcript level (ESTs or Genbank RNA database sequences) supporting this hypothesis, results from the LCL DE analysis indicate that 4 genes are associated with HsInv0124 in *cis* (*IFITM2, IFITM3, RP11-326C3.11* and *RP11-326C3.12*), and the inversion-eQTL analyses corroborate these findings. This analysis provides evidence for *IFITM2, IFITM3, RP11-326C3.11* and *RP11-326C3.12* being differentially expressed in LCLs, with *IFITM2, IFITM3* affected at a transcript level, and *IFITM2, RP11-326C3.7* and *RP11-326C3.12* being also differentially expressed in other tissues (see section IV3.2).

HsInv0124 is not associated to a particular haplotype in Europeans and Africans, and we do not observe any variant in high LD with the inversion in these populations (**Figure 55**). An examination of the nucleotide variation pattern in HsInv0124 region in Europeans suggested that the inverted rearrangement displays signals of recurrence (Aguado et al. 2014), which has been confirmed (M. Gayà, personal communication). Indeed, because of the lack of tag SNPs for HsInv0124 in non-Asians, no SNP-based approach can be performed to genotype the inversion in these populations. Therefore, the inversion rearrangement should be genotyped *in situ* in order to look for associations of inversion genotype with particular gene expression profile in different tissues. The aforementioned arguments suggest that HsInv0124 is not affecting genes in *cis* because of strong linkage with a causal haplotype. A more plausible scenario is that the inversion exerts a functional impact by means of positional effect. Nevertheless, among the 5 affected genes in *cis*, only

the alternative transcript ENST00000399815 of gene *IFITM2* is undoubtedly broken by HsInv0124. The paralog RNA genes *RP11-326C3.11* and *RP11-326C3.*7 intronic regions overlap but do not bridge HsInv0124 BPs, therefore we do not have evidence of a breakage effect by the inversion. The remaining candidate RNA gene *RP11-326C3.12* sequence is clearly not affected by the rearrangement as it is located > 15 kb downstream HsInv0124 BP2. A possible hypothesis is that the inversion affects one of these candidate genes, which in turn is involved in the modulation of expression of the remaining ones. Another possible hypothesis is that HsInv0124 is in high LD with an uncharacterized structural variant responsible for the gene expression changes. Interestingly, we observe that three polymorphisms located inside the inversion (rs72867737, rs77612739, rs12421894) mirror the inversion effect on the expression pattern of genes *IFITM2*, *IFITM3* and *RP11-326C3.11* in LCL, as they alter the expression of the particular exonic regions of *IFITM2* and *IFITM3* genes in the same direction and magnitude as the inversion. Two of these variants (rs72867737, rs77612739) are in perfect LD (r^2 = 1) with HsInv0058 in CHB and JPT populations and not linked to the inversion (r^2<0.2) in European, African and other Asian populations. We hypothesize that inside the inverted region there could exist a regulatory element that modulates the expression of *RP11-326C3.11*, *RP11-326C3.*7, *RP11-326C3.12* (upstream) and *IFITM2*, *IFITM3* (downstream) genes in *cis*. The function of this regulatory region could be altered by variants affecting its sequence, like the SNPs rs72867737, rs77612739, rs12421894, or its position with respect to regulated genes, like the inversion HsInv0124.





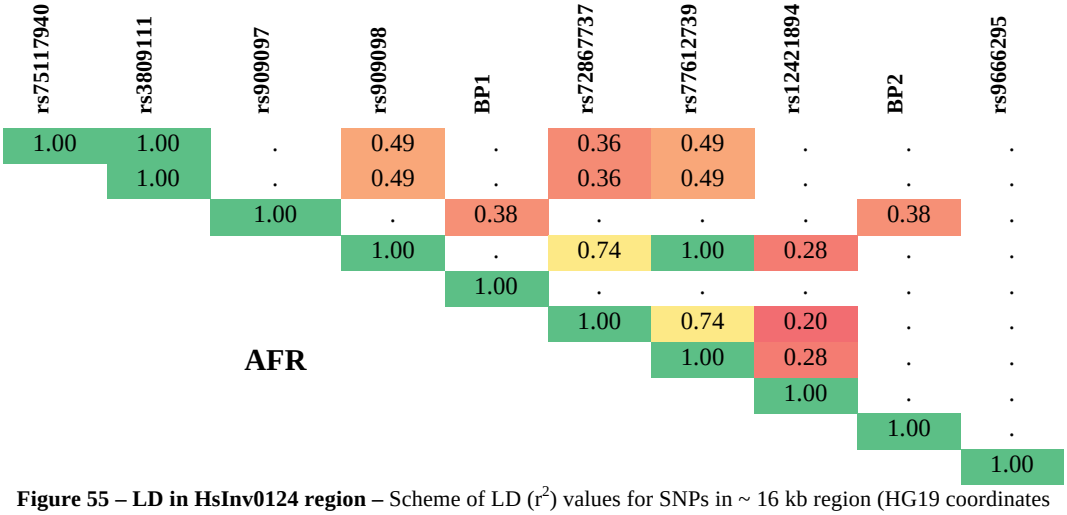


Figure 55 – LD in HsInv0124 region – Scheme of LD (r^2) values for SNPs in ~ 16 kb region (HG19 coordinates chr11:307025-323222) containing HsInv0124. LD values displayed for different populations [ASN (CHB, JPT), EUR (CEU, TSI), AFR (YRI, LWK)]. The inverted fragment is delimited by BPs (BP1, BP2) and contains 3 SNPs, 2 of which (rs72867737, rs77612739) are tag SNPs for HsInv0124 in Asians. $r^2 < 0.2$ is indicated by a dot.

If we analyze non-NAHR mediated inversions with absence of IRs at the breakpoint regions, we find some clear examples of inversions undoubtedly affecting protein-coding genes and causing a potential functional impact (HsInv0102, HsInv0201 and HsInv0379). In all these cases gene expression changes associated with the inversion, both in *cis* and *trans,* are described here or elsewhere (Puig et al., in preparation).

One of the best examples is HsInv0379, which has been validated and studied in detail within our group (Puig et al., in preparation). In this case, the inversion disrupts the zinc finger gene *ZNF257* by affecting the 1st methionine of the protein that is part of a Kruppel-associated box (KRAB) domain (Puig et al., in preparation). The KRAB domain is present in about a third of zinc finger proteins containing C2H2 fingers and is involved in protein-protein interactions (Kim et al. 1996). This domain is generally encoded by two exons, encoding two different subdomains named KRAB-A and KRAB-B, KRAB-containing proteins are thought to have critical functions in cell proliferation and differentiation, apoptosis and neoplastic transformation (Urrutia 2003). HsInv0379 is expected to impair *ZNF257* function by suppressing gene expression, as it inverts the gene promoter region (Puig et al., in preparation). This may have possible consequence at systemic level, as the gene is lowly (avg. expression = 0.87 FPKMs) but widely expressed in many tissues (**Figure 60**, **Appendix**). Interestingly, HsInv0379 inversion is absent in non-Asian populations, which suggests that its relatively low frequency may be related to the inversion causing a functional detrimental effect. Unfortunately, the inversion was filtered out in the LCL DE analysis, as in the genotyped individuals of the expression datasets (Stranger 2007, Stranger 2012, Geuvadis), the inversion was only present in heterozygosity in East Asians (CHB+JPT), but with insufficient frequency (MAF<5%). There exist three tag-SNPs (rs142395395, rs142395395, rs150182828) in perfect LD with the inversions in both CHB and JPT populations, and an additional one in only JPT. However, results of the inversion-eQTL analysis report no gene expression changes associated to these variants, but this was expected as most of eQTL studies derive from European-ancestry samples, where the inversion (and therefore the surrogate tag-SNP) is monomorphic.

HsInv0102 is another clear example of an inversion directly affecting gene expression by a positional effect, causing a mutation that alters a gene transcribed sequence. Specifically it inverts an internal untranslated exon of an alternative isoform of the *RHOH* gene (Villatoro et al., in preparation). Although the inversion does not directly modify the gene protein coding sequence, it could still affect the

gene at a protein level by altering its expression levels, changing its cellular location or by other mechanisms. Results from LCL DE analyses at exon level confirm that HsInv0102 inverted allele negatively correlates with the expression of *RHOH* alternatively spliced exon in both CEU+TSI and YRI populations, and the association has been experimentally validated (Villatoro et al., in preparation) (see section IV3.3). *RHOH* is also expressed in several non-LCL tissues (e.g. lung, skin) for which eQTL data is available. However, inversion-eQTL analysis has not been performed for HsInv0102 due to the lack of tag-SNPs (HsInv0102 is not in high LD with any polymorphism in CEU, TSI, JPT, CHB, YRI and LWK populations). Therefore, to explore HsInv0102 impact on *RHOH* or other genes expression in non-LCL tissues, direct genotyping of the inversion is required, as in the HsInv0124 case. *RHOH* belongs to the Rho family of small GTP-binding proteins, which control a variety of signaling pathways regulating cytoskeleton organization, proliferation, adhesion and migration in eukaryotic cells. Recently, *RHOH* has been identified as a hypermutable gene locus in human lymphomas and has been implicated in suppression of Rac-mediated signaling in cell lines (Troeger et al. 2012). As there is evidence that HsInv0102 mutates *RHOH*, the hypothetical association of this inversion with lymphomas should be explored. Additionally, results of the LCL DE analysis indicate that 36 genes could be affected in *trans* by HsInv0102 inversion. Association tests based on protein-protein interactions of different nature between *RHOH* and this set of genes have been performed with `STRING` (**Figure 56**). The interactions include direct (physical) and indirect (functional) associations and they are inferred from high-throughput experiments, shared genomic context, coexpression and literature. We find no evidence for associations or interactions between *RHOH* and these HsInv0102 *trans* candidates, but there is experimental evidence for protein-protein interactions between 3 HsInv0102 *trans*-affected candidates (*MMP7, SLC3A2* interact with *CD44)* in different studies (Ishimoto et al. 2011; Lau et al. 2012; Yu et al. 2002). However, the significance of these results is not clear and further work should be carried out to explain the real effect of the inversion on these genes.

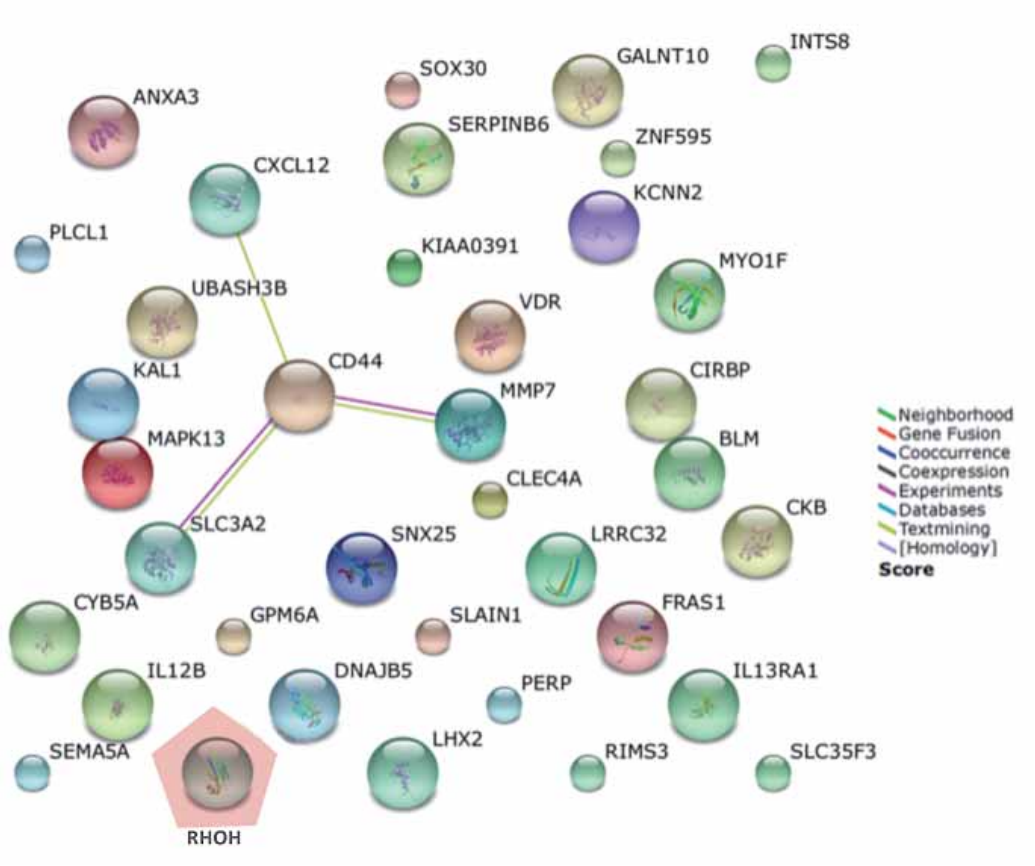


Figure 56 – HsInv0058 DE candidate-genes protein-protein interactions - Protein-protein interactions of different nature between HsInv0058 DE candidate genes are shown. Analysis performed with `STRING` applying intermediate-confidency threshold (as defined by default). *CD44*, *MMP7* and*SLC3A2* show a protein-protein interaction detected by different assays (anti-tag coimmunoprecipitation assay, affinity Capture-Western assay and anti bait coimmunoprecipitation) assay) in 3 different studies. *RHOH* gene highlighted in red.

HsInv0201 is associated with a deletion of a protein-coding exon of gene *SPINK14*. According to the inversion-eQTL results, HsInv0201 inverted allele correlates with *SPINK14* low expression in colon, lung, skin and thyroid, although *SPINK14* is poorly expressed in all mentioned tissues and gene expression seems to be null in the inverted homozygotes (see section IV3.3). Specifically, HsInv0201 associated deletion at BP1 overlaps with the 3rd protein-coding exon of *SPINK14* isoform ENST00000356972. This isoform codes for the 97 aa length *SPINK14* protein ENSP00000349459. *SPINK14* function is unknown, but it is predicted to be a serine peptidase inhibitor because of the presence of a serine protease domain (InterPro id: IPR002350) between aa 34 and 97 of isoform ENSP00000349459. The deletion of the exon associated to HsInv0201 inverted allele causes 46 aminoacids of

the protein sequence to be missing (from aminoacid 38 to aminoacid 83) and alters its ORF that changes the aminoacid sequence spanning half of its original length (47 aa) and causing a premature opal stop codon (**Figure 57**) that produces the loss of 4 additional aa at the N-terminal part. As the mutation alters the predicted functional domain of *SPINK14* protein, HsInv0201 may associate to *SPINK14* loss of function with possible disease implications.

```
1   ATGGCCAAATCTTTCCCAGTATTCTCACTTTTGTCCTTTATCTTGATACATTTGGTGTTA
1   -M--A--K--S--F--P--V--F--S--L--L--S--F--I--L--I--H--L--V--L-
1   -M--A--K--S--F--P--V--F--S--L--L--S--F--I--L--I--H--L--V--L-

61  TCTTCTGTTTCAGGCCCTAGACACTGGTGGCCACCACGTGGAATTATTAAGGTGAAATGT
21  -S--S--V--S--G--P--R--H--W--W--P--P--R--G--I--I--K--V--K--C-
21  -S--S--V--S--G--P--R--H--W--W--P--P--R--G--I--I--K-.........

121 CCATATGAGAAAGTAAACTTGAGCTGGTACAATGGAACGGTCAACCCCTGCCCTGGCTTA
41  -P--Y--E--K--V--N--L--S--W--Y--N--G--T--V--N--P--C--P--G--L-
    ............................................................

181 TATCAACCCATCTGCGGCACCAATTTTATAACCTATGATAATCCCTGCATTCTGTGTGTT
61  -Y--Q--P--I--C--G--T--N--F--I--T--Y--D--N--P--C--I--L--C--V-
    ............................................................

241 GAGAGCTTGAAATCTCATGGAAGAATCAGGTTTTACCATGATGGAAAATGTTAG
81  -E--S--L--K--S--H--G--R--I--R--F--Y--H--D--G--K--C--*-
38  .........E--I--S--W--K--N--Q--V--L--P--*-
```

Figure 57 - ***SPINK14* coding sequence –** Scheme of *SPINK14* isoform ENST00000356972 cDNA (top line of the blocks in black, blue and red) with the corresponding translated protein sequence (in black). The deletion of an exon (in red) by HsInv0201 rearrangement causes an altered, shorter protein sequence (in brown) and affects a functional domain (underlined).

HsInv0201 seems to associate with the expression of other *SPINK* genes in *cis*. Similarl to HsInv0030, we have looked for significant association between HsInv0201 surrogate SNP (rs13360182), which is in perfect LD with the inversion in Europeans, and gene expression changes in pre-computed GTEx eQTL associations (tissue datasets with N>60) through GTEx portal tool. Interestingly, we have found that the HsInv0201 inversion associates to expression changes of only 1 gene: *SPINK6*, located 33,061 bp downstream *SPINK14* TSS and 28448 bp downstream inversion midpoint (**Figure 58**). This association is only found in esophagus mucosa tissue (GTEx eQTL data). Further investigation should be carried out to ascertain the causes of *SPINK* genes expression changes associated to HsInv0201 rearrangement,

perhaps caused by a compensatory effect of overexpressing some *SPINK* due to inactivation of *SPINK14*.

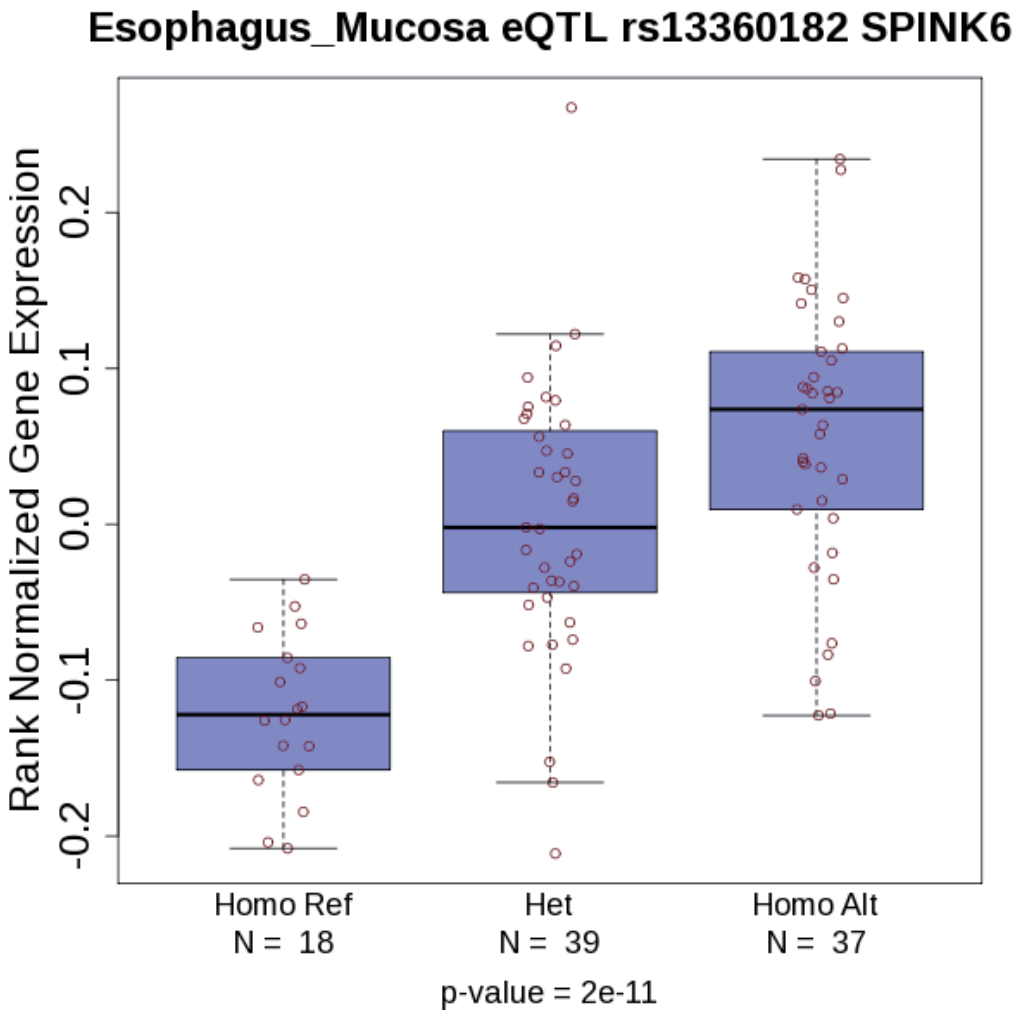


Figure 58 – *SPINK6* eQTL association – Association between *SPINK6* expression and rs13360182. Homo Ref corresponds to HsInv0201 *Std* homozygotes, Het corresponds to HsInv0201 heterozygotes and Homo Alt corresponds to HsInv0201 homozygotes for *Inv* allele. Sample size (N) is indicated below each genotype category. Data obtained from GTEx portal (see URLs). rs13360182 is in perfect LD with HsInv0201 in Eurasian populations.

We also observe evidences of inversions affecting exonic sequences of long non-coding RNAs (lincRNAs). This is the case of HsInv0340, which disrupts the RefSeq validated lincRNA *LINC00395* by inverting the first three 5' exons of the isoform ENST00000451570. Although there exist other isoform models that would

not be affected by the inversion as they lack the three 5' exons overlapping the inversion, it remains unclear wether these isoforms exist independently of the inversion or they have been annotated from samples homozygous or heterozygous for HsInv0340 inverted allele, and therefore are disrupted products of the longest isoform. However, the existance of these isoforms seems improbable as the inversion seems to invert the gene promoter region, as in HsInv0379-*ZNF257*. The first 5' exon of *LINC00395* overlaps in opposite orientation with the first 5' exon of the olfactory receptor pseudogene *OR7E156P*. Both genes show a similar expression profile, and there is evidence that they are expressed mainly in testis (**Figure 60**, **Appendix**), but *LINC00395* function, if any, remains unclear. Interestingly, HsInv0340 derived allele is the most frequent one, corresponds to the standard conformation (global freq. > 83%) and in Eurasian populations is almost completely prevalent (freq. > 98%). We have not found any association between HsInv0340 and *LINC00395* expression, but HsInv0340 has not been interrogated in the tissues were the gene is expressed, because the inversion lacks a tag-SNP.

We observe 7 cases of inversions affecting intronic regions; 5 of which overlap with protein-coding genes (HsInv0006, HsInv0059, HsInv0098, HsInv0105, HsInv409 overlap *DSTYK*, *GABRR1*, *ULK4*, *C7orf10* (also known as *SUGCT*), and *NLGN4X*, respectively). 2 other inversions overlap with validated and putative lincRNAs (HsInv0061, HsInv0374 overlap *RP1-60O19.1*, *AC005562.1*, respectively). In none of the cases the inversion affects splicing sites of the overlapping genes and the functional impact that the inversions could exert on gene structure and expression is unclear. However, in one instance (HsInv0059) we find a correlation with changes in gene expression and in another case (HsInv409) we detect patterns of possible association with disease (section IV4.2).

HsInv0059 appears to affect *GABRR1* expression as it is negatively correlated with the inversion derived allele in several tissues, but mainly in different types of artery tissue (aorta, gland, coronary) and in esophagus mucosa tissue (data not shown). HsInv0059 occurs together with a deletion of 617 bp at its BP1, and thus is unknown whether the inversion or the deletion is the causal variant for the changes in gene expression. *GABRR1* is a member of the rho subunit family of GABA receptors, which are ligand-gated chloride channels of GABA, the major inhibitory neurotransmitter in the mammalian brain. Several transcript variants encoding different isoforms have been found for *GABRR1*, and the gene has been recently associated with neurological diseases such as tremor and epilepsy (Luo et al. 2012). Therefore, the pathological implications of the HsInv0059 rearrangement caused by

the alteration of *GABRR1* expression should be explored. However, we have not found any statistically significant association of HsInv0059 with disease in the performed GWAS studies survey.

Finally, we found a case in which a particular inversion (HsInv0058) seems to be strongly associated with expression changes of several genes, but further analysis suggest that gene expression changes associated with the inversion are not due to a positional effect but more probably to high LD between the inverted allele and a causal haplotype. This hypothesis is supported by the apparent absence of gene or regulatory sequences affected by HsInv0058 and by the strong LD in MHC, the genomic region where the inversion occurred (see section IV3.1). Regarding the association that HsInv0058 has with disease, we find that HsInv0058 tag-SNPs associate with several hypersensitivity and autoimmune diseases such as type I diabetes mellitus, hypothyroidism, asthma, Stevens-Johnson syndrome and toxic epidermal necrolysis (TEN), drug-induced liver injury due to flucloxacillin, ulcerative colitis, Crohn's disease, asthma and also to non-autoinmmune diseases such as narcolepsy. It also associates with complex phenotypes such as height and glycated hemoglobin levels, the latter being an altered trait in diabetes mellitus. However, the association does not seem to becausal. As described in section IV3.1, HsInv0058 is linked to certain MHC haplotypes and to the expression of various MHC genes in several tissues. Hence, it is not surprising to find HsInv0058 associated with autoimmune diseases, as most of these diseases are proven to be mediated by MHC genes. Examples include Psoriasis (*PSORS1*, *HLA-C*0602*), which is the disease that presents the strongest association with HsInv0058 (tag-SNP rs2844645, p-value = 7.29E-19), narcolepsy (*HLA-DQB1*0602*, *HLA-DRB1*1501*) and type I diabetes mellitus (*HLA-DRB1*, *HLA-DQB1*, *HLA-DAB1*) among others. However, after examining the significance of HsInv0058 linked variants in the respective original GWAS studies, none of the candidate associations appeared to be significant and other SNPs seem to be the causal variants of the disorders, so most probably HsInv0058 does not play a role in the diseases mentioned.

To summarize, In order to unveil the functional impact of human polymorphic inversions, further work needs to be carried out by validating results of this work at a gene and protein expression level. In addition to this, it is necessary to perform follow-up and complementary studies involving the analysis of the expression of different categories of non-coding genes. Finally, it is also pertinent to look for inversion associations with other molecular traits besides gene expression and to identify inversion-generated gene chimeras.

Inversions and disease

To explore the possible associations between inversions and diseases, we have made use of the large amount of GWAS data published to date, using a valuable web resource that catalogues, centralizes and makes GWAS data accessible (Beck et al. 2014). We have carried out this analysis in a very similar fashion to the inversion-eQTL analysis: first, identifying inversion surrogates (SNPs in high LD with the inversion) and second, assuming that any significant association attributed to the SNPs is mirrored by the inversion. However, instead of gene expression, as in inversion-eQTL analysis, in this case the trait of interest is association with disease.

After discarding false positives (see section IV4.2), the only inversion that presents a significant association with disease is HsInv0409. This inversion has tag-SNPs in multiple populations; one of them (rs1882409) presents significant association with amyotrophic lateral sclerosis (ALS) in two different GWAS studies (see section IV4.2). However, rs1882409-ALS association may not be indicative of HsInv0409-ALS association, as the SNP is in perfect LD with Eastern Asians (CHB+JPT) but in moderate (0.58) LD with Europeans (CEU), and the ALS study was carried out on individuals of the latter population.

HsInv0409 inverts an intron of *NLGN4X*, which encodes a protein that belongs to a family of neuronal cell surface proteins. Members of this family may act as splice site-specific ligands for beta-neurexins and may be involved in the formation and remodelling of central nervous system synapses. *NLGN4X* is highly expressed, among others, in several central and peripheral nervous system tissues (brain, nerve) (**Figure 59**) so we looked for evidences that ALS associated, HsInv0409 tag-SNP variant (rs1882409) could be an eQTL for *NLGN4X* in these tissues. Although *NLGN4X* eQTLs exist in blood cells, we have found no evidence of genetic associations with *NLGN4X* expression in nervous system tissues. However, because HsInv0409 locates in chromosome X and eQTLs have been generally overlooked in sex chromosomes, we cannot discard the possibility of variant rs1882409 regulating *NLGN4X* expression in tissues that are potentially affected by ALS and therefore suggesting a possible association with HsInv0409.

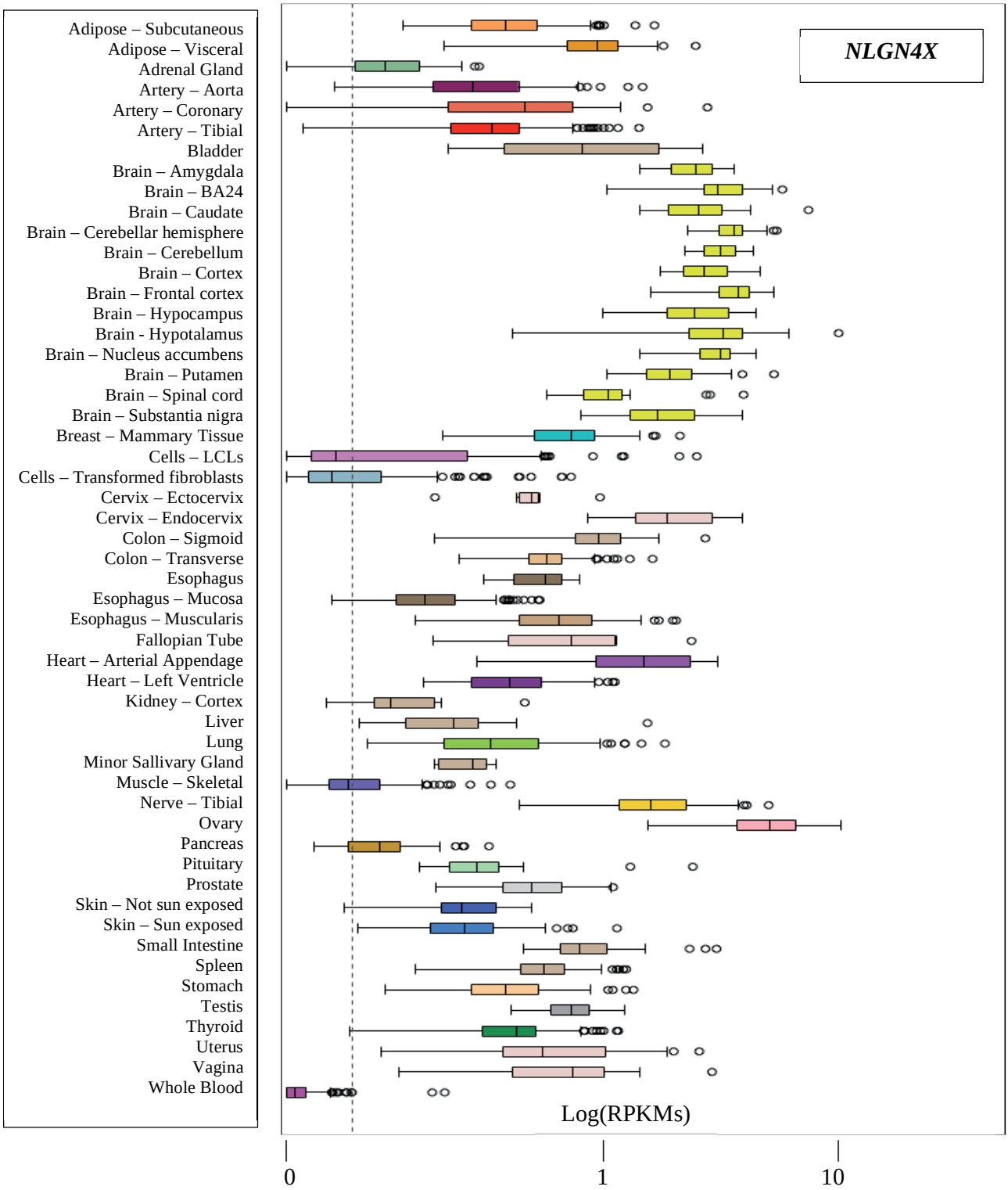


Figure 59 - *NLGN4X* expression profile in multiple tissues - Image adapted from GTEx database. Expression values are transformed (see GTEx project documentation).

Overall, results from our GWAS survey suggest that some inversions (e.g. HsInv0409) may play a putative role in disease by affecting the expression of a gene in *cis*. However, in our approach inversion-disease associations have been inferred by inversion tag-SNPs that may not reflect the correct inversion genotype in some cases, due to SNP miscalling events or due to specific population genetic background and LD patterns in disease case samples that differ from the reference ones used to assess inversion tag-SNPs (e.g. HapMap, 1000GP populations). Another problem of this approach is the coverage of SNP arrays used for GWAS studies. Even if an inversion is associated to a disease investigated in a GWAS study and have a perfect tag-SNP in the GWAS population, if the SNP is not present in the microarray the inversion will be overlooked. Hence, genome-wide association studies directly genotyping the inversions in cases and control cohorts are a mandatory step to validate putative inversion and disease associations. In addition, changes in expression of genes associated to disease-related inversions require experimental validation to prove the legitimacy of the associations found.

VI FUTURE DIRECTIONS

Here we have identified a few polymorphic inversions in the human genome that may affect gene expression in a tissue-specific manner. However, all the analysis performed is based on finding statistically significant associations of inversion genotypes, inferred directly or indirectly, with gene expression changes. Further, the expression data has not been generated but obtained from dozens of publicly available expression datasets. Only in one case (HsInv0102) we replicated the association experimentally. Therefore, the next step (on-going) should be to verify the remaining candidate inversion-gene associations; reproducing the observed expression change patterns by genotyping inversions *in situ* in an alternative set of samples, experimentally measuring (e.g. by RT-PCR) the expression of candidate genes and re-testing the association of inverted alleles with gene expression. However, prior to this it would be necessary to identify the subset of inversions associated with gene expression changes that have a causal role, as in some cases (e.g. HsInv0058) we have observed that a particular haplotype inherited together with the inverted allele seems to be causing the alteration of gene expression. However, in these cases, the interaction between the haplotype and the inversion in causing the change of gene expression should be investigated, as the inversion may still have an independent additive effect. Alternatively, the inversion and the haplotype could have different effects in combination than individually (epistatic relationship). Nevertheless, that can only be easily addressed in cases where there is strong but incomplete inversion-haplotype linkage disequilibrium.

In this work, we have investigated the changes in expression on several gene types, both coding and non-coding, and have included also pseudogenes. However we have missed other kind of gene types such us the generally called small RNAs (sRNAs), that include micro-RNAs (miRNAs), endogenous small interfering RNAs (endo-siRNAs) and piwi-interacting RNAs (piRNAs). Recently, a work aiming to determine sRNA-QTLs has been published (Lappalainen et al. 2013) reporting 65 genes that have a significant QTL out of 644 analyzed genes (FDR = 10%). The small RNA-Seq expression dataset used in this work, composed of 452 lymphoblastoid cell lines is publicly available (arrayexpress id E-GEUV-2). As the samples are part of the 1000GP, their SNP genotypes are known and so it is straightforward to perform an inversion-sRNA-QTL study here, which would be a follow-up of the work reported here. Additionally, we could also test the differential expression associated withthe inversion by direct genotyping, but this is not a very straightforward approach.

Another functional effect that inversions can cause, not directly investigated in this work, is the formation of chimeric genes by exchanging gene exons at the inversion boundaries or by exonizing originally non-coding regions. An illustrative example for that is HsInv0030, which causes the formation of CTRB1/CTRB2 quimeras (Pang et al. 2013). Therefore, a follow-up study tackling this issue should be performed, as it is predicted that other inversions of our analyzed set could also produce quimeric genes. An example of that is HsInv0124 alteration of *IFITM2/IFITM3* locus, supported by differential expression of *IFITM2* and *IFITM2* particular exons on both LCL DE and inversion-eQTL analysis.

Although in this work we have focused on trying to assess inversion effects on gene expression, there are many other genomic features that can be potentially analyzed. The means to identify and quantify these features genome-wide in a large number of samples, which allows the identification of possible genetic determinants exist today. For instance, it has been recently discovered that DNAse I hypersensitivity trait loci (dsQTL) are a major determinant of human expression variation (Degner et al. 2012) and that methylation QTLs (mQTLs) are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels (Banovich et al. 2014). Therefore, follow-up studies aiming to identify inverted rearrangements affecting these traits could be performed. In addition, the magnitude of change of expression is not the only gene expression feature that can be altered by genetic variants. Recently, several studies aiming to determine loci that affect gene expression variability QTLs (*evQTLs*) have been carried out (Brown et al. 2014; G. Wang et al. 2014). These studies have focused on studying SNPs, and therefore it is unclear the potential role of inversions in affecting this particular trait. Therefore inversion-oriented evQTL studies may be carried out.

Finally, the ultimate objective of this study is to assess the functional impact of inversions. Therefore, it should be investigated how changes in gene expression affect the expression of the gene at the protein level and how this translates into an alteration of the protein function. For that, functional analysis should be conducted, for instance by generating the inversion on wild-type (homozygotes for standard allele) cells and measuring changes in their phenotype, at a morphological and molecular level.

VII CONCLUSIONS

1. GRIAL, to date the only PEM based algorithm specifically designed to predict inversions, performs with more accuracy and efficiency compared to other PEM based SV-detection methods, particularly in refining inversion BPs.

2. Predicting inversions from PEM data yields a high false positive rate. The usage of several filters both at pre and post-mapping stages of the process, coupled with manual inspection of complex cases is crucial to discard false positive predictions.

3. In general, human polymorphic inversions tend to be located in intergenic regions. However, in the dayaset analysed dataset here, a considerable amount of inversions overlap with genes, and 13.6% of the cases affect gene exons, several of which have been confirmed and are being studied in detail in the laboratory.

4. The functional impact of inversions with BPs overlapping with paralogous genes in IRs is heterogeneous, and depends on the level of identity of the overlapped genes CDS. Here, we have identified a case (HsInv0124) of an inversion that affects the expression of paralogous genes (*IFITM2/IFITM3*) at the transcript level.

5. Non-recurrent inversions often have tag-SNPs that can be used as surrogates to search for inversion associations with quantitative traits at a molecular level, such as gene expression and methylation, among others. Here, making use of this, we have identified 13 inversions affecting the expression of 36 genes in 10 different tissues and cell lines.

6. The two approaches employed to identify inversion associations with gene expression (LCL DE analysis and inversion-eQTL) are complementary and have allowed us to identify in total 19 inversions that seem to affect the expression of 43 genes in several tissues. A subset (N = 11) of the associations found in lymphocyte-derived cells are consistent as they have been identified by both approaches.

7. Inversions may be linked to haplotypes as a result of recombination inhibition. We have identified a case of an inversion (HsInv0058) in the MHC that is associated with gene expression as a result of linkage disequilibrium with a putative causal haplotype.

8. We have validated the methodology used for differential expression analysis by reproducing known associations between two well-studied inversions (17q21.31, 8p23.1) and 6 genes expressed in lymphocyte-derived cell lines. We have also identified novel associations of 17q21.31 inversion with the expression of 6 genes: *KANSL1, KANSL-AS1, ARL17B, WNT3, NSF,* and *ARHGAP27.* Besides, we observe that at least 2 of these genes (*ARL17B, KANSL-AS*1) are associated to 17q21.31 structural rearrangements.

9. The association of inversions with disease can be studied by GWAS, typically for inversions with tag-SNPs. Results from our GWAS survey identify one inversion (HsInv0409) associated with amyotrophic lateral sclerosis in two different GWAS studies.

VIII APPENDIX

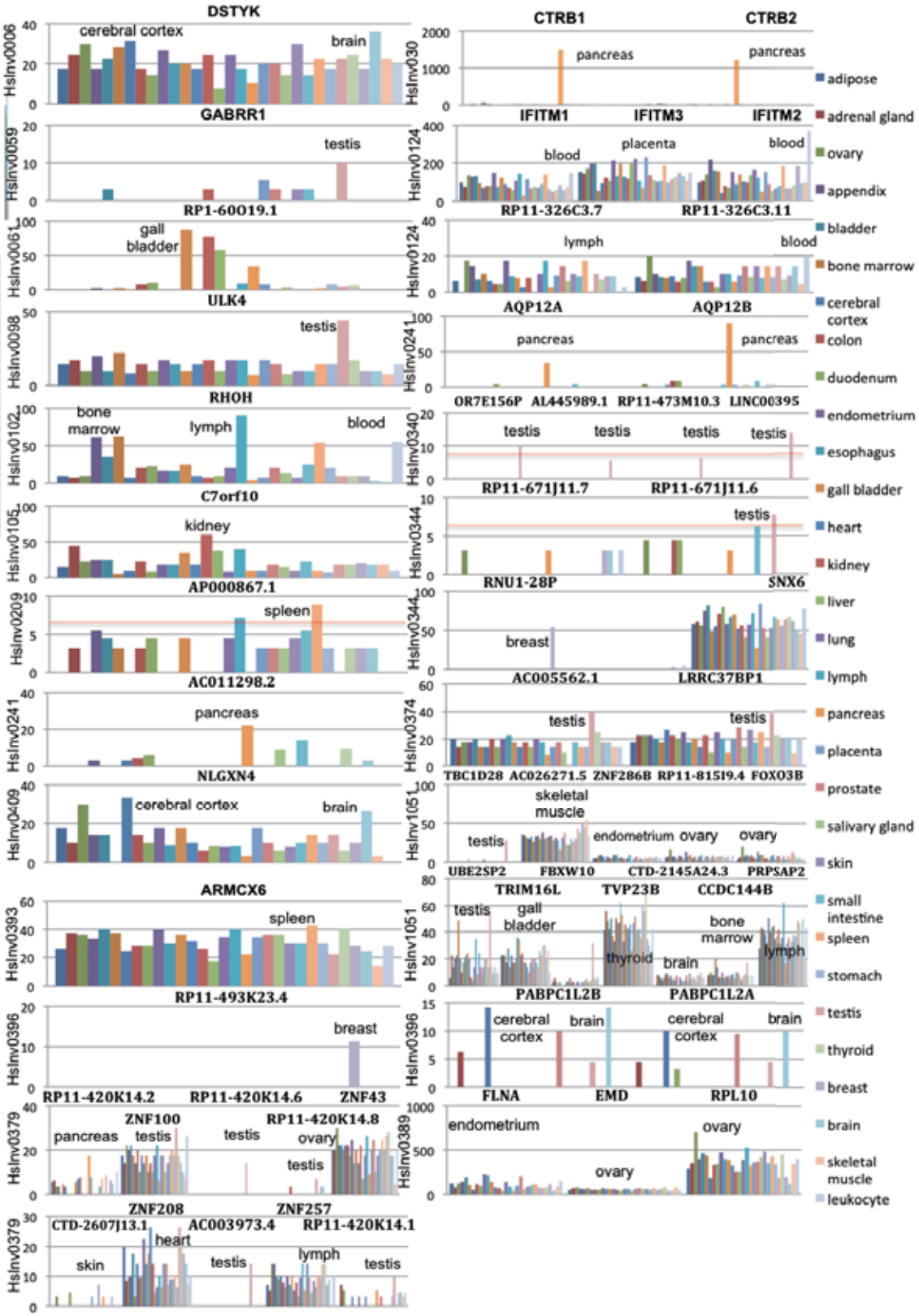


Figure 60 – Expression of genes contained or overlapping inversions – Expression values in $(100xFPKMs)^{1/2}$**.** Red line indicates low expression threshold (FPKM = 0.5). Gene top-expressed tissue/s labeled. Expression data of 27 tissues of 95 samples obtained from (Fagerberg et al. 2013) and additional breast, brain, skeletal muscle and leukocyte expression data obtained from Illumina bodyMap 2.0 project.

Inversion id	Gene	Gene type	Transcript	Transcript type	Effect	%id paralogs (transcript,CDS , protein)
HsInv0006	*DSTYK*	protein_coding ***	ENST00000367160	known protein coding	inverted intron	
HsInv0006	*DSTYK*	protein_coding ***	ENST00000367161	known protein coding	inverted intron	
HsInv0006	*DSTYK*	protein_coding ***	ENST00000367162	known protein coding	inverted intron	
HsInv0030	*CTRB1*	protein coding ***	ENST00000361017	known protein coding	overlapping BP (5' exon exchanged)	100,100,100
HsInv0030	*CTRB1*	protein coding ***	ENST00000495583	putative protein coding	overlapping BP (unchanged)	
HsInv0030	*CTRB2*	protein coding ***	ENST00000303037	known protein coding	overlapping BP (5' exon exchanged)	
HsInv0030	*CTRB2*	protein coding ***	ENST00000562387	putative protein coding	overlapping BP (unchanged)	
HsInv0030	*CTRB2*	protein coding ***	ENST00000562106	novel protein coding	overlapping BP (unchanged)	
HsInv0030	*CTRB2*	protein coding ***	ENST00000567767	putative protein coding	overlapping BP (unchanged)	
HsInv0030	*CTRB2*	protein coding ***	ENST00000481611	known retained intron	overlapping BP (unchanged)	
HsInv0030	*CTRB2*	protein coding ***	ENST00000565656	known retained intron	overlapping BP (5' exon inverted)	100,100,100
HsInv0059	*GABRR1*	protein_coding ***	ENST00000454853	known protein coding	inverted intron	
HsInv0059	*GABRR1*	protein_coding ***	ENST00000435811	known protein coding	inverted intron	
HsInv0059	*GABRR1*	protein_coding ***	ENST00000369451	putative protein coding	inverted intron	
HsInv0059	*GABRR1*	protein_coding ***	ENST00000457434	known nonsense mediated decay	inverted intron	
HsInv0059	*GABRR1*	protein_coding ***	ENST00000481493	putative processed transcrip	inverted intron	
HsInv0061	*RP1-60O19.1*	lincRNA **	ENST00000602621	known lincRNA	inverted intron	

Inversion id	Gene	Gene type	Transcript	Transcript type	Effect	%id paralogs (transcript,CDS , protein)
HsInv0061	*RP1-60O19.1*	lincRNA **	ENST00000436659	known lincRNA	inverted intron	
HsInv0061	*RP1-60O19.1*	lincRNA **	ENST00000428750	known lincRNA	inverted intron	
HsInv0098	*ULK4*	protein_coding ***	ENST00000301831	known protein coding	inverted intron	
HsInv0102	*RHOH*	protein coding ***	ENST00000508513	known protein coding	inverted exon [nc]	
HsInv0105	*C7orf10*	protein coding ***	ENST00000401647	novel protein coding	inverted intron	
HsInv0105	*C7orf10*	protein coding ***	ENST00000335693	known protein coding	inverted intron	
HsInv0105	*C7orf10*	protein coding ***	ENST00000464028	Known processed transcript	inverted intron	
HsInv0105	*C7orf10*	protein coding ***	ENST00000460466	Known processed transcript	inverted intron	
HsInv0105	*C7orf10*	protein coding ***	ENST00000309930	known protein coding	inverted intron	
HsInv0124	*IFITM2*	protein coding ***	ENST00000399815	putative protein coding	inverted 3’ exon	
HsInv0124	*RP11-326C3.7*	antisense *	ENST00000526612	known antisense	overlapping BP (undef)	99.5,-,-
HsInv0124	*RP11-326C3.11*	antisense *	ENST00000602429	known antisense	overlapping BP (undef)	99.5,-,-
HsInv0124	*RP11-326C3.11*	antisense *	ENST00000602756	known antisense	overlapping BP (undef)	
HsInv0124	*RP11-326C3.11*	antisense *	ENST00000508004	known antisense	overlapping BP (undef)	
HsInv0201	*SPINK14*	protein coding ***	ENST00000356972	known protein coding	inverted exon [c]	
HsInv0209	*KRTAP5-11*	protein coding ***	ENST00000398530	known protein coding	overlapping BP (monoexon exchanged?)	85,84,61.5
HsInv0209	*KRTAP5-11*	protein coding ***	ENST00000526239	known protein coding	overlapping BP (3' exon exchanged?)	
HsInv0209	*KRTAP5-10*	protein coding ***	ENST00000398531	known protein coding	overlapping BP (monoexon exchanged?)	
HsInv0209	*KRTAP5-10*	protein coding ***	ENST00000376536	known protein coding	overlapping BP (3' exon exchanged?)	85,84,61.5
HsInv0209	*AP000867.14*	pseudogene	ENST00000511464	Known processed pseudogene	overlapping BP (monoexon exchanged?)	
HsInv0209	*AP000867.1*	protein coding *	ENST00000343767	known protein coding	overlapping BP (3' exon exchanged?)	
HsInv0241	*AQP12A*	protein coding ***	ENST00000429564	known protein coding	overlapping BP (5' exon exchanged)	99,99,98
HsInv0241	*AQP12A*	protein coding ***	ENST00000337801	known protein coding	overlapping BP (5' exon	

Inversion id	Gene	Gene type	Transcript	Transcript type	Effect	%id paralogs (transcript,CDS , protein)
					exchanged)	
HsInv0241	*AQP12A*	protein coding ***	ENST00000474778	putative processed transcript	overlapping BP (unchanged)	
HsInv0241	*AQP12A*	protein coding ***	ENST00000471878	putative processed transcript	overlapping BP (unchanged)	
HsInv0241	*AQP12A*	protein coding ***	ENST00000460527	putative processed transcript	overlapping BP (exon exchanged)	
HsInv0241	*AQP12B*	protein coding ***	ENST00000407834	Protein coding	overlapping BP (5' exon exchanged)	99,99,98
HsInv0241	*AQP12B*	protein coding ***	ENST00000414322	known nonsense mediated decay	overlapping BP (5' exon exchanged)	
HsInv0241	*AQP12B*	protein coding ***	ENST00000413999	known nonsense mediated decay	overlapping BP (5' exon exchanged)	
HsInv0241	*AQP12B*	protein coding ***	ENST00000452886	known nonsense mediated decay	overlapping BP (5' exon exchanged)	
HsInv0241	*AQP12B*	protein coding ***	ENST00000459806	putative processed transcript	overlapping BP (unchanged)	
HsInv0278	*TRNA_Val*	tRNA	-	tRNA	overlapping BP (unchanged)	
HsInv0278	*TRNA_Leu*	tRNA	-	tRNA	overlapping BP (unchanged)	
HsInv0340	*LINC00395*	lincRNA*	ENST00000451570	antisense	inverted 5' exons [nc,nc,nc]	
HsInv0344	*RP11-671J11.7*	processed transcript **	ENST00000553697	Known processed transcript	overlapping BP (unchanged)	
HsInv0344	*RP11-671J11.6*	lincRNA **	ENST00000556693	known lincRNA	overlapping BP (unchanged)	
HsInv0344	*RNU1-27P*	snRNA [NA]	ENST00000383869	known snRNA	overlapping BP (unchanged)	
HsInv0344	*RNU1-28P*	snRNA [NA]	ENST00000383861	known snRNA	overlapping BP (unchanged)	
HsInv0344	*RP11-671J11.4*	antisense **	ENST00000554608	known antisense	overlapping BP (unchanged)	
HsInv0344	*SNX6*	protein coding ***	ENST00000396526	known protein coding	overlapping BP (unchanged)	

Inversion id	Gene	Gene type	Transcript	Transcript type	Effect	%id paralogs (transcript,CDS , protein)
HsInv0344	*SNX6*	protein coding ***	ENST00000396534	known protein coding	overlapping BP (unchanged)	
HsInv0374	*AC005562.1*	processed transcript [NA]	ENST00000398849	Known processed transcript	inverted intron	
HsInv0374	*AC005562.1*	processed transcript [NA]	ENST00000431308	Known processed transcript	inverted intron	
HsInv0374	*AC005562.1*	processed transcript [NA]	ENST00000440026	Known processed transcript	inverted intron	
HsInv0374	*LRRC37BP1*	pseudogene	ENST00000398851	Known transcribed unprocessed pseudogene	overlapping BP (unchanged)	
HsInv0379	*ZNF257*	protein coding ***	ENST00000435820	known nonsense mediated decay	inverted 5' exons [c,c]	
HsInv0379	*ZNF257*	protein coding ***	ENST00000597927	putative protein coding	inverted 5' exon [nc]	
HsInv0379	*ZNF257*	protein coding ***	ENST00000594363	putative protein coding	inverted 5' exon [c]	
HsInv0379	*ZNF257*	protein coding ***	ENST00000594947	known protein coding	inverted 5' exon [c]	
HsInv0379	*ZNF257*	protein coding ***	ENST00000597796	known nonsense mediated decay	inverted 5' exons [c,c]	
HsInv0379	*ZNF257*	protein coding ***	ENST00000596471	known nonsense mediated decay	inverted 5' exons [nc,nc]	
HsInv0379	*ZNF257*	protein coding ***	ENST00000600162	putative protein coding	inverted 5' exon [c]	
HsInv0379	*RP11-420K14.1*	pseudogene	ENST00000600907	Known unprocessed pseudogene	inverted 5' exon [nc]	
HsInv0389	*RPL10*	protein_coding ***	ENST00000474786	known retained intron	overlapping BP (5' exon exchanged)	
HsInv0393	*RP4-545K15.3*	pseudogene	ENST00000539247	known protein coding	overlapping BP (unchanged)	
HsInv0393	*ARMCX6*	protein_coding ***	ENST00000361910	known protein coding	overlapping BP (unchanged)	
HsInv0393	*ARMCX6*	protein_coding ***	ENST00000538627	known protein coding	overlapping BP (unchanged)	
HsInv0393	*ARMCX6*	protein_coding ***	ENST00000497931	Known processed transcript	overlapping BP (unchanged)	
HsInv0393	*ARMCX6*	protein_coding ***	ENST00000467089	Known processed transcript	overlapping BP (unchanged)	
HsInv0393	*ARMCX6*	protein_coding ***	ENST00000462302	Known processed transcript	overlapping BP (unchanged)	
HsInv0396	*RP11-493K23.1*	antisense *	ENST00000416989	known antisense	overlapping BP (unchanged)	
HsInv0396	*RP11-493K23.4*	antisense *	ENST00000454388	known antisense	overlapping BP (unchanged)	
HsInv0396	*PABPC1L2B*	protein_coding ***	ENST00000373521	known protein coding	overlapping BP (undef)	98,100,100
HsInv0396	*PABPC1L2B*	protein_coding ***	ENST00000538388	known protein coding	overlapping BP (undef)	
HsInv0396	*PABPC1L2A*	protein_coding ***	ENST00000373519	known protein coding	overlapping BP (undef)	

Inversion id	Gene	Gene type	Transcript	Transcript type	Effect	%id paralogs (transcript,CDS , protein)
HsInv0396	*PABPC1L2A*	protein_coding ***	ENST00000453389	known protein coding	overlapping BP (undef)	98,100,100
HsInv0409	*NLGN4X*	protein_coding ***	ENST00000381095	known protein coding	inverted intron	
HsInv0409	*NLGN4X*	protein_coding ***	ENST00000381093	known protein coding	inverted intron	
HsInv0409	*NLGN4X*	protein_coding ***	ENST00000275857	known protein coding	inverted intron	
HsInv0409	*NLGN4X*	protein_coding ***	ENST00000381092	known protein coding	inverted intron	
HsInv0409	*NLGN4X*	protein_coding ***	ENST00000538097	known protein coding	inverted intron	
HsInv0409	*NLGN4X*	protein_coding ***	ENST00000469740	Known processed transcript	inverted intron	
HsInv1051	*CCDC144B*	pseudogene	ENST00000285176	Transcribed unprocessed pseudogene	inverted 5' exons [nc,nc]	
HsInv1051	*CCDC144B*	pseudogene	ENST00000425214	Processed transcript	overlapping BP (undef)	
HsInv1051	*CCDC144B*	pseudogene	ENST00000442583	Processed transcript	overlapping BP (undef)	
HsInv1051	*CCDC144B*	pseudogene	ENST00000445752	Retained intron	inverted 5' exons [nc,nc]	
HsInv1051	*CCDC144B*	pseudogene	ENST00000447265	Transcribed unprocessed pseudogene	inverted 5' exons [nc,nc]	
HsInv1051	*CCDC144B*	pseudogene	ENST00000450277	Processed transcript	inverted 5' exons [nc,nc]	
HsInv1051	*CCDC144B*	pseudogene	ENST00000455629	Processed transcript	inverted 5' exons [nc,nc]	
HsInv1051	*CCDC144B*	pseudogene	ENST00000457330	Retained intron	inverted 5' exons [nc,nc]	
HsInv1051	*PRPSAP2*	protein coding ***	ENST00000441887	known protein coding	inverted 5' exons [nc,nc] (unchanged)	
HsInv1051	*AC107982.4*	pseudogene	ENST00000507171	Known unprocessed pseudogene	overlapping BP (undef)	

Table 29 – Transcripts overlapping inversion BPs - Inversion identifier (Inversion id) corresponds to invFEST id (Martínez-Fundichely et al. 2013). Gene and transcripts ids and annotation (Gene type, Transcript type) obtained from Ensembl v.75 and GENCODE. The asterisk marks refer to the reliability of the gene. For RNA genes, one asterisk indicates that the only support is from a single EST, the best supporting EST is suspicious to be an artefact or no single transcript supports the model structure. Two asterisks indicate that the best supporting mRNA is flagged as suspect, the support is from multiple ESTs or all splice junctions of the transcript are supported by at least one non-suspect mRNA. NA indicates that the reliability of transcript has not been analysed. For protein genes, one asterisk indicates a predicted putative protein, with no experimental evidences supporting the prediction. Two asterisks indicate that the protein is supported at a transcript level. Three asterisks indicate that the protein is supported at a protein level. In bold, genes that with average expression > 0.5 FPKMs in (Fargerberg et al. 2013). NC: non-coding. C: coding. Undef: inversion BPs not clearly defined, overlap with genic sequence cannot be determined.

IX BIBLIOGRAPHY

Aberg, Karolina et al. 2010. “Genomewide Association Study of Movement-Related Adverse Antipsychotic Effects.” *Biological Psychiatry* 67(3): 279–82.

Aguado, Cristina et al. 2014. “Validation and Genotyping of Multiple Human Polymorphic Inversions Mediated by Inverted Repeats Reveals a High Degree of Recurrence” ed. Gregory S. Barsh. *PLoS Genetics* 10(3): e1004208.

Ahn, S.-M. et al. 2009. “The First Korean Genome Sequence and Analysis: Full Genome Sequencing for a Socio-Ethnic Group.” *Genome Research* 19(9): 1622–29.

Alkan, Can, Bradley P. Coe, and Evan E. Eichler. 2011. “Genome Structural Variation Discovery and Genotyping.” *Nature Reviews Genetics* 12(5): 363–76.

Altshuler, David M. et al. 2010. “Integrating Common and Rare Genetic Variation in Diverse Human Populations.” *Nature* 467(7311): 52–58.

Alves, Joao M, Alexandra M Lopes, Lounès Chikhi, and António Amorim. 2012. “On the Structural Plasticity of the Human Genome: Chromosomal Inversions Revisited.” *Current genomics* 13(8): 623–32.

Anderson, Carl A et al. 2011. “Meta-Analysis Identifies 29 Additional Ulcerative Colitis Risk Loci, Increasing the Number of Confirmed Associations to 47.” *Nature Genetics* 43(3): 246–52.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2014. “HTSeq – A Python Framework to Work with High-Throughput Sequencing Data.” *bioRxiv*. http://biorxiv.org/content/early/2014/02/20/002824 (July 17, 2014).

Antonacci, Francesca et al. 2009. “Characterization of Six Human Disease-Associated Inversion Polymorphisms.” *Human Molecular Genetics* 18(14): 2555–66.

———. 2010. “A Large and Complex Structural Polymorphism at 16p12.1 Underlies Microdeletion Disease Risk.” *Nature Genetics* 42(9): 745–50.

Arlt, Martin F et al. 2011. “Comparison of Constitutional and Replication Stress-Induced Genome Structural Variation by SNP Array and Mate-Pair Sequencing.” *Genetics* 187(3): 675–83.

Artigas, María Soler et al. 2011. “Genome-Wide Association and Large-Scale Follow up Identifies 16 New Loci Influencing Lung Function.” *Nature Genetics* 43(11): 1082–90.

Ayala, Diego, Rafael F. Guerrero, and Mark Kirkpatrick. 2013. “Reproductive Isolation and Local Adaptation Quantified for a Chromosome Inversion in a Malaria Mosquito.” *Evolution* 67(4): 946–58.

Bachtrog, Doris. 2006. “A Dynamic View of Sex Chromosome Evolution.” *Current Opinion in Genetics & Development* 16(6): 578–85.

Barnett, Derek W. et al. 2011. “BamTools: A C++ API and Toolkit for Analyzing and Managing BAM Files.” *Bioinformatics* 27(12): 1691–92.

Beck, Tim et al. 2014. “GWAS Central: A Comprehensive Resource for the Comparison and Interrogation of Genome-Wide Association Studies.” *European Journal of Human Genetics* 22(7): 949–52.

Bell, Jordana T. et al. 2011. “DNA Methylation Patterns Associate with Genetic and Gene Expression Variation in HapMap Cell Lines.” *Genome Biology* 12(1): R10.

Boettger, Linda M, Robert E Handsaker, Michael C Zody, and Steven A McCarroll. 2012. “Structural Haplotypes and Recent Evolution of the Human 17q21.31 Region.” *Nature genetics* 44(8): 881–85.

Bondeson, M. L. et al. 1995. “Inversion of the IDS Gene Resulting from Recombination with IDS-Related Sequences Is a Common Cause of the Hunter Syndrome.” *Human Molecular Genetics* 4(4): 615–21.

Brem, Rachel B., John D. Storey, Jacqueline Whittle, and Leonid Kruglyak. 2005. “Genetic Interactions between Polymorphisms That Affect Gene Expression in Yeast.” *Nature* 436(7051): 701–3.

Brem, R. B. 2002. “Genetic Dissection of Transcriptional Regulation in Budding Yeast.” *Science* 296(5568): 752–55.

Brown, A. A. et al. 2014. “Genetic Interactions Affecting Human Gene Expression Identified by Variance Association Mapping.” *eLife* 3(0): e01381–e01381.

De la Chapelle, A. et al. 1974. “Pericentric Inversions of Human Chromosomes 9 and 10.” *American Journal of Human Genetics* 26(6): 746–66.

Charlesworth, B. 1991. “The Evolution of Sex Chromosomes.” *Science* 251(4997): 1030–33.

Chen, Ken et al. 2009. “BreakDancer: An Algorithm for High-Resolution Mapping of Genomic Structural Variation.” *Nature Methods* 6(9): 677–81.

Chowdhury, Reshmi et al. 2009. “Genetic Analysis of Variation in Human Meiotic Recombination” ed. Gregory P. Copenhaver. *PLoS Genetics* 5(9): e1000648.

Conrad, Donald F. et al. 2010. “Origins and Functional Impact of Copy Number Variation in the Human Genome.” *Nature* 464(7289): 704–12.

Consortium, 1000 Genomes Project, and others. 2010. “A Map of Human Genome Variation from Population-Scale Sequencing.” *Nature* 467(7319): 1061–73.

Coyne, Jerry A. 2004. *Speciation*. Sunderland, Mass: Sinauer Associates.

Daly, Ann K et al. 2009. “HLA-B*5701 Genotype Is a Major Determinant of Drug-Induced Liver Injury due to Flucloxacillin.” *Nature Genetics* 41(7): 816–19.

Darai-Ramqvist, E. et al. 2008. “Segmental Duplications and Evolutionary Plasticity at Tumor Chromosome Break-Prone Regions.” *Genome Research* 18(3): 370–79.

Degner, Jacob F et al. 2012. “DNase I Sensitivity QTLs Are a Major Determinant of Human Expression Variation.” *Nature* 482(7385): 390–94.

Dimas, A. S. et al. 2009. “Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner.” *Science* 325(5945): 1246–50.

———. 2012. “Sex-Biased Genetic Effects on Gene Regulation in Humans.” *Genome Research* 22(12): 2368–75.

Dobzhansky, Theodosius. 1970. *Genetics of the Evolutionary Process*. New York: Columbia University Press.

Van Doorn, G. S., and M. Kirkpatrick. 2007. “Turnover of Sex Chromosomes Induced by Sexual Conflict.” *Nature* 449(7164): 909–12.

Dunham, Ian et al. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489(7414): 57–74.

Eichler, Evan E. et al. 2007. “Completing the Map of Human Genetic Variation.” *Nature* 447(7141): 161–65.

ENCODE Project Consortium. 2004. “The ENCODE (ENCyclopedia Of DNA Elements) Project.” *Science (New York, N.Y.)* 306(5696): 636–40.

———. 2007. “Identification and Analysis of Functional Elements in 1% of the Human Genome by the ENCODE Pilot Project.” *Nature* 447(7146): 799–816.

Eriksson, Nicholas et al. 2012. “Novel Associations for Hypothyroidism Include Known Autoimmune Risk Loci.” *PLoS ONE* 7(4): e34442.

Van Es, Michael A. et al. 2007. “ITPR2 as a Susceptibility Gene in Sporadic Amyotrophic Lateral Sclerosis: A Genome-Wide Association Study.” *Lancet Neurology* 6(10): 869–77.

Everitt, Aaron R et al. 2012. “IFITM3 Restricts the Morbidity and Mortality Associated with Influenza.” *Nature* 484(7395): 519–23.

Fagerberg, Linn et al. 2013. “Contribution of Antibody-Based Protein Profiling to the Human Chromosome-Centric Proteome Project (C-HPP).” *Journal of Proteome Research* 12(6): 2439–48.

Fairfax, B. P. et al. 2014. “Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression.” *Science* 343(6175): 1246949–1246949.

Feuk, Lars et al. 2005. “Discovery of Human Inversion Polymorphisms by Comparative Analysis of Human and Chimpanzee DNA Sequence Assemblies.” *PLoS Genetics* 1(4): e56.

———. 2010. “Inversion Variants in the Human Genome: Role in Disease and Genome Architecture.” *Genome Medicine* 2(2): 11.

Feuk, Lars, Andrew R. Carson, and Stephen W. Scherer. 2006. “Structural Variation in the Human Genome.” *Nature Reviews. Genetics* 7(2): 85–97.

Franke, Andre et al. 2010. “Genome-Wide Meta-Analysis Increases to 71 the Number of Confirmed Crohn’s Disease Susceptibility Loci.” *Nature Genetics* 42(12): 1118–25.

Gaffney, Daniel J et al. 2012. “Dissecting the Regulatory Architecture of Gene Expression QTLs.” *Genome biology* 13(1): R7.

Génin, Emmanuelle et al. 2011. “Genome-Wide Association Study of Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis in Europe.” *Orphanet Journal of Rare Diseases* 6(1): 52.

Gibbs, J Raphael et al. 2010. “Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain.” *PLoS genetics* 6(5): e1000952.

Giglio, S. et al. 2001. “Olfactory Receptor-Gene Clusters, Genomic-Inversion Polymorphisms, and Common Chromosome Rearrangements.” *American Journal of Human Genetics* 68(4): 874–83.

González, Juan R. et al. 2014. “A Common 16p11.2 Inversion Underlies the Joint Susceptibility to Asthma and Obesity.” *The American Journal of Human Genetics* 94(3): 361–72.

Grundberg, Elin et al. 2012. “Mapping Cis- and Trans-Regulatory Effects across Multiple Tissues in Twins.” *Nature genetics* 44(10): 1084–89.

———. 2013. “Global Analysis of DNA Methylation Variation in Adipose Tissue from Twins Reveals Links to Disease-Associated Variants in Distal Regulatory Elements.” *The American Journal of Human Genetics* 93(5): 876–90.

GTEx Consortium. 2013. “The Genotype-Tissue Expression (GTEx) Project.” *Nature genetics* 45(6): 580–85.

Gu, Wenli, Feng Zhang, and James R Lupski. 2008. “Mechanisms for Human Genomic Rearrangements.” *PathoGenetics* 1(1): 4.

Hakonarson, Hakon et al. 2007. “A Genome-Wide Association Study Identifies KIAA0350 as a Type 1 Diabetes Gene.” *Nature* 448(7153): 591–94.

’t Hart, L. M. et al. 2013. “The CTRB1/2 Locus Affects Diabetes Susceptibility and Treatment via the Incretin Pathway.” *Diabetes* 62(9): 3275–81.

Hinds, D. A. 2005. “Whole-Genome Patterns of Common DNA Variation in Three Human Populations.” *Science* 307(5712): 1072–79.

Hoffmann, Ary A., and Loren H. Rieseberg. 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: 21–42.

Hormozdiari, Fereydoun, Can Alkan, Evan E Eichler, and S Cenk Sahinalp. 2009. “Combinatorial Algorithms for Structural Variation Detection in High-Throughput Sequenced Genomes.” *Genome research* 19(7): 1270–78.

Iafrate, A. John et al. 2004. “Detection of Large-Scale Variation in the Human Genome.” *Nature Genetics* 36(9): 949–51.

Innocenti, Federico et al. 2011. “Identification, Replication, and Functional Fine-Mapping of Expression Quantitative Trait Loci in Primary Human Liver Tissue.” *PLoS genetics* 7(5): e1002078.

International HapMap Consortium. 2003. “The International HapMap Project.” *Nature* 426(6968): 789–96.

International Multiple Sclerosis Genetics Consortium et al. 2007. “Risk Alleles for Multiple Sclerosis Identified by a Genomewide Study.” *The New England Journal of Medicine* 357(9): 851–62.

Ishimoto, Takatsugu et al. 2011. “CD44 Variant Regulates Redox Status in Cancer Cells by Stabilizing the xCT Subunit of System Xc(-) and Thereby Promotes Tumor Growth.” *Cancer Cell* 19(3): 387–400.

Johnson, Andrew D., and Christopher J. O’Donnell. 2009. “An Open Access Database of Genome-Wide Association Results.” *BMC medical genetics* 10: 6.

Krimbas, Costas B., and Powell, Jeffrey R. 1992. “Drosophila Inversion Polymorphism." *CRC Press*

De Jong, Simone et al. 2012. “Common Inversion Polymorphism at 17q21.31 Affects Expression of Multiple Genes in Tissue-Specific Manner.” *BMC genomics* 13: 458.

Kang, Hyun Min et al. 2008. “Efficient Control of Population Structure in Model Organism Association Mapping.” *Genetics* 178(3): 1709–23.

Kennington, W. Jason, Linda Partridge, and Ary A. Hoffmann. 2006. “Patterns of Diversity and Linkage Disequilibrium within the Cosmopolitan Inversion In(3R)Payne in Drosophila Melanogaster Are Indicative of Coadaptation.” *Genetics* 172(3): 1655–63.

Kidd, Jeffrey M. et al. 2008. “Mapping and Sequencing of Structural Variation from Eight Human Genomes.” *Nature* 453(7191): 56–64.

Kidd, Jeffrey M et al. 2010. “A Human Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms.” *Cell* 143(5): 837–47.

Kim, Min-Sik et al. 2014. “A Draft Map of the Human Proteome.” *Nature* 509(7502): 575–81.

Kim, S. S. et al. 1996. “A Novel Member of the RING Finger Family, KRIP-1, Associates with the KRAB-A Transcriptional Repressor Domain of Zinc Finger Proteins.” *Proceedings of the National Academy of Sciences of the United States of America* 93(26): 15299–304.

Kingsmore, Stephen F. et al. 2008. “Genome-Wide Association Studies: Progress and Potential for Drug Discovery and Development.” *Nature Reviews Drug Discovery* 7(3): 221–30.

Kirkpatrick, Mark. 2010. “How and Why Chromosome Inversions Evolve.” *PLoS Biology* 8(9): e1000501.

Kirkpatrick, Mark, and Nick Barton. 2006. “Chromosome Inversions, Local Adaptation and Speciation.” *Genetics* 173(1): 419–34.

Kleinjan, D. 1998. “Position Effect in Human Genetic Disease.” *Human Molecular Genetics* 7(10): 1611–18.

Koike, Asako et al. 2009. “Genome-Wide Association Database Developed in the Japanese Integrated Database Project.” *Journal of Human Genetics* 54(9): 543–46.

Korbel, Jan O et al. 2007. “Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome.” *Science (New York, N.Y.)* 318(5849): 420–26.

———. 2009. “PEMer: A Computational Framework with Simulation-Based Error Models for Inferring Genomic Structural Variants from Massive Paired-End Sequencing Data.” *Genome Biology* 10(2): R23.

Lai, Zhao et al. 2005. “Extensive Chromosomal Repatterning and the Evolution of Sterility Barriers in Hybrid Sunflower Species.” *Genetics* 171(1): 291–303.

Lakich, Delia, Haig H. Kazazian, Stylianos E. Antonarakis, and Jane Gitschier. 1993. “Inversions Disrupting the Factor VIII Gene Are a Common Cause of Severe Haemophilia A.” *Nature Genetics* 5(3): 236–41.

Lango Allen, Hana et al. 2010. “Hundreds of Variants Clustered in Genomic Loci and Biological Pathways Affect Human Height.” *Nature* 467(7317): 832–38.

Lappalainen, Tuuli et al. 2013. “Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans.” *Nature* 501(7468): 506–11.

Lau, Eric et al. 2012. “PKCε Promotes Oncogenic Functions of ATF2 in the Nucleus While Blocking Its Apoptotic Function at Mitochondria.” *Cell* 148(3): 543–55.

Lawson-Yuen, Amy, Juan-Sebastian Saldivar, Steve Sommer, and Jonathan Picker. 2008. “Familial Deletion within NLGN4 Associated with Autism and Tourette Syndrome.” *European journal of human genetics: EJHG* 16(5): 614–18.

Lee, Jennifer A., Claudia M. B. Carvalho, and James R. Lupski. 2007. “A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders.” *Cell* 131(7): 1235–47.

Leek, Jeffrey T, and John D Storey. 2007. “Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis.” *PLoS genetics* 3(9): 1724–35.

Lee, M. N. et al. 2014. “Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells.” *Science* 343(6175): 1246980–1246980.

Levy, Samuel et al. 2007. “The Diploid Genome Sequence of an Individual Human.” *PLoS biology* 5(10): e254.

Listgarten, Jennifer, Carl Kadie, and David Heckerman. 2010. “Correction for Hidden Confounders in the Genetic Analysis of Gene Expression.” *Proceedings of the National Academy of Sciences*. http://www.pnas.org/content/early/2010/08/30/1002425107 (July 14, 2014).

Lucas Lledó, José Ignacio, and Mario Cáceres. 2013. “On the Power and the Systematic Biases of the Detection of Chromosomal Inversions by Paired-End Genome Sequencing” ed. Zhanjiang Liu. *PLoS ONE* 8(4): e61292.

Lucas-Lledó, José, David Vicente-Salvador, Cristina Aguado, and Mario Cáceres. 2014. “Population Genetic Analysis of Bi-Allelic Structural Variants from Low-Coverage Sequence Data with an Expectation-Maximization Algorithm.” *BMC Bioinformatics* 15(1): 163.

Luo, Cheng et al. 2012. “Disrupted Functional Brain Connectivity in Partial Epilepsy: A Resting-State fMRI Study” ed. Olaf Sporns. *PLoS ONE* 7(1): e28196.

MacDonald, J. R. et al. 2014. “The Database of Genomic Variants: A Curated Collection of Structural Variation in the Human Genome.” *Nucleic Acids Research* 42(D1): D986–92.

Martinez-Fundichely, A. et al. 2014. “InvFEST, a Database Integrating Information of Polymorphic Inversions in the Human Genome.” *Nucleic Acids Research* 42(D1): D1027–32.

McKernan, K. J. et al. 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): 1527–41.

McVean, Gil A. et al. 2012. “An Integrated Map of Genetic Variation from 1,092 Human Genomes.” *Nature* 491(7422): 56–65.

Medvedev, Paul, Monica Stanciu, and Michael Brudno. 2009. “Computational Methods for Discovering Structural Variation with next-Generation Sequencing.” *Nature Methods* 6(11 Suppl): S13–20.

Mills, Ryan E. et al. 2011. “Mapping Copy Number Variation by Population-Scale Genome Sequencing.” *Nature* 470(7332): 59–65.

Miyagawa, Taku et al. 2008. “Variant between CPT1B and CHKB Associated with Susceptibility to Narcolepsy.” *Nature Genetics* 40(11): 1324–28.

Moffatt, Miriam F. et al. 2010. “A Large-Scale, Consortium-Based Genomewide Association Study of Asthma.” *New England Journal of Medicine* 363(13): 1211–21.

Montgomery, Stephen B. et al. 2010. “Transcriptome Genetics Using Second Generation Sequencing in a Caucasian Population.” *Nature* 464(7289): 773–77.

Mudhasani, Rajini et al. 2013. “IFITM-2 and IFITM-3 but Not IFITM-1 Restrict Rift Valley Fever Virus.” *Journal of virology* 87(15): 8451–64.

Myers, Amanda J et al. 2007. “A Survey of Genetic Human Cortical Gene Expression.” *Nature genetics* 39(12): 1494–99.

Navarro, Arcadi, and Nick H. Barton. 2003. “Accumulating Postzygotic Isolation Genes in Parapatry: A New Twist on Chromosomal Speciation.” *Evolution; International Journal of Organic Evolution* 57(3): 447–59.

Nica, Alexandra C et al. 2010. “Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations.” *PLoS genetics* 6(4): e1000895.

O’Neill, R. J., M. D. B. Eldridge, and C. J. Metcalfe. 2004. “Centromere Dynamics and Chromosome Evolution in Marsupials.” *The Journal of Heredity* 95(5): 375–81.

Onishi-Seebacher, Megumi, and Jan O. Korbel. 2011. “Challenges in Studying Genomic Structural Variant Formation Mechanisms: The Short-Read Dilemma and beyond.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 33(11): 840–50.

OpenStax community. 2014. “Chromosomal Basis of Inherited Disorders.” *OpenStax-CNX*. http://cnx.org/content/m44483/latest/?collectin=col11448

Ortíz-Barrientos, Daniel, Jane Reiland, Jody Hey, and Mohamed A. F. Noor. 2002. “Recombination and the Divergence of Hybridizing Species.” *Genetica* 116(2-3): 167–78.

Pang, Andy W et al. 2010. “Towards a Comprehensive Structural Variation Map of an Individual Human Genome.” *Genome Biology* 11(5): R52.

Pang, Andy Wing Chun et al. 2013. “Mechanisms of Formation of Structural Variation in a Fully Sequenced Human Genome.” *Human Mutation* 34(2): 345–54.

Perner, Sven et al. 2008. “EML4-ALK Fusion Lung Cancer: A Rare Acquired Event.” *Neoplasia (New York, N.Y.)* 10(3): 298–302.

Pickrell, Joseph K et al. 2010. “Understanding Mechanisms Underlying Human Gene Expression Variation with RNA Sequencing.” *Nature* 464(7289): 768–72.

Pique-Regi, Roger et al. 2011. “Accurate Inference of Transcription Factor Binding from DNA Sequence and Chromatin Accessibility Data.” *Genome Research* 21(3): 447–55.

Rausch, T. et al. 2012. “DELLY: Structural Variant Discovery by Integrated Paired-End and Split-Read Analysis.” *Bioinformatics* 28(18): i333–39.

Redon, Richard et al. 2006. “Global Variation in Copy Number in the Human Genome.” *Nature* 444(7118): 444–54.

Rice, William R. 1987. “The Accumulation of Sexually Antagonistic Genes as a Selective Agent Promoting the Evolution of Reduced Recombination between Primitive Sex Chromosomes.” *Evolution* 41(4): 911–14.

Rieseberg, Loren H. 2001. “Chromosomal Rearrangements and Speciation.” *Trends in Ecology & Evolution* 16(7): 351–58.

Salm, Maximilian P A et al. 2012. “The Origin, Global Distribution, and Functional Impact of the Human 8p23 Inversion Polymorphism.” *Genome research* 22(6): 1144–53.

Schadt, Eric E. et al. 2008. “Mapping the Genetic Architecture of Gene Expression in Human Liver” ed. Goncalo Abecassis. *PLoS Biology* 6(5): e107.

Schymick, Jennifer C. et al. 2007. “Genome-Wide Genotyping in Amyotrophic Lateral Sclerosis and Neurologically Normal Controls: First Stage Analysis and Public Release of Data.” *Lancet Neurology* 6(4): 322–28.

Sebat, Jonathan et al. 2004. “Large-Scale Copy Number Polymorphism in the Human Genome.” *Science (New York, N.Y.)* 305(5683): 525–28.

Shaikh, T. H. et al. 2000. “Chromosome 22-Specific Low Copy Repeats and the 22q11.2 Deletion Syndrome: Genomic Organization and Deletion Endpoint Analysis.” *Human Molecular Genetics* 9(4): 489–501.

Sharp, Andrew J., Ze Cheng, and Evan E. Eichler. 2006. “Structural Variation of the Human Genome.” *Annual Review of Genomics and Human Genetics* 7(1): 407–42.

Sindi, Suzanne S et al. 2012. “An Integrative Probabilistic Model for Identification of Structural Variation in Sequencing Data.” *Genome Biology* 13(3): R22.

Soda, Manabu et al. 2007. “Identification of the Transforming EML4–ALK Fusion Gene in Non-Small-Cell Lung Cancer.” *Nature* 448(7153): 561–66.

Soranzo, Nicole et al. 2010. “Common Variants at 10 Genomic Loci Influence Hemoglobin A_1(C) Levels via Glycemic and Nonglycemic Pathways.” *Diabetes* 59(12): 3229–39.

Spirito F. 1998. *Species and Speciation*. New York: Oxford New York : Oxford Univer-sity Press.

Stankiewicz, Paweł, and James R. Lupski. 2010. “Structural Variation in the Human Genome and Its Role in Disease.” *Annual Review of Medicine* 61(1): 437–55.

Stebbins, G. L. 1958. “The Inviability, Weakness, and Sterility of Interspecific Hybrids.” *Advances in Genetics* 9: 147–215.

Stefansson, Hreinn et al. 2005. “A Common Inversion under Selection in Europeans.” *Nature genetics* 37(2): 129–37.

Stegle, Oliver et al. 2012. “Using Probabilistic Estimation of Expression Residuals (PEER) to Obtain Increased Power and Interpretability of Gene Expression Analyses.” *Nature Protocols* 7(3): 500–507.

Steinberg, Karyn Meltz et al. 2012. “Structural Diversity and African Origin of the 17q21.31 Inversion Polymorphism.” *Nature genetics* 44(8): 872–80.

Stevison, Laurie S., Kenneth B. Hoehn, and Mohamed A. F. Noor. 2011. “Effects of Inversions on within- and between-Species Recombination and Divergence.” *Genome Biology and Evolution* 3: 830–41.

Stranger, Barbara E, Alexandra C Nica, et al. 2007. “Population Genomics of Human Gene Expression.” *Nature genetics* 39(10): 1217–24.

Stranger, Barbara E, Matthew S Forrest, et al. 2007. “Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes.” *Science (New York, N.Y.)* 315(5813): 848–53.

Stranger, Barbara E et al. 2012. “Patterns of Cis Regulatory Variation in Diverse Human Populations.” *PLoS genetics* 8(4): e1002639.

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): 235–37.

The International HapMap Consortium. 2005. “A Haplotype Map of the Human Genome.” *Nature* 437(7063): 1299–1320.

Troeger, A. et al. 2012. “RhoH Is Critical for Cell-Microenvironment Interactions in Chronic Lymphocytic Leukemia in Mice and Humans.” *Blood* 119(20): 4708–18.

Turner, Daniel J. et al. 2006. “Assaying Chromosomal Inversions by Single-Molecule Haplotyping.” *Nature Methods* 3(6): 439–45.

Tuzun, Eray et al. 2005. “Fine-Scale Structural Variation of the Human Genome.” *Nature Genetics* 37(7): 727–32.

Urrutia, Raul. 2003. “KRAB-Containing Zinc-Finger Repressor Proteins.” *Genome Biology* 4(10): 231.

Veyrieras, Jean-Baptiste et al. 2008. “High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation.” *PLoS genetics* 4(10): e1000214.

Wang, G., E. Yang, C. L. Brinkmeyer-Langford, and J. J. Cai. 2014. “Additive, Epistatic, and Environmental Effects Through the Lens of Expression Variability QTL in a Twin Cohort.” *Genetics* 196(2): 413–25.

Wang, Jun et al. 2008. “The Diploid Genome Sequence of an Asian Individual.” *Nature* 456(7218): 60–65.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature Reviews Genetics* 10(1): 57–63.

White, M. J. D. 1978. *Modes of Speciation*. San Francisco: W. H. Freeman.

Wilson, M. G., J. W. Towner, G. S. Coffin, and I. Forsman. 1970. “Inherited Pericentric Inversion of Chromosome No. 4.” *American Journal of Human Genetics* 22(6): 679–90.

Xia, K. et al. 2012. “seeQTL: A Searchable Database for Human eQTLs.” *Bioinformatics* 28(3): 451–52.

Yang, Tsun-Po et al. 2010. “Genevar: A Database and Java Application for the Analysis and Visualization of SNP-Gene Associations in eQTL Studies.” *Bioinformatics (Oxford, England)* 26(19): 2474–76.

Yu, Wei-Hsuan, J. Frederick Woessner, John D. McNeish, and Ivan Stamenkovic. 2002. “CD44 Anchors the Assembly of matrilysin/MMP-7 with Heparin-Binding Epidermal Growth Factor Precursor and ErbB4 and Regulates Female Reproductive Organ Remodeling.” *Genes & Development* 16(3): 307–23.

Zeller, Tanja et al. 2010. “Genetics and beyond--the Transcriptome of Human Monocytes and Disease Susceptibility.” *PloS one* 5(5): e10693.