Positive selection in humans: from single genes to interaction maps

Pierre Luisi

TESI DOCTORAL UPF / ANY 2014

DIRECTORS DE LA TESI

Jaume Bertranpetit and Hafid Laayouni
Departament of Experimental and Health Sciences

Universitat Pompeu Fabra
Barcelona

For my family…

You find the world is a very puzzling place and if you are willing to be puzzled, you can learn. [...] Learning comes from asking ”Why do things work like that, not some other way?”

Is the man who is tall happy?
An Animated Conversation with Noam Chomsky
Documentary by Michel Gondry
NOAM CHOMSKY

Acknowledgments

> OH! I get by with a little help from my friends.
>
> *S*gt pepper's lonely hearts club band
> THE BEATLES

I wished the moment to acknowledge all the people who supported me to accomplish my PhD would never come. Finalizing my thesis feels like the end of an incredible period, rather than an achievement. The time I have spent in Barcelona, including at the IBE, has been really amazing. I had a lot of pleasure and I have grown both scientifically and personally during those last years. I owe it to all the incredible people I have interacted with, and it would be difficult to exhaustively thank everyone sufficiently.

First of all, I wish to naturally thank my supervisors. Jaume and Hafid you have been supportive in so many ways. You made working with you so easy. I really appreciate everything you have done to make my life easier in Barcelona. From the extra beginning you proved that humanely you care about your students. Hafid, you helped me to set up in Barcelona when the first wage lasted to come, then you always offered your support when I had some uneasy moments to go through. Jaume, you tried to find a solution so that my partner at that time could stay in Barcelona. I am sure not that all the supervisors would do what you both did. I am very thankful for that. More importantly for this thesis, I learnt a lot from your supervision. You both have been very patient when I lost faith in some projects. You accepted my directness when I was beset by impatience and desperation. You took time to listen to all of my doubts and worries. Especially you Hafid, you supported me to answer those doubts and find solutions when I felt a project was blocked or even not worth to follow. Jaume, I really enjoyed learning from your vision of science, always trying to put our scientific activities within a bigger biological frame, always anticipating the knowledge and technical outbreaks to come. I also really

wish to thank both of you for all the freedom you gave me to follow some projects with co-workers outside the IBE. I have learnt a lot during this PhD, both from the independence I had and directly from yourselves. I could not expect better supervisors, and I feel ready to keep on in my scientific career thanks to all what I have learnt under your supervision.
I would like to thank former and present members of the Evolutionary System Biology Group led by Jaume Bertranpetit for their scientific support and their kindness during those years. Thank to the ones who shared my office, Gio, Ludo, Brandon, Marc and Bego, for standing all my —organized — chaos and my overflowing energy in front of the bright screen.
I wish to thank the "Boosting team", Gio, Marc and Johannes for this very interesting project. It really boosted my learning process!
I also have to thank everybody in BioEvo. Working there has been amazing. A special thank goes to Txema, Jordi and Angel for this very robust cluster you have put together and you manage efficiently, it makes our work so easier. Thanks to Arcadi, Diego and David for the useful comments on the interactome paper. Thanks Ferran for your tips on the Postdoc positions.

I want to thank people outside the IBE with whom I have worked. First, I naturally have to thanks David Alvarez-Ponce. I started my PhD while you were finishing yours, you are now Assistant Professor at the University of Nevada, USA. You have moved forward quite fast and I have learnt a lot from your rigour. Your contribution in the papers we co-signed is amazing and made them of much higher quality, as much for the content as for the style. Thanks Mario Fares for your contribution in the interactome paper.
I would like to thanks members of Anna Di Rienzo lab for their welcome at the University of Chicago back in 2012. I would also like to thank Oscar Lao at Erasmus MC University, Rotterdam, the Netherlands, for the exciting on-going project on selection on standing variants, as well as Tom Gilbert at the Centre for GeoGenetics, Copenhagen, Denmark, for the invitation to Copenhagen and for having trusted in me for the promising project on ancientDNA.

Last of the people I have been working with, but not least, I am very thankful to Audrey, Emmanuelle and Blandine for including me in your projects on pharmacogenetics. I especially have a thought to Blandine for all the hours trying to solve some issues by chat. Your thesis has been fraught with difficulties, but your forceful personality allowed you to keep on. I admire you for that. Thanks you Audrey for all those years working together. It is such a pleasure to work with someone like you. Thanks for having trusted in me while I was still a bachelor student, for all advices in my career, all the invitations to Paris and the on-going projects with the UMR-216. Thanks for having brought me to know this research unit, and thanks for having suggested me as a workshop facilitator at the University of Ghana. I am sure our collaboration is just at its beginning and I am looking forward to working with you on future projects. Thanks Laure for the wonderful work you have done on the *HLA-G* study.

Gracias a la banda del Tupper//Taper/Toper/Täper/Tuper club por todos estos magníficos almuerzos compartiendo comida. Cada día, el almuerzo fue una experiencia increíble. Estoy muy orgulloso de ser parte de este experimento social, donde el grupo vale, donde el individualismo no tiene lugar. Aunque al principio eran todos compañeros de trabajo nada más (dedicatoria a María!), hemos crecido una amistad muy fuerte y me siento muy cercano a todos los miembros, sean pasados o presentes. Hemos tenido momentos bien divertidos almorzando en esta terraza pero también afuera del PRBB! Gracias a (por orden alfabético, así no hay celos posibles... sé que todos son un poco susceptibles) Alicia, Diego, Elena, Fede, Johannes, Juan, Katharina, Ilaria, Lara, Lisa, Marc, Marco, María, Miruna, Nino y Vero. Les deseo seguir así mucho tiempo, no sin envidia. Para citar a un miembro (le dejaré el beneficio del anonimato): "El taperclub es lo mejor que me ha pasado en el doctorado... by far!". Prometo que cuando esté de visita en Barcelona les cocinaré mi especialidad, una leyenda del éxito culinario: los míticos púlpitos!
Quiero agradecer especialmente a Diego por todas las horas escuchando mis pajas mentales, hablando de las tuyas y sobre todo por el apoyo en cualquier situacíon; a Elena (y Marc seu) por —entre muchas más cosas

— Granada y por dejarme dormir en su sofá tantas veces; a Marc y María, así como a Diego (y Abril), por la ayuda cuando estaba buscando hogar (para variar); a Nino por las horas de palas en la playa y por Sicilia. Gracias Marc (apareces demasiado en estos agradecimientos, eh!) por todas las pausas "cigarouuu". Hemos sublimado nuestro espíritu crítico sobre nuestros proyectos, ciencia y política. Sólo para estas pausas, valió la pena empezar a fumar tan tarde... Ahora que me voy, lo puedo dejar!
Gracias a los Anomalocaris por este año en la liga de basket, especialmente a Diego, Ignasi y Javi. Me la pasé muy bien. Ojalá volvamos a jugar juntos pronto. Gracias a todos los locos que "madrugan" para jugar un partido de fútbol antes del "curo" una vez a la semana, a los del IMIM y del CRG por los partidos semanales de basket, a todos que bajan a jugar a beachvolley en el verano y a Javi por los demasiados pocos partidos de tenis.

Quiero agradecer a la gente con quien me ha encantando vivir estos años. Karla, por esos años felices juntos, gracias por haber probado tanto para que funcionara nuestra historia. Así va la vida, suerte en tu nuevo camino. Gracias a Tania y Martha por haber conseguido hacerme sentir en casa desde el primer día. Se me hace difícil marchar de Barcelona también porque me encanta vivir con ustedes dos, mamasitas!
Gracias Nata, el año que hemos coincidido en Barcelona fue genial. Nos vemos en Perú prontito.

Je veux aussi remercier MG Gold pour cette fameuse chanson dont je tairai le titre... Merci aux loustiques, mes amis de longue date, vous me permettez de me vider la tête régulièrement. Merci donc Alex, Bestion, Banette, Guyves, Laurent, Manu, Mehdi, Mehdi (l'autre), Nico, Seb et Zig pour les moments intenses que l'on passe à chaque rencontre.

Enfin, je veux remercier le clan Luisi! Je sais la chance que j'ai de faire partie d'une telle famille, une famille soudée et où les choix des uns et des autres sont respectés. Merci Anne et Michel pour tout le soutien que vous m'avez toujours apporté et pour votre amour. Je suis vraiment fier de l'éducation que voue nous avez inculquée et de me sentir si proche de vous. Merci Gates et Titou pour les bons moments que l'on passe ensemble. C'est assez malheureux qu'ils soient si peu fréquents, mais je pense à vous! Merci à Grand-Père et Grand-Mère pour tout l'amour que vous nous avez donné.

Encore merci à tous!
Thanks again to everyone!
Otra vez, gracias a todos!

Pierre Luisi Barcelona, Spain Septembre 6, 2014

This research was supported by a Ph.D. fellowship from "Acción Estratégica de Salud, en el marco del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2008–2011" from Instituto de Salud Carlos III.

Abstract

From Darwin's *Origin of the Species* to the recent wealth in genomic data, many biologists have focused their research on understanding how natural selection has shaped the variability among and within species. Although theoretical and empirical advances have been remarkable, most biological mechanisms underlying the molecular basis of human adaptation remain to be elucidated. The selectionist view of adaptation accounted for the bias towards independent gene evolution. Most published studies aiming at detecting positive selection using either polymorphism or divergence data have been performed using a gene-candidate or a genome-wide scan approach, as described in the two first articles presented here. However, gene evolution is largely influenced by the biological context in which the encoded protein performs its intrinsic function(s). The phenotype, not the genotype, is at the interface with natural selection. Thus, in order to understand gene evolution, and particularly when considering adaptive selection, it is crucial to reduce the gap between genotype and phenotype. Genes and proteins do not act in isolation, but rather interact one with others in order to perform a given biological function. Therefore, when studying natural selection at molecular level one promising framework is to consider gene networks, as described in the two last articles of the present thesis. Analyses of gene networks describing the Insulin/TOR transduction signalling cascade and the whole protein-protein physical interaction map hold very striking results. Namely, genes acting at the core of both networks, thus having either more effect on a given phenotype or more pleiotropic effects within the organism, are more likely to be targeted by recent positive selection, as inferred using polymorphism data.

Resumen

Desde el “Origen de las Especies” de Darwin a la reciente revolución genómica, muchos biólogos han centrado su investigación en la comprensión de cómo la selección natural ha dado forma a la variabilidad entre y dentro de las especies. Aunque, los avances teóricos y empíricos han sido notables, la mayoría de los mecanismos biológicos que subyacen a las bases moleculares de la adaptación biológica aún no están suficientemente esclarecidos. La visión seleccionista de adaptación marcó el sesgo de los estudios evolutivos hacia el análisis de genes individuales. La mayoría de estudios publicados destinados a la detección de la selección positiva utilizando datos de polimorfismo o de divergencia se han realizado utilizando un gen candidato o un enfoque de exploración genómica, como se describe en los dos primeros artículos presentados en la presente tesis. Sin embargo, la evolución de genes está muy condicionada por el contexto biológico en el que cada gen realiza su función intrínseca, siendo el fenotipo, y no el genotipo, su materia primaria. Por lo tanto, a fin de comprender la evolución de genes, y en particular cuando se considera la evolución adaptativa, es crucial reducir la brecha entre el genotipo y el fenotipo. Los genes y las proteínas no actúan de manera aislada, sino que interactúan entre sí con el fin de realizar una función biológica determinada. Por lo tanto, un marco prometedor al estudiar la selección natural a nivel molecular seria considerar las redes de genes, como se describe en los dos últimos artículos de la presente tesis. Los análisis de los datos de polimorfismo genético, tanto de los genes que componen la vía de la insulina, cómo de los todos los genes descritos en los mapas físicos de interacción proteína-proteína tienen resultados muy sorprendentes: los genes que actúan en el núcleo de ambas redes, teniendo así más efecto sobre un determinado fenotipo o más efectos pleótropicos dentro del organismo, tienen más probabilidades de ser el blanco de la selección positiva reciente.

Resum

Des del ”Origen de les Espècies”de Darwin fins a la recent revolució genòmica, molts biòlegs han centrat la seva investigació a la comprensió de com la selecció natural ha donat forma a la variabilitat existent entre i dins de les espècies. Tot i que els avenços teòrics i experimentals han estat notables, la majoria dels mecanismes biològics subjacents a les bases moleculars de la adaptació biològica no estan prou aclarits.

Els estudis evolutius per entendre l'adaptació estan esbiaixats cap a la comprensió de l'acció de gens individuals. La majoria d'estudis publicats destinats a la detecció de la selecció positiva (adaptativa) utilitzant dades de polimorfisme o de divergència, s'han realitzat utilitzant o bé gens candidats o bé un enfocament d'escaneig de tot el genoma, tal com es descriu en els dos primers articles presentats en aquesta tesi.

No obstant això, l'evolució de gens està molt condicionat pel context biològic en el qual la proteïna codificada per cada gen realitza la seva pròpia funció. El fenotip, no el genotip, és a la interfície directa amb la selecció natural. Per tant, per tal per entendre l'evolució dels gens, i en particular quan es considera la selecció adaptativa, és crucial reduir la separació entre el genotip i el fenotip. Els gens i les proteïnes no actuen de manera aïllada, sinó que interactuen uns amb altres per tal de realitzar una funció biològica determinada. Per tant, per l'estudi de la selecció natural a nivell molecular, un marc prometedor és considerar les xarxes de gens, tal com es descriu en els dos últims articles de la present tesi.

Les anàlisis de la xarxa de gens que descriuen la cascada de transducció de senyals de la insulina/TOR i del conjunt total del mapa d'interaccions físiques proteïna-proteïna en humans tenen resultats molt sorprenents. De fet, els gens que actuen en el nucli de totes dues xarxes (i que per tant tenen més impacte en un determinat fenotip i més efectes pleiotròpics dins de l'organisme), tenen més probabilitats de ser la diana de la selecció positiva recent que no pas els gens amb menys interaccions.

PREFACE

> Most achievements in science are to a certain degree group efforts.
>
> *Speech at the Nobel Banquet in Stockholm, December 10, 1960*
> WILLARD LIBBY

In 1859, with his masterpiece *The Origin of the Species* Charles R. Darwin laid the cornerstone of evolutionary biology. Nevertheless, it is not until the 1920s that the field properly began with the visionary work from few theoreticians. Indeed, at that time and in the following decades, Ronald R. Fisher, Sewall Wright and J. B. S Haldane developed the *modern evolutionary synthesis* through the formulation of the mathematical background for population genetics. Since then, this field has been long lasting. The insights from population genetics into evolutionary biology are extraordinary. To broaden the understanding on the main evolutive forces at play, cross-talks between theoretical development and empirical observations have proved to be essential. During almost one century of population genetics and evolutionary biology, lively discussions have been frequently rekindled thanks to many empirical and theoretical breakthroughs. In the last few years, evolutionary biology and population genetics have been living a very exciting moment. Indeed, with the advent of high-throughput technologies to produce large amount of data with increasing confidence, the so-called "*-omics*" era could begin. Great amount of data from genomics, interactomics, metabolomics, transcriptomics, epigenomics, etc., are now available. Such wealth in data may seem overwhelming and much effort is still required to process and fully understand it. Although this is challenging, biology is now moving from a traditionally reductionist view and we are more and more able to consider many layers of complexity to answer many interesting biological issues and interrogate and/or improve the models traditionally used. Particularly, evolutionary biology is now on the path to leave behind the gene-centric view which led the field for many years, as more information on the gene function and context can be included to attempt to bridge the gap between genotype and phenotype.

Considering the biological pathways in which genes participate is one of the emerging frameworks for evolutionary biology studies. Very few studies on how natural selection acts within gene networks have been published to date. Specifically, the impact of positive selection across gene networks has been overlooked. The present thesis introduces the first study of the relationship between gene adaptive evolution and the position occupied by the protein it encodes in a given functional pathway. Then, it focuses on a study at much larger scale which demonstrates how challenging it is to fully understand the molecular mechanisms driving adaptive evolution.

The kind of analyses presented in this thesis deeply rely on an accurate representation of the interactions among proteins. However, although technical and technological efforts have been made to produce such an amount of data, it remains a relatively significant number of errors. When studying small-scale networks, such as those representing specific biological pathways, the errors can be addressed retrieving information from the literature. Therefore, earlier efforts by many researchers makes possible technical and technological progresses to improve each day the accuracy of the produced data. A special thought also goes to all the persons who manually curate the databases by retrieving information from decades of efforts in biochemistry and molecular biology. The modest contribution to the field of evolutionary biology presented in this thesis would not have been possible without all this people.

Contents

Abbreviations

AVK : Anti Vitamine K
Bp : Base pair
BGS : Background Selection
CEU : European ancestry population from Utah
CHB : Han Chinese population from Beijing
CDS : Coding Sequence
CCDS : Conserved Coding Sequence
CNV : Copy Number Variant
CMS : Composite Multiple Score
DAF : Derived Allele Frequency
EHH : Extended Haplotype Homozygosity
ENCODE : Encyclopedia of DNA Elements
eQTL : expression Quantitative Trail Locus
FAF : Final Allele Frequency
FDR : False Discovery Rate
FGM : Fisher's Gemotric Model of Adaptation
FPR : False Positive Rate
GWAS : Genome-Wide Association Study
iHS : integrated Haplotype Score
IS : Individual Site
JPT : Japanese population from Tokyo
Kb : Kilobase
KYA : Thousand Years Ago
KEGG: Kyoto Encyclopedia of Genes and Genomes

HGDP : Human Genome Diversity Panel
HPRD : Human Protein Reference Database
IT pahway : Insulin/TOR transduction pathway
LD : Linkage Disequilibrium
LRH : Long Range Haplotype
MAF : Minor Allele Frequency
MAPK : Mitogen Activated Protein Kinase
Mb : Megabase
MCMC : Markov Chain Monte Carlo
MK test : McDonald-Kreitman test
MRCA : Most Recent Common Ancestor
mRNA : messenger RNA
MYA : Million Years Ago
mtDNA : mitochondrial DNA
NGS : Next-Generation Sequencing
PANTHER : Protein Analysis Through Evolutionary Relationships
PIN : Protein-protein Interaction Network
PPI : Physical Protein-protein Interaction
PS : Pooled Sites
QTL : Quantitative Trait Locus
SFS : Site Frequency Spectrum
SNP : Single Nucleotide Polymorphism
SNV : Single Nucleotide Variant
SV : Structural Variant
SVM : Support Vector Machine
TF : Transcription Factor
TLR : Toll-Like Receptor
UTR : Untranslated Region
XP-EHH : Cross-population Extended Haplotype Homozygosity
Y2H : Yeast 2-Hybrids
YRI : Yoruba population from Nigeria

List of Figures

List of Tables

Part I

Introduction

Chapter 1

BACKGROUND

> Nothing in biology makes sense except in the light of evolution.
>
> THEODOSIUS DOBZHANSKY

> Darwin would have loved DNA.
>
> LINDELL BROMHAM

1.1 The rise of *Homo Sapiens* and its history.

Modern human lineage (*Homo Sapiens*) diverged from its closer living relative, the chimpanzee, about six million years ago (6 MYA). During all those years, many different ancient hominin lineages appeared and disappeared, some of them being ancestors of modern humans [1]. Modern humans emerged some 200 thousand years ago (200 KYA) somewhere in Africa. The oldest fossil that has been classified as being the remain of a modern human was found in Ethiopia and dates to about 195 KYA [2]. The relationship and boundaries among the hominin lineages remain

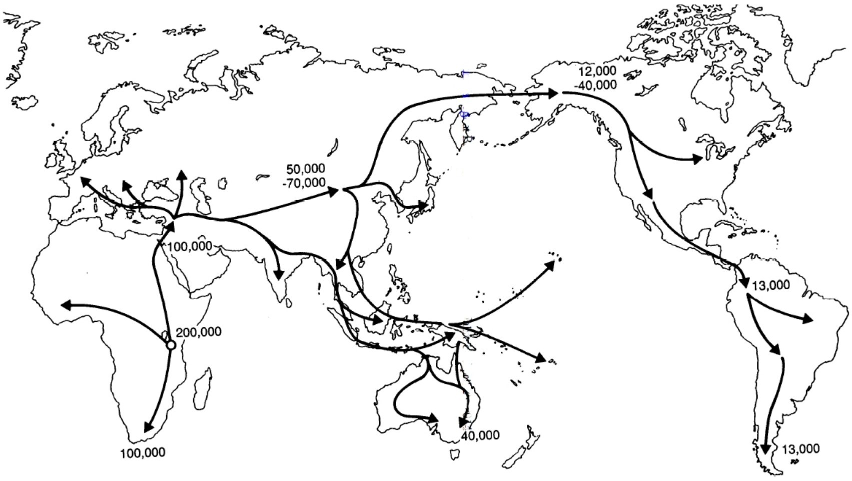


Figure 1.1: A map of prehistoric diaspora of modern humans. From [5].

much debated and are currently revisited thanks to the Next-Generation Sequencing (NGS) technology (for a review see [1]). Indeed, recent studies appointed to some gene flow between some past modern human populations and our extinct relatives Neandertals and Denisovans (for a review see [3]). Although it is still unclear what are the specific morphological features that make modern humans different from the other lineages, some consensus have been reached: the globular shape of the skull and the face with its particular degree of retraction are modern human specific [4]. The fact that modern human fossils dated before 45 KYA have never been found out of Africa suggest that Homo sapiens migrated to Eurasia and beyond much after its first appearance in Africa. Although the routes followed while migrating out-of-Africa remain debated, we now acknowledge that modern humans reached the Americas from Siberia ~ 15-20 KYA, Oceania from East Asia ~ 50 KYA and the Pacific islands from nearby continental lands ~ 5 KYA (Figure 1.1). Archaeological and genetic data are consistent enough to accept this view of modern human diaspora across the globe [3]. Around ~ 10 KYA occurred the Mesolithic-

Neolithic transition during which agricultural life-style appeared independently in several regions across the globe. This dramatic change from nomadic hunter-gatherer communities to more sedentary agriculturist ones allowed a dramatic expansions of human populations and cultural and social revolution with extensive technological improvements. Much effort have been put into building demographic models to explain both the human expansion over different time-scales, and genetic data have made possible to test them. By looking at genetic diversity and reconstructing genetic phylogenies, we can infer the root (i.e. the common ancestor) of our lineages and thus, trace back our origin. Maternal or paternal specific markers, namely markers located on mitochondrial DNA (mtDNA) and Y-chromosome were the most used because besides being sex-specific, and thus having a simple mode of inheritance, are also non-recombinant, facilitating the analyses. The out-of-Africa hypothesis have been confirmed because the human genetic diversity decrease with distance from Africa and populations out-of-Africa also present unique variants thought to be gained after their migration.

The recent wealth in genetic diversity data and the unprecedented power of computational modeling approaches allowed an increase power to infer the history of the human populations. Those demographic models include several past demographic events which are responsible for the observed variation patterns and genetic diversity in the current human population across the globe. We can distinguish among two families. The first kind of models considers that the human expansion occurred through several founding events, where an initial population increases in size and, in turn, a subset of individuals found a subsequent population and so on and so forth [6–9]. The second class of models assumes a single out-of-Africa event to Europe and Asia with subsequent population bottlenecks and expansions. They also include migration among populations that can vary across time. Most of these models only consider three populations representing African, European and Asian continents and, therefore, are obviously quite simplistic. However, they describe sufficiently past human demography for many purposes in genetic studies focusing on populations from these three continents. The most used model of this kind has been so

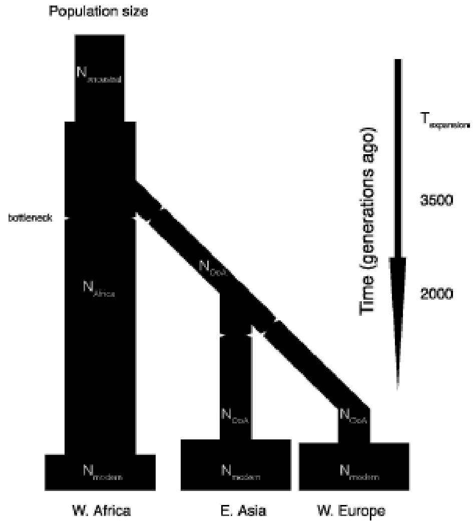


Figure 1.2: A model describing the modern human expansion with one single out-of-Africa events. From [10].

far the one calibrating on HapMap III genotype data (www.hapmap.org) using COSI coalescent simulator [10] and shown in Figure 1.2. Recently other models with more complex demographic histories or more populations have been implemented (e.g. in [11, 12]) as described in Section 1.5.3.

1.2 Human genetic variation.

1.2.1 Types of genetic variation.

Although the genetic differences among two individuals has consequences on the phenotypic variability, the genetic contribution to phenotypes has not been fully established and it is one of the main challenges in the 21st century. Even though the phenotypic variability among individuals seems important, any randomly pair of individuals in the world have on average only 0.1% sequence difference. This means that two human genomes share 99.9% of their variants. Recent available sequencing of human DNA provided a detailled description of the variants segregating in the genome of healthy individuals. Among the variants, there are substitutions and insertions/deletions, and they can be divided into three categories (Figure 1.3) according to their corresponding number of base pairs (bps): (1) structural variants (SVs) encompassing from few kilobases (> 10Kb) to few megabase, including large deletions and insertions, inversions, macrosatellites and Copy Number Variants (CNVs); (2) SVs encompassing few hundreds of base pairs such as medium sized insertions and deletions and minisatellites (repeats of 10-100 bps); and (3) variants of few bps such as small insertions and deletions, microsatellites (repeats of 2-6 bps) and single base pair substitutions, called Single Nucleotide Variants (SNVs) or traditionally Single Nucleotide Polymorphism (SNPs).

In the present thesis, most analyses have been performed using SNVs which are the most common and the most studied type of variation in the human genome. Single nucleotide polymorphisms are divided into two kinds. The transitions substitute either a pyrimidine to another pyrimidine (C to T or T to C) or a purine to another purine (A to G or G to A) while the transversions substitute a purine for a pyrimidine, or vice-versa. It has been observed 2-fold enrichment of segregating transitions as compared to transversions. The potential explanation would be that one purine (or pyrimidine) can be altered to the other purine (or pyrimidine), while it is impossible to chemically alter a purine to a pyrimidine (and vice versa). Another explanation could be that enzymes involved in DNA replication and correction are not able to correct transitions as well as transversions. The mutation rate can be estimated through comparative genomics (phylogenetic estimation), analysis of the frequency of new disease loci in human populations (direct estimation) or using biochemical knowledge of the DNA replication process (biochemical method). Depending on the studies, the mutation rate differ but has been estimated to be of the order of 10^{-8} per base per generation [9, 13, 14]. The mutation rate varies across the genome. For example, the CpG dinucleotide is a mutation hotspot, with a mutation rate ~ 10-fold higher than other base pairs and with a strong tendency to mutate to TpG or CpA because of the higher rate for transition than for transversion.

Beyond the genotypes, one can study the haplotypic variability. Haplotypes are the combinations of alleles that are inherited together because they are carried by the same chunk of the chromosome which has not been cut by any recombination event during meosis. Two loci are in linkage disequilibrium (LD) if there is a specific combination of their alleles that are observed on the same haplotype more often than expected at random. Knowing the haplotypes provide valuable information about ancestry and inheritance to perform evolutionary studies. Estimating haplotypes experimentally appears to be harsh, time-consuming and quite expensive. As a consequence, many computational algorithms have been implemented to infer the haplotypes from genotypes. Those algorithms are mainly Markov Chain Monte Carlo (MCMC) methods within a Bayesian framework [15]. Mutation at genotypic level is obviously responsible for creating new haplotypes. However, recombination is the main force driving haplotypic diversity. Recombination rates are not uniformly distributed along the genome: there are recombination hotspots. Therefore, allelic combinations are shaped in a haplotype-block manner. Recombination hotspots are differently distributed in the genome according to the population, hence, recombination events are responsible for haplotypic diversity among populations [16]. Finally, gene conversion (non reciprocal transfer of genetic variants from one chromosome to the other), where one allele does not change whereas the other one converts to the same state as the unchanged allele, is also responsible for haplotypic variability.

1.2.2 Available polymorphism data.

The first draft of the human genome was released in 2001 thanks to two independent sequencing efforts [17, 18]. This draft, as well as the following ones, does not consist in the sequence of *the* human genome. Indeed, it

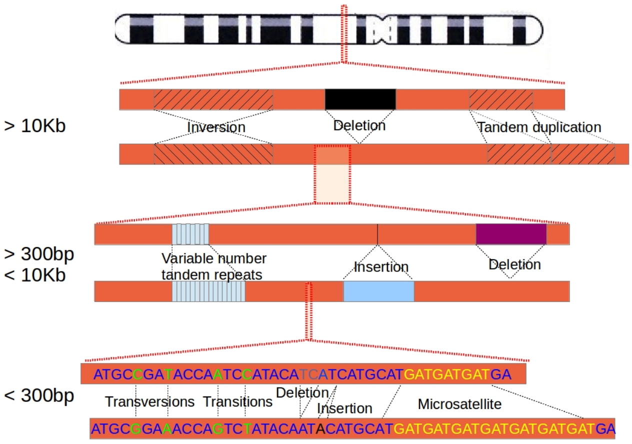


Figure 1.3: Types of genetic variation. Different genetic variations segregating in a genome and classified according to their size in base pairs.

has been retrieved from a mosaic of many different genomes from different individual sequences. During the assembly of the first human genome, around 4 Millions SNVs were discovered. Those SNVs, represents single nucleotide differences among the individuals used for the assembly.
Since then, several project provided public access to genotype data from samples in worldwide populations. The work described in the present thesis has been mostly performed on three main databases. A brief description of each as well as their own strengths and drawbacks is then required.

Human Genome Diversity Panel.

The Human Genome Diversity Project (HGDP), led by Luca Cavalli-Sforza and Allan Wilson, began in 1991. This project aimed at collecting, analyzing and making available a broad set of human samples all around the globe . In 2002, a panel (called HGDP with now the P standing for Panel instead of Project) made available 1,064 cell lines from individuals from 51 populations representing the seven main geographic areas (Sub-

Saharan Africa, Middle-East and North Africa, Europe, Central-South Asia, East Asia, Oceania and America) [19]. In 2008, Li *et al.* genotyped 1,043 of those individuals distributed across 51 populations, on the Illumina HumanHap650K Beadchips. This array include 650 thousands markers chosen to maximize *tagging* of additional common SNPs that are in LD with the genotyped SNPs. Those markers were described to tag in European, Asian and African samples more than 90%, 88% and 67% of SNPs with Minor Allele Frequency (MAF) above 5%, respectively. The main drawback of this data results from the type of markers included. Indeed, Li *et al.* (2008) [9] reported a bias towards highest heterozygosity in Europe, with heterozygosity level being lower in Middle-East, Central-South Asia and three hunter-gatherers groups in Africa and followed by East Asia. This bias, known as *ascertainment bias*, is the systemic distorsion of the allele frequency spectrum due to a *a priori* discovery of the polymorphism segregating in a reduced sample. Thus, when genotyping individuals from other populations, especially isolated by distance from the ascertainment sample, one most certainly does not catch all the genetic variation (more on *ascertainment bias* in 1.5.1).

International HapMap Project.

In 2003 started the International HapMap Project. This project aimed at developping a variant map of the human genome to describe the common patterns of genetic variation. During the Phase I and II, ~ 3.1 Million SNPs have been genotyped in 270 individuals from three different populations [20]. The samples were retrieved in a Yoruba (YRI) population in Nigeria, an European ancestry population in Utah, USA (CEU), a Han Chinese population in Beijing and Japanese from Tokyo (CHB+JPT). There were different technologies used to discover new SNPs and type them in the three samples. In these two phases the ascertainment scheme is therefore difficult to assess since it depends on the genotyping technology. For the Phase III of the project, the number of sampled populations has been increased up to 14. However, this phase does not provide as much variants as in the first two phases. Indeed, the samples were genotyped only on Affymetrix Genome-Wide Human SNP Array 6.0 with however an easier ascertainment scheme to assess. This chip also includes *tag-SNPs*, i.e. SNPs that catch the variation in the surrounding regions because they are in LD with ungenotyped variants. As for HGDP, Hapmap data mostly captures common variation since it provides the genotypes for SNPs segregating with a MAF $> 5\%$. Moreover, HapMap project released accurate genetic map (giving information on the past recombination events across the genome) and a wealth of information about the patterns of LD in human populations.

1000 Genomes Project.

Following the HapMap project, the 1000 Genomes Project provided more insight into the human variation [21]. Making profit of the emergence of NGS technologies, the 1000 Genomes project aims to provide a catalog of human genomic variation by sequencing ~ 2,500 individuals in 27 populations. Now, with the Phase I release, already over 1000 individuals have been sequenced and ~ 41 Millions SNVs have been discovered using both whole-genome sequencing at low coverage (2-6 X) and limited targeted exon sequencing at higher coverage (50-100 X). They describe 98% variants (both SNVs and indels) segregating with a MAF $> 1\%$ in a given population. They also implemented *in silico* genome-wide phasing and imputation, meaning that the haplotypes are provided and ungenotyped SNVs have been inferred. The data also describes detectable Copy Number Variants (CNVs). So far, this is the most detailed catalog of human variation available. However, one must note that the coverage used for sequencing strongly affect the power to detect rare variants, thereby, the allele frequency spectrum observed for exonic regions will be different than for the rest of the genome.

1.3 What determines genetic diversity levels within species?

One main goal in studying species evolution is to determine which are the forces producing and maintaining genetic diversity in natural populations. Such knowledge contributed to the development of the neutral theory of molecular evolution which has mostly been attributed to the work of Motoo Kimura [22] in the 50s and 60s. The *neutral theory of molecular evolution* may in turn be used as the null hypothesis for many evolutionary analyses in order to assess whether a population have evolved under natural selection accounting for some specific molecular patterns observed throughout the genome [23–25]. The *Hardy-Weinberg principle* [26, 27] which describes the conditions a sexual population has to meet to be at equilibrium gives straight forward insights into the evolutionary forces in action. The so-called *Hardy-Weinberg equilibrium* states that allele frequency will remain equal across generations if the following criteria are fulfilled (1) diploidy and individuals can only reproduce through sexual mating; (2) generations are non overlapping; (3) allele frequencies are equal in both sexes; (4) there is no mutation; (5) the population is panmixic (individuals mate randomly); (6) there is no migration from or to another population; (7) the size of the population is infinitely large, and (8) natural selection is not active.

Since we are interested in human evolution, diploidy and sexual mating are met in any cases. It is complicated to infer the influence of the absence overlapping generations, but we will consider this criteria as granted. The allele frequency equality among sex is not guaranteed, especially for sexual chromosomes and, for this reason, many genetic studies focus either on sexual or autosomal chromosomes separately. Mutation is a rare process: it has been estimated that the average single nucleotide substitution is in the order of 10^{-8} per base per generation [9, 13, 14]. However, those rare changes accumulate across generations and are the raw material for evolution to occur. Therefore, mutation can not be discarded when studying evolution but it is often assume that this mechanism does not account for observed genetic differences among genomic regions, and thereby its rate is considered to be uniform across the genome although it is not (e.g. see [28]). Panmixia condition relies on the absence of any mating restrictions among the individuals so that they can mate randomly. Specific environmental, behavioural, hereditary or social interactions may account for population structure and, thus, prevent random mating. Moreover, the absence of migration from one population to the studied one guarantees no new allele supply in the gene pool. Finally, violation of the last two conditions regarding population size and absence of natural selection are the most studied by evolutionary biologists. If a population has an infinitely large size, the random sampling of the alleles from one generation's gene pool to be passed to the next one is unbiased: the alleles present at one generation are a representative sample of the alleles at the previous one. The main contribution of the neutral theory of molecular evolution was to describe how, in finite populations, this random allele sampling from one generation to the next one leads to a significant fluctuation in allele frequencies across generations, a phenomena well-known as *genetic drift* and discussed in 1.3.2. On the other hand, if one allele is evolving under natural selection, that is its odds to segregate through generations is lower or higher than for the other alleles, its frequency will decrease or increase. We will discuss the different modes of natural selection in 1.3.1.

1.3.1 Natural selection in action.

> This preservation of favorable variations and the rejection of injurious variations, I call Natural Selection.
>
> *On the origin of species by means of Natural Selection, or the preservation of favoured races in the struggle for life*
> CHARLES R. DARWIN

As stated in the quotation above, (Darwinian) natural selection targets a heritable trait that provides greater or lower chances for an organism to reproduce, and/or to survive, in a given environment. This evolutionary process is therefore directional: while an allele responsible for any advantageous trait will be selected for and, thus, increase in frequency in the population, an allele encoding a prejudicial phenotype will be selected against and purged from the population. This concept, introduced in 1858 simultaneously by Charles R. Darwin and Alfred R. Wallace [29, 30], has been at the core of the study of evolution and biological research. However, since then there has been passionate debate concerning its relative importance among other evolutionary processes, the prevalence of adaptive traits and how they are originated in natural populations. For natural selection to be effective, Darwin suggested that a population must present three features. First, as mentioned in the title of the book presenting his theory, individuals within a population must "struggle for life", meaning that more individuals are born than the number that can actually survive in the population. Second, individuals should vary in their ability to reproduce (or survive until reproductive age), so that only the fittest ones would be more likely to have offspring, and thus, to transmit their characteristics to the next generation. A concept that the liberal theorist, Albert Spencer, interpreted as the "*survival of the fittest*". Although this expression was then used by Charles Darwin himself in latter versions of The Origin of the Species, it does not accurately describe the process of natural selection which acts on reproductive differences among individuals. Third, the variation in reproductive success must be heritable. Darwin could demonstrate the existence in natural population of the two first requirements. However, he was unfortunately unaware of Gregor Mendel's work on the law of inheritance and could not provide any suitable model of inheritance. Alternatively, he suggested a blending inheritance model in which the offspring is a fusion of its parent's characters. This model was not suitable for his theory because such blending inheritance would remove quickly any variability in the population, avoiding natural selection to act. The absence of a suitable model of inheritance prejudiced the natural selection theory and brought controversy in the field. We had to wait until the 20th century, for Mendel's work to be broadly known.

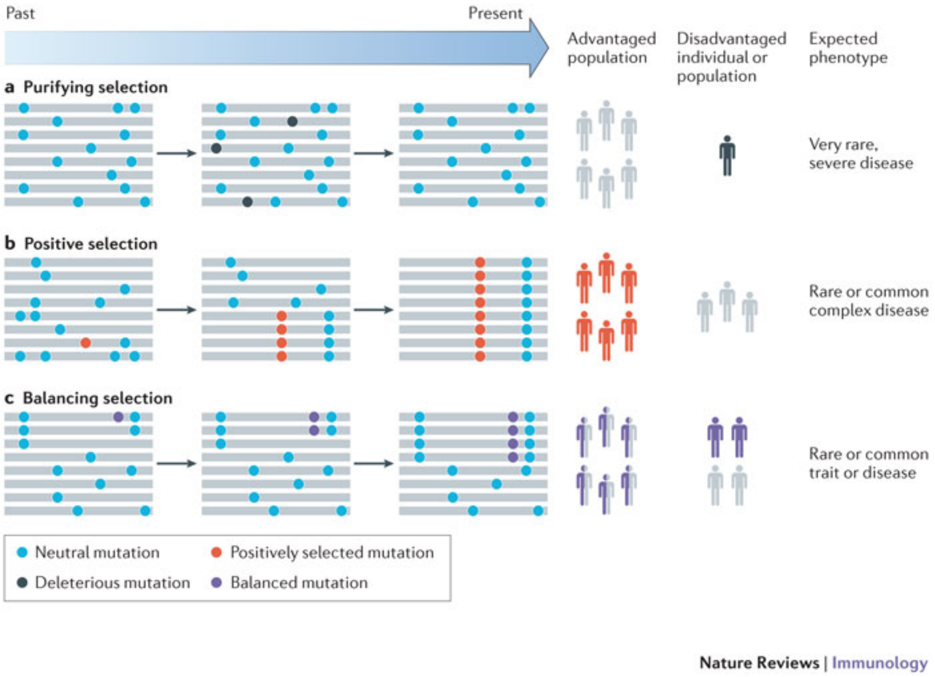

Past
Present
Advantaged population
Disadvantaged individual or population
Expected phenotype
a Purifying selection
Very rare, severe disease
b Positive selection
Rare or common complex disease
c Balancing selection
Rare or common trait or disease
Neutral mutation
Positively selected mutation
Deleterious mutation
Balanced mutation
Nature Reviews | Immunology

Figure 1.4: The different modes of natural selection. From [31].

Along with the *Hardy-Weinberg principle*, Mendel's law of inheritance could finally provided the missing link in Darwin's theory: variability can be maintained in the population when random mating occurs in a sufficiently large population. On the other hand, when mating is not random anymore and the fitness of the individuals is not the same, natural selection occurs leading to a reduction in variability, just as predicted by Darwin.

Consequently, at the dawn of the 20th century, natural selection began to be fully accepted by biologists and several theorists gave birth to the *modern synthesis of evolutionary theory*, also referred as *neo-Darwinism*. Notably, John D. S. Haldane, Ronald A. Fisher and Sewall Wright began impressive theoretical work considered as the founding principles of population genetics. In their extensive work, they demonstrated how natural selection on a phenotype induced by a single or multiple loci could result in rapid changes in loci frequency within a population, leading in turn to the phenotype evolution [32–35]. In this setting background paradigm, phenotype is placed as the target of natural selection. This theoretical background led to the selectionist view in which natural selection is considered by far the main mechanism of evolutive changes in a population. Natural selection is a generic name which accounts for four different evolutionary processes: sexual, purifying, positive and balancing selection. Sexual selection is a mode in which random mating does not occur due to the choice of the reproductive partners based on some specific phenotypes. The three other modes are related to the odds of an individual to survive until reproduction and to reproduce. Therefore, sexual selection is not to be considered together with the three others: the reason why an allele encoding the phenotype targeted by the different modes of selection is selected for or against differs. The following will focuses on the description of natural selection where environment represents the main selective force, and, sexual selection will not be considered.

1. **Purifying selection** (Figure 1.4a.), also referred as *negative selection* or *stabilizing selection*, is the evolutionary process by which deleterious mutations are removed from the population's genetic pool. It ensures that organisms remain well-fitted to their environment and prevents from the spread of any damaging mutations across generations. It is considered as the most effective type of selection because mutations with a functional consequence seem to be more likely to decrease than to increase the fitness. Thus, purifying selection is believed to be widespread in functionally important genes or regulatory elements.
2. **Positive selection** (Figure 1.4b.), also referred as *Darwinian selection* or *adaptive selection* has been considered one of the most important driving forces for phenotypic variability among species or populations. The concept is straightforward: any beneficial mutation will increase in frequency in the population, thus allowing the adaptation of individuals to new environments.
3. **Balancing selection** (Figure 1.4c.) sustains the segregation of dif-

ferent alleles in a population. In opposition to positive and negative selection, this mode of selection avoids alleles to reach fixation, and thereby, favors genetic diversity. Alleles under balancing selection cannot be strictly classified as deleterious or beneficial to the environment since it would depend on other factors. Indeed, four main processes can lead to an excess of polymorphisms. First the overdominance, i.e. when the heterozygote genotype is the fittest. Second, frequency-dependent selection, i.e. when an allele becomes deleterious or beneficial depending on its frequency in the population. Third, fluctuating selection, i.e. when selection coefficients vary over time and/or space. Fourth, pleiotropy when the selective variant affects multiple traits with different effects.

1.3.2 The neutral theory: the role of genetic drift in evolution.

> This neutral theory claims that the overwhelming majority of evolutionary changes at the molecular level are not caused by selection acting on advantageous mutants, but by random fixation of selectively neutral or very nearly neutral mutants through the cumulative effect of sampling drift (due to finite population number) under continued input of new mutations.
>
> *The neutral theory of molecular evolution: A review of recent evidence*
> MOTOO KIMURA

Since its introduction at the beginning of the 20th century, the *modern synthesis of evolutionary theory* and its selectionist view was each time more popular until the 1970s. However, it was based on pioneer work

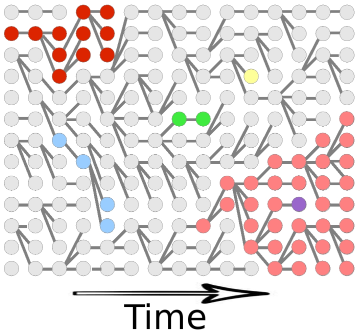


Figure 1.5: *Genetic drift* in a finite population. Different coloured circles represent different alleles. Lines represent the transmission of an allele from one generation to next one. *Genetic drift* drives increases and decreases in allele frequencies by random sampling of the parental alleles to the offsprings.

from Wright, Fisher and Haldane who also considered an other source of variability within and among populations: a stochastic selectively neutral process in opposition to the deterministic evolution through natural selection. One of the strongest assumption in *Hardy Weinberg principle* is certainly the infinite population size, particularly for humans. Indeed, as briefly described in 1.1, modern human expansion has likely occurred through several founding events and the human populations suffered several bottleneck. The *effective population size*, *Ne* [36], a measure of the constant size for an idealized population that represents the past history of the population of interest, it is relatively small in humans. Indeed, it has been estimated that *Ne* is ~ 10,000, a quite striking number if compared to the more than 7 billions individuals peopling the world nowadays. Introduced by Sewall Wright in 1931, the *effective population size* represents the harmonic mean of population sizes among generations, and thus, allows to describe the amount of *genetic drift* a population has suffered. As stressed before, *genetic drift*, is the allele frequency fluctuation across generations due to random sampling of the gametes from one generation to engender the next one (Figure 1.5). The Wright-Fisher model [33, 36] was the first proposed to explain the diffusion of allele within a

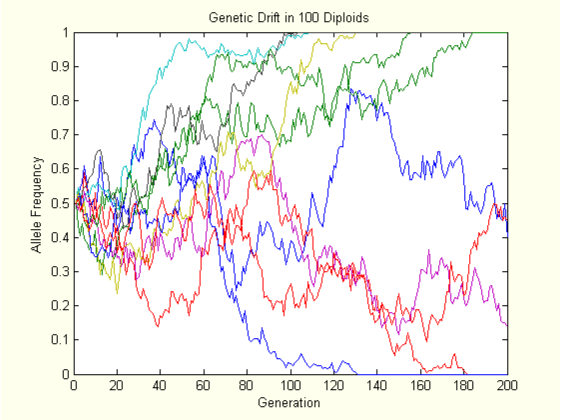


Figure 1.6: Simulated allele frequency trajectories under *genetic drift*. The fate of the change is random.From `http://pandasthumb.org`

population due to *genetic drift*. In this famous discrete time model, they proposed a simple equation to describe the probability to observe an allele at frequency p_n at generation n when segregating at frequency p_{n-1} at the previous generation. This model is rather simplistic since it relies on assumptions found in the *Hardy-Weinberg principle* (diploidy, absence of selection and mutation, no overlapping generations, panmixia) and the absence of recombination among variants. However, it remains extensively used to estimate the rate of evolution of a population through *genetic drift*. The equation for diallelic variants they proposed is the following:

$$\begin{aligned} P(p_n = \frac{k}{2N} \mid p_{n-1}) &= \binom{2N}{k} p_{n-1}^k (1-p_{n-1})^{2N-k} \\ \Leftrightarrow & \frac{(2N)!}{k!(2N-k)!} p_{n-1}^k (1-p_{n-1})^{2N-k}, \end{aligned} \tag{1.1}$$

where, N is the actual size of the population and k the number of copies of the allele observed at generation n. When comparing the observed p to the one expected under this model across many variants, one straightforward outcome is the amount of *genetic drift* due to the differences between N and Ne.

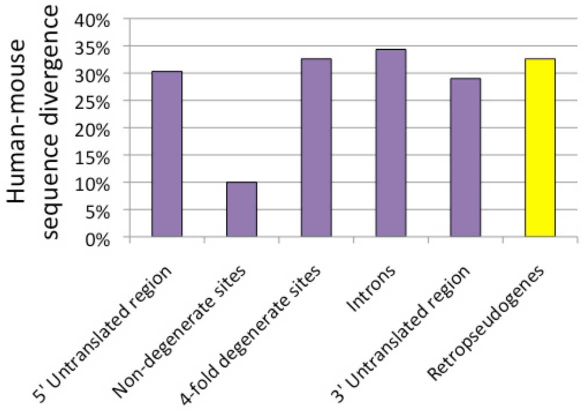


Figure 1.7: Sequence divergence among different site classes. Less changes are observed at non-synonymous sites than at site with no or little effect on protein function. From [37].

However, until the 1960s. The predominant view in evolutionary biology was that natural selection is playing the dominant role in explaining the observed gene pool. According to this view, the differences between species were assumed to occur mostly from advantageous mutations that had been fixed by positive selection (Section 1.3.1). On the other hand, the observed relatively important amount of polymorphism within populations, was explained by the action of balancing selection (Section 1.3.1) or by a transient change towards fixation of the advantageous alleles. Hence, neutral (non adaptive) processes were overlooked because of their assumed little contribution. Based on the recent view of large amounts of polymorphism, Kimura challenged the *selectionist theory* in the late 1960s. Indeed, Kimura observed that genetic variability was more frequent than expected and, proposed his now recognized *neutral theory of molecular evolution* [23]. The main claim in his theory was that most genetic variants observed in a population are neutral, i.e. has no phenotypic effect for the individuals carrying them. Thus, the observed amount of diversity is mainly the results of the interplay of *genetic drift* and mutation (Figure 1.6).

One strength of Kimura's theory was not to be exclusively neutralist. Indeed, Kimura also acknowledged that most new mutations are deleterious if occurring in functionally important regions and, thereby, are quickly removed from the population by purifying selection (Section 1.3.1). Therefore, such mutations do not contribute, or contribute little, to the divergence among different species sequence and to polymorphisms within species. Moreover, in this theory, the role positive selection is not rejected, but rather Kimura stated that it was rare. He also anticipated the controversy he would generate in the field and provided strong predictions to be tested with the actual data. The most famous and used one would be that, under the neutral theory of molecular evolution, more changes during divergence between species sequences are expected in functionally less important regions. When Kimura proposed his neutral theory in 1968, only a few protein sequences were available. We had to wait until the wealth larger amount of DNA sequence data in the 1980s to validate such prediction. In 1991, Kimura then published a review [38] in which he reported observations supporting his theory: (1) amino acid substitutions with similar biochemical properties are more often observed than radical changes, because of their lower effect on protein function; (2) there are more synonymous substitutions (causing no change in amino acids) than non-synonymous ones; (3) the evolutive rate at non coding regions (e.g. introns) is higher than for coding ones; (4) non-coding sequences, such as introns, evolve at a high rate similar to that of synonymous sites, and (5) pseudogenes evolve at high rate somewhat similar to the rate observed for third-codon positions, also known as codon wobble positions (Figure 1.7). While these observations are consistent with the neutral theory, they contradict the selectionist one: if most substitutions were adaptive, we would observe more substitutions in regions with important function than in regions where changes have little or no effect on phenotype, such as pseudogenes, non-coding sequences and synonymous sites.

All together, Kimura's *theory of neutral molecular theory* "only" reconsider the dominance of the evolutionary forces in action: the *effective* *population size* does matter. In order to conclude the action of selection at a given locus, one must reject the null hypothesis which states that the molecular patterns observed around this locus are the result of *genetic drift*. A new theoretical background to detect the action of alternative evolutionary processes disrupting neutral evolution was born! [23–25]

1.4 Statistical approaches to identifying signals of positive selection.

As introduced in Section 1.3.1, positive selection at the genomic level is the process through which an allele that determine an advantageous trait will increase rapidly in frequency, potentially until it reaches fixation. The allele frequency trajectory in the population through the action of positive selection depends on two main factors: the strength of the selective pressure and the number of generations since it started. The strength of positive selection is measured by the selection coefficient, s, defined as the increased percentage of offspring that the individual carrying the advantageous genotype at each generation, compared to individuals with alternative genotypes. A higher selection coefficient involves the advantageous allele to increase quicker in frequency, and thereby, to reach fixation in a shorter time. On the other hand, the speed of the increase tends to decline with the frequency of the advantageous allele in the population: the selection coefficient relates the relative advantage of individuals carrying the advantageous genotype compared to all the others. As a consequence, the allele frequency trajectory is non-linear but rather depends on the number of generations since the allele began to increase in frequency through the action of positive selection.

The allele frequency shift comes with some typical molecular footprints used to detect selective events in the genome. Usually, we distinguish between two method families according to the kind of data analysed.

1. **Using divergence data**, i.e. sequences from different species, one can identify substitutions in the genome that are different across the species due to a past selective events that contributed to the species divergences.
2. **Using polymorphism data**, i.e. sequence or genotype data from different populations within a same species, one can explore the nucleotidic and haplotypic diversity within and among populations.

The different molecular patterns left by a selective event are not maintained forever in the genome, and those footprints allow to infer how many generations have past since the selective events occurred. (Figure 1.8).

1.4.1 Using divergence data.

Estimating molecular evolutionary rates.
Using divergence data, the calculation of the ratio d_N/d_S (referred to as ω) is the most commonly used method for determining the nature of the selective forces acting on a protein-coding gene. For that purpose using the orthologous sequences for several species, d_N is the rate of nucleotide substitutions that have occurred per nonsynonymous site in the sequence while d_S is the rate of substitutions per synonymous site. Based on the assumption that synonymous substitutions are largely neutral, in opposition to non-synonymous ones (see Section 1.3.2), the ratio of substitution rates between these two site classes is then taken as an indicator of the strength of non-neutral selective forces acting on the gene during the species evolution since their divergence. When d_N/d_S = 1, the gene is said to have evolved neutrally, while $d_N/d_S < 1$ is seen as the footprint of purifying selection constraining the gene evolution because non-synonymous substitutions have been removed from the population at a greater rate than synonymous substitutions. On the other hand, d_N/d_S > 1 would imply that positive selection has occurred, with more fixed non-synonymous substitutions than expected under neutral evolution (as inferred by the rate of synonymous substitution).

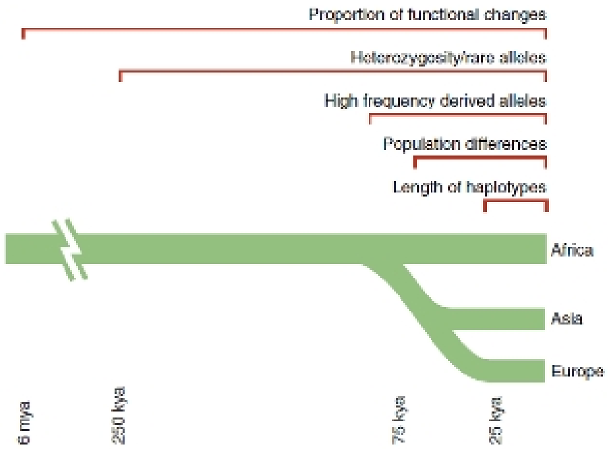

Proportion of functional changes
Heterozygosity/rare alleles
High frequency derived alleles
Population differences
Length of haplotypes
Africa
Asia
Europe
6 mya
250 kya
75 kya
25 kya

Figure 1.8: Time scales for the signature of selection. The signatures of selection persist over different time-scales. From [39].

Different methods have been implemented to estimate the d_N/d_S. The early ones, such as the one by Nei and Gojobori (1986) [40], rely on a basic counting. Although rather simple, they illustrate well the spirit of the methods. At each codon, the numbers of synonymous and nonsynonymous sites are calculated. For each position, i, in a codon, the fraction f_i of observed synonymous changes is computed; the numbers of synonymous (s) and nonsynonymous sites (n) are then obtained by $s = \sum_{i=1}^{3} f_i$ and $n = 3 - s$. The total number of synonymous (S) and nonsynonymous (N) sites in a sequence are the sum of s and n, respectively, across the whole sequence, i.e. the whole set of codons. The number of synonymous (s_d) and nonsynonymous changes (n_d) per codon between two sequences are then counted. At this step, multiple mutational paths between two codons are considered equally likely, thus, the resulting counts is computed as the average between all possible paths. The sum of these counts give the total number of synonymous (S_d) and nonsynonymous (N_d) changes across the sequence. $p_s = S_d/S$ and $p_n = N_d/N$ are used as an estimation of the proportions of synonymous (p_S) and nonsynonymous (p_N) differences,respectively. Finally, the rate of synonymous substitutions (d_S) and nonsynonymous substitutions (d_N) can be computed following the formula 1.2 [41]:

$$d = -\frac{3}{4} log_e \left(1 - \frac{4}{3}p\right) \tag{1.2}$$

where p is either p_S or p_N. More recent methods also include more parameters to account for differences among different types of substitutions. For example, transversion changes do not occur as frequently as transition (transition/tranversion rate bias) and there are different chemical property differences among amino acids. The number of sequences analyzed has increased a lot since 1986, and therefore several statistical approaches have been suggested to estimate more efficiently d_N and d_S, such as Markov-process model [42] and Bayesian approach [43].

Tests of adaptive selection.

A gene can be overall constrained in its evolution while some specific codons being evolving under positive selection. As a consequence, such approach requires a very strong trend of positive selection to produce a value greater than one. Therefore, several tests for positive selection at specific codon sites, instead of working at the gene-level, using divergence data have been suggested. They can be divided into two main families of tests: the individual site (IS) and the pooled site (PS) tests. The first IS method was proposed in 1999 [44] and relies on the construction of a phylogenetic tree in order to count the total number of synonymous and nonsynonymous substitutions across all branches of the tree. A signal of positive selection at a given codon would then be a significantly greater number of nonsynonymous than the number of synonymous substitutions. Improvements came from both likelihood and Bayesian-based implementation of this framework [45–47]. On the other hand, the PS methods rely on nested models in order to perform likelihood-ratio tests between the

null model where all sequence alignments are assumed to present either a $d_N/d_S < 1$ or $d_N/d_S = 1$ (i.e. sites are either evolving neutrally or under purifying selection, respectively) and the alternative model where another site class is considered (namely some sites are fit by the model to have evolve under positive selection, that is with a $d_N/d_S > 1$) [48, 49]. For example, one of the most used test compares models M7a and M8. The M7a model estimates the likelihood of the sequence alignment to fit seven site classes, one with $d_N/d_S = 1$ and six with $d_N/d_S < 1$, with a β distribution to model the d_N/d_S values for the sites. Next, the same data is fitted to the M8 model which considers one more site class allowing $d_N/d_S > 1$. Finally, a likelihood ratio between the best likelihoods under each model allows to test whether adding putative positive selection to the model explain better the data [48, 49]. Further methods using a Bayesian framework can also be used to test *a posteriori* which are the specific codons that have been targeted by positive selection [48–50].
As mentionned before, the main advantage of these codon-specific tests over the simple calculation of d_N/d_S at gene level, is that they take into account that individual sites to evolve at different rates. Nevertheless, these methods remain generally little sensitive to detect positive selection [51]. A particular concern is that they require sustained strong events of positive selection while those may be rare during evolution and the common case seems to be brief episodes of selection. Consequently, in order to detect such episodic adaptive events, several methods have been developed in which evolutionary rates can vary not only between codons but also between phylogenetic branches [52–54]. These methods are more sensitive to detect positive selection [54, 55].

1.4.2 Using polymorphism data.

In 1974, Maynard Smith and Haigh [56] proposed a model to explain the molecular mechanisms at play when positive selection acts on a variant. In this model, now referred as the *hard sweep model*, they described the phenomena of *genetic hitchhiking* which results from positive selection

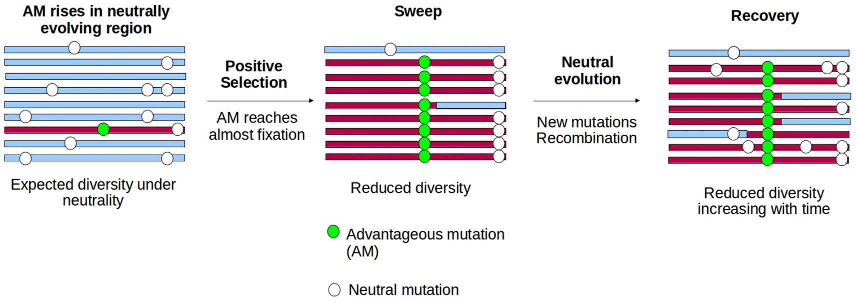

AM rises in neutrally evolving region
Expected diversity under neutrality
Positive Selection
AM reaches almost fixation
Sweep
Reduced diversity
Neutral evolution
New mutations Recombination
Recovery
Reduced diversity increasing with time
Advantageous mutation (AM)
Neutral mutation

Figure 1.9: Molecular patterns in genomic region suffering selective sweep. Before the selective sweep in a neutrally evolving region, an adaptive mutation (green circle) arises on one chromosome. During the selective sweep the frequency of the adaptive allele and its linked variants rapidly increase in frequency. After the sweep, adaptive and linked alleles are fixed, variability is lost. During the recovery phase, new mutations begin to appear in different chromosomic backgrounds by recombination and mutation.

driving a quick increase in frequency of an initially rare and beneficial allele up to (or close to) fixation. This *selective sweep* occurs so quickly that recombination is not efficient to cut the haplotype where the selected variant arose, and thus, all the variants carried by this haplotype also increase in frequency (Figure 1.9). Therefore, under the *hard sweep model*, ones expect a decreased in genetic diversity in the surrounding genomic region. The size of the region affected by such a sweep is proportional to the ratio of the strength of selection to the rate of recombination [56–58]. Thus, the reduction in levels of diversity within the genome is determined by the distribution of selection coefficients and the number of selective events in unlinked genomic regions. A selective sweep drives a quicker shift in allele frequency than what is expected under *genetic drift*, but recombination may occur, and thus, neutral alleles further away from the selected site may not be driven all of the way to fixation, resulting in a temporary excess of high-frequency derived alleles at intermediate distance away from the selected site after the selective event [59–61]. Once the sweep is over, the genomic region enters a recovery phase during which it regains neutral diversity levels through new mutations. Therefore, sweep leaves a strong skew towards low frequency alleles, a pattern that persists for many generations [60–62]. The rate of sweeps can be important enough that *hitchhiking* dominates *genetic drift*, especially in large populations, as the source of stochasticity for neutral alleles [56, 57, 63]: the *genetic draft* [63].

Maynard Smith and Haigh gave the theoretical background to most of the tests implemented so far to detect signatures of selection at molecular level using polymorphsim data. Many of those tests have been ran on 1000 Genomes data, the latest publicly available polymorphism data in worldwide populations (see Section 1.2.2 for a description of the data) by Pybus et al. (2014) [64] (see Chapter IV) and rely on three main features expected to be present in a genomic region surrounding a selected allele: an important linkage disequilibrium, a skewed Site Frequency Spectrum (SFS) and an excessive genetic differentiation among populations. The list of methods classified by methods is given in Table 1.1.

Tests based on long haplotypes.

Positive selection creates high levels of LD in the region surrounding the selected variant: for a similar shift in allele frequency, less recombination events takes place when there is *selective sweep* than pure *genetic drift* since the shift in allele frequency is much quicker in the former case. The Long Range Haplotype (LRH) test is commonly used to detect such signal [65]. However, this test does not take into account the recombination rate heterogeneity across the genome. To overpass this limitation, other tests have been implemented and are based on the the Extended Haplotype Homozygosity (EHH) decay which measures the decay of the haplotype homozygosity observed when moving away from the selected variant because hitchhiking of neutral allele is weaker (see Figure 1.10 for a schematic representation of EHH decay calculation). First, the Cross-Population Extended Haplotype Homozygosity ($XPEHH$) compares the EHH decay observed in a population of interest to a reference one [39]. Second the integrated Haplotype Score (iHS) compares within

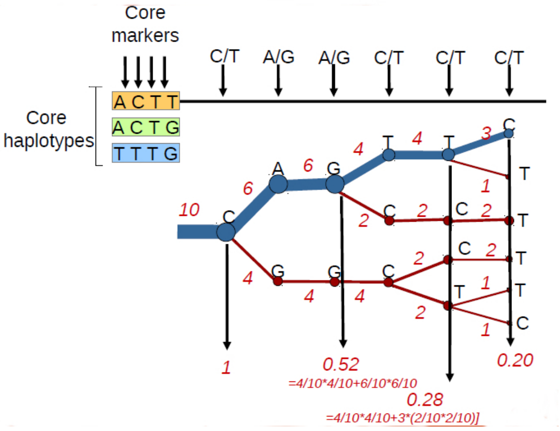


Figure 1.10: Extended Haplotype Homozygosity decay. Moving away from the variant of interest, the haplotypes bifucarte and the haplotype carrying the core markers are less and less frequent. Thickness of the blue lines represent the frequency of the haplotype (haplotype counts in red). The haplotype homozygosities when considering different number of variants are given at the bottom.

the same population the EHH decay observed for the derived and ancestral alleles. Those comparisons correct automatically for recombination rate heterogeneity across the genome. Only recent selective sweeps (˜ < 30 KYA) can be characterized by the presence of long haplotype blocks: elder sweeps are however not identifiable with this footprint since recombination would have time to shuffle the haplotype blocks.

Tests based on Site Frequency Spectrum.
The Site Frequency Spectrum (SFS) is the representation of the number of alleles observed in a sample belonging to different frequency classes for a given set of polymorphic sites. As mentioned before, genetic hitchhiking will drive neutral alleles located on the haplotype carrying the selected allele to high frequency, leading to a reduced diversity, an excess of rare alleles, an excess of derived alleles at high frequency and a scarcity of alleles at intermediate frequency compared to what is expected in a neutrally evolving region (Figure 1.11). The excess of rare alleles can

be formally tested by the famous Tajima’s *D* [66] and persists for a long time (up to ~ 250 KYA) during the recovery phase. Moreover, if ancestral state of the variants is available, one can also test for the expected excess of high frequency derived allele (Figure 1.11), with tests such as Fay and Wu’s *H* [59]. This specific excess vanishes more rapidly as recombination relieves neutral variants to evolve under pure *genetic drift*. This patterns can be detected it up to ~ 80 KYA after the sweep occured.

Tests based on genetic differentiation.
When a population faces a new environment, positive selection may act on mutations that help the individual to better adapt to this new environment. Therefore, to detect the alleles responsible of local adaptation, i.e. that has not occured in all the populations, one approach is to study genetic differentiation among populations. Traditionally the most used statistics is the fixation index, F_{ST}, first introduced by Wright and declined into different versions. Using Cockerham and Weir’s formula, F_{ST} can be viewed as the the proportion of genetic diversity due to allele frequency differences among populations:

$$F_{ST} = \frac{\sigma_a^2}{\sigma_w^2 + \sigma_b^2 + \sigma_a^2} \tag{1.3}$$

where σ_w^2, σ_a^2 and σ_b^2 are the intra-individual, inter-population and within population inter-individual variances, respectively.
F_{ST} ranges from 0 to 1, with 0 when there is no differentiation (complete panmixia) and 1 indicating complete differentiation of the populations. Although, high F_{ST} can putatively be attributed to the action of positive selection in one population, this approach is often criticised because of its sensitivity to population structure, demographic history, ascertainment bias and minor allele frequency (for a review see [82]). The ΔDAF score (the differences of derived allele frequency between one population and a reference) is another genetic differentiation index which suffers the same limitation. However, the use of the derived allele state allows to identify the population where positive selection occurred. Further meth-

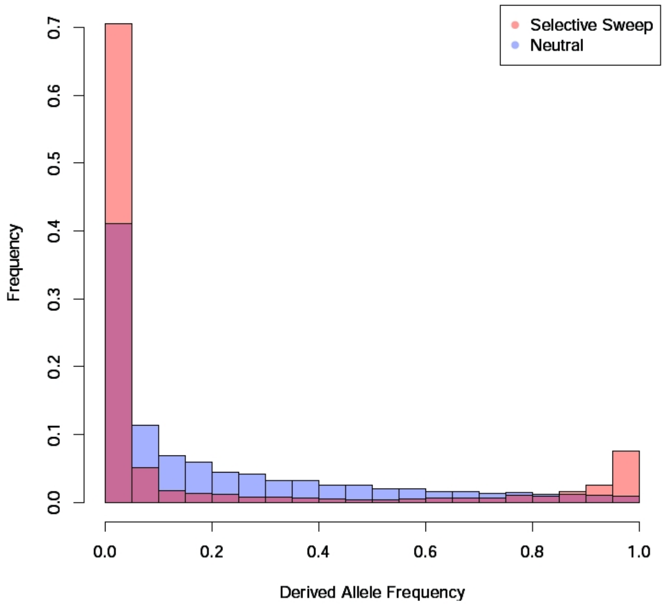


Figure 1.11: Site Frequency Spectrum (SFS) under different evolutionary models. The Unfolded SFS represents the number of derived alleles observed within different frequency classes. A region that have evolved under positive selection presents an excess of rare variant and of derived alleles at high frequency is expected (sweep in red). During the recovery phase, the former pattern will remain due to new mutations arising in the region while the later is lost more rapidly. Based on coalescent simulations of 100Kb regions (from Pybus *et al. 2014*) with no selection (3000 neutral replicates in blue) and with a recent selective sweep in an European population driving an advantageous up to fixation (300 selective sweep replicates in red) using COSI with demographic model calibrated by Schaffner *et al.* (2005) [10].

ods using genetic differentiation pattern have been developed. For example, the Cross-Population Composite Likelihood Ratio test ($XPCLR$) implemented by Chen *et al.* (2010) [79] relies on a null model of *genetic drift* and an alternative one with *selective sweep* and makes profit of the genomic context around the selected allele to detect genomic regions with long chunks being differentiated among populations due to hitchhiking, making this method more robust to demography than individual SNP based methods, such as F_{ST} and ΔDAF.

1.5 Practical challenges in detecting positive selection using polymorphism data.

The work presented in this thesis mostly focuses on the impact of positive selection in the human genome at the intraspecific level, and, as a consequence, the analyses have mostly been performed using polymorphism data. Therefore, it seems interesting to describe some challenges potentially met when detecting positive selection using polymorphism data and approaches to overcome them. First, looking for the different genomic footprints left by positive selection can result difficult in the genotype data used (1.5.1). Second, those footprints may also result from other mechanisms (1.5.2 and 1.5.3) and are specific to the *hard sweep model* while other modes of adaptation leave much subtle signals at molecular level (1.5.5).

Table 1.1: Statistics implemented by Pybus *et al.* (2014) [64] and available in as UCSC tracks in the 1000 Genomes Selection Browser 1.0 at `http://hsb.upf.edu/`.

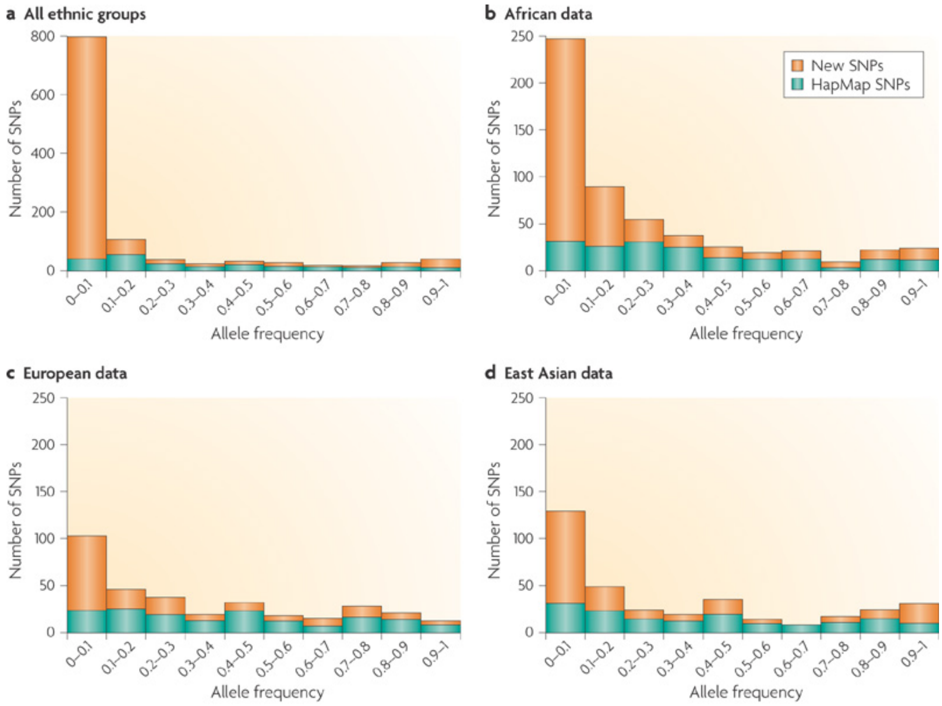
Method family	**Method**	**Reference**
Site Frequency Spectrum	$Tajima's\ D$	[66]
	CLR	[67]
	$Fay\ and\ Wu's\ H$	[59]
	$Fu\ and\ Li's\ D^*$	[68]
	$Fu\ and\ Li's\ F^*$	[68]
	R^2	[69]
Long haplotypes	$XPEHH$	modified from [70]
	ΔiHH	modified from [71]
	iHS	modified from [71]
	$EHH\_average$	modified from [65]
	EHH_max	modified from [65]
	$Wall's\ B$	[72]
	$Wall's\ Q$	[73]
	$Fu's\ F$	[74]
	DH	[75]
	Za	[76]
	ZnS	[77]
	ZZ	[76]
Population Differentiation	F_{ST}	[78]
	$XPCLR$	[79]
	ΔDAF	[80]
Descriptive statistics	Segregating Sites	
	Singletons	
	π (Nucleotide diversity)	[81]
	DAF (Derived Allele Frequency)	
	MAF (Minor Allele Frequency)	

1.5.1 Distortions due to ascertainment bias.

As mentioned in Section 1.2.2, most genotype data used to study population diversity present relatively important ascertainment bias. The ascertainment bias is an intrinsic feature of genotyping technologies which are still extensively used because they are simpler, cheaper and much faster than sequencing approaches. The ascertainment bias results from the *a priori* selection of the SNPs to be genotyped. Therefore, the genotype information in the population of interest will not be produced for all the segregating sites but only for the ones present in the so-called discovery sample which generally has reduced size. Hence, the ascertained SNPs are likely to segregate in the population at intermediate or high frequencies (Figure 1.12) since the probability to identify a SNP is a function of its frequency, and as a consequence, common SNPs are easier to detect than rare ones.

Furthermore, ascertainment bias can also be caused by a selection of SNP in a discovery sample that does not represent the studied population in term of genetic variability. For example, many arrays used SNPs discovered in European samples but then are used to genotype populations worldwide in which SNPs on the array are not polymorphic. Moreover, populations do not share all the variation: there are private SNPs segregating, i.e. SNPs not shared with other populations [85]. Usually, the SNPs for new designed arrays are selected from public database such as HapMap (www.hapmap.org) which in turn present an ascertainment bias of their own.

Usually, SNPs are selected to be genotyped in a population of interest with some of the following criteria: (1) having a MAF above a given threshold, usually relatively high in discovery samples representing either one or several populations of interest; (2) select one SNP every a given number of base pairs; (3) select many SNPs in targeted regions. The criteria used affect the ascertainment bias, and it is difficult to asses *a posteriori* its extent when using genotyping array designed by others. However, recent efforts have been made to design genotyping array with reduced ascertainment bias. For example, Illumina made available the Omni Family of



Nature Reviews | Genetics

Figure 1.12: Ascertainment Bias in HapMap populations. New SNPs were discovered by Wall *et al.* [84] through sequencing of 40 intergenic regions in 90 individuals from 6 different ethnic groups. The figure shows the number of SNPs in the HapMap data (green) compared with the number of SNPs that were discovered by resequencing and that were not present in the HapMap data (orange), categorized by derived allele frequency. It can be seen that the HapMap data have greater SNP ascertainment bias for African than for European or Asian populations. In particular, African populations have many low-frequency alleles that are not well represented. From [83].

Microarrays which include up to five million markers per sample and extensive coverage of new variants identified by the 1000 Genomes Project, i.e. SNPs discovered through NGS in samples from worldwide populations. Moreover Patterson *et al.* designed the *Affymetrix Human Origins array* with clearly documented ascertainment specifically for population genetics study [86].

The ascertainment bias has a direct effect on many statistics to detect positive selection using polymorphism data [87]. First, and the most straightforward, SFS-based statistics are distorted by the artefactual excess of common variants in genotyping arrays. Second, the tests based on genetic differentiation, such as the F_{ST} index, rely on a measure of genetic variance within and among the populations. Hence, if the SNPs genotyped within different populations present different ascertainment bias, the distribution of the index of genetic differentiation will be distorted. In the first decade of this century, much effort have been made to implement other methods less sensitive to the ascertainment bias with the development of haplotype-based statistics [39, 71]. However, such methods rely on an accurate estimate of LD patterns within a genomic region for the population of interest in order to infer whether there is a particularly EHH [88]. If the genotyping array only contains common variants and particularly chosen to *tag* the variability from another population, the observed LD patterns in the studied population are unlikely to be representative of the real ones. For example, for HGDP data (Section 1.2.2), in African populations for which the genotyped SNPs tag only 67% of SNPs with Minor Allele Frequency (MAF) above 5%, the power to detect positive selection is lower than for European where 90% of such SNPs are tagged. On the other hand, it has been proved that haplotype diversity is more representative than individual SNP heterozygosity in the HGDP data [16], suggesting that the ascertainment schemes affect more individual variants than haplotypes.

Nowadays, more and more studies obtain genotype information through NGS which does not suffer any ascertainment bias. However, the SFS is highly dependent on the coverage (the read depth) used for sequencing. Indeed, the power to detect rare variants increase with coverage [21].

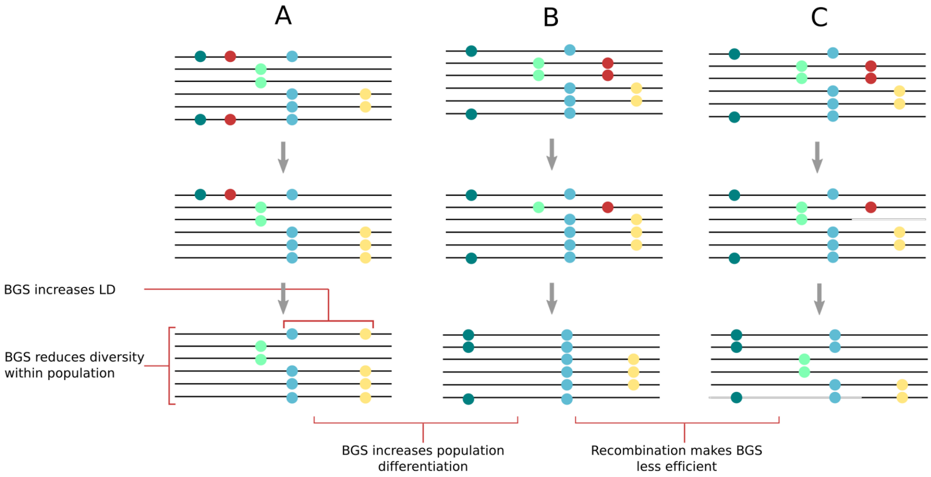

A
B
C
BGS increases LD
BGS reduces diversity within population
BGS increases population differentiation
Recombination makes BGS less efficient

Figure 1.13: Background Selection (BGS) and molecular diversity. Deleterious mutations (shown in red) are eliminated from the population together with neutral mutations on the same haplotype through BGS. This mechanism leads to reduced diversity. Initial neutral diversity is identical in all cases (A–C). Comparison of cases (A) and (B) shows that different BGS episodes will contribute to populations' genetic differentiation. Comparison of cases (B) and (C) shows that recombination reduces the effect of BGS, maintaining diversity, and reducing linkage disequilibrium (LD) as well as population differentiation (compare final states in [A] and [C]). From [89].

Moreover, the genotype information may also depend on the sequencing centre and technology and the SNP calling algorithm used. Therefore, for population genetics study, one may be cautious when merging data from different datasets and on the coverage across the genome.

1.5.2 The confounding factor of background selection.

Background selection (BGS) is a process by which neutral variation are removed from the population if they are linked to deleterious ones [90]. Therefore, BGS reduces levels of polymorphism in regions with many functional elements and low recombinations (Figure 1.13). The lower level of polymorphisms in extended region is often attributed to the action of positive selection because it is a molecular pattern expected under the *hard sweep model*. It is consequently important to correct for BGS, and one straightforward approach when analyzing protein-coding regions is to look for lower levels of neutral variation near functional substitutions, i.e. at functional sites where a mutation was fixed in a set of species, which is the clearest genomic evidence for positive selection while not being expected under BGS. However, this approach is biased towards protein-coding regions and would just detect events of positive selection acting on mutations with *a priori* known function. An alternative to this approach, would be to correct for several genomic variables that correlate with BGS, such as levels of recombination rate and functional constraint. Measuring functional constraint is not straightforward but one can use the density of coding sequences (CDS), density of conserved coding sequences (CCDS) and conserved non-coding sequences, as well as the density of untranslated regions (UTRs). Moreover, Enard *et al.* recently found that GC content have a strong influence on levels of neutral diversity [91]. Although BGS has been seen as mimicking positive selection a molecular level since the article by Charlesworth *et al.* in 1993 [90], it has been proved that tests based on LD (namely $XPEHH$ [39] and iHS [71]) are insensitive to BGS [91, 92], and therefore, their extreme deviations may directly be attributed to recent *hard sweeps*.

1.5.3 Demography can mimic positive selection.

As described in Section 1.3, many mechanisms can affect the genetic diversity in population or species, among which several demographic processes can lead to molecular patterns expected under a positive selection scenario (Table 1.2).

Migration and structure.

The neutral model assumes that any cross-gender individual pair has the same probability to reproduce in the population. However, there may be population subdivision due to geographic distance, social, linguistic or economical barriers (e.g. in India with the caste system). Those barriers to random mating are likely not to be absolute, and a number of migrants can move between subpopulations at each generation. When there is a population subdivision one is not aware of, and thus panmixia is improperly assumed, the genetic variability is higher than expected with an excess of variants at intermediate frequency. On the other hand, if there are migration from an external population to the studied population, an higher variability with an excess of rare variants is expected.

Population expansion.

During population expansion, a new generation has a greater number of individuals than the previous one. Well-described population expansion events occurred after the Neolithic transition. One possible explanation is that the agriculturist way of life may have provided a more reliable mode of subsistence and allowed settlements to increase in size. A population with expansion will show an excess of singletons and mutations at low frequency compared to a population with constant size due recent mutations which have not increased in frequency through *genetic drift* and remain almost individual specific [93]. This also implies a lower genetic variability than expected for the population size.

Population bottleneck.

A bottleneck is the phenomena through which population size decreases all the sudden, followed by a recovery or even an increase of the original population size in a few generations. One striking example is the Black Death plague faced by Asian and European populations in the 14th century. Plague is thought to be responsible responsible several large epidemics with death rates of up to 30–50% of the European population and lingering thereafter in Europe for several centuries [94]. Many alleles from the original population, mostly at low frequency, will disappear during the decreasing size phase, thus reducing the genetic variability. During the recovery phase, as for population expansion, an excess of rare variant will arise.

Founder effect.
A founder effect when a small subpopulation leaves its former habitat to establish a new one. This can be seen as a particular case of bottleneck. It is likely how modern humans colonized geographic areas out-of-Africa [95]. One more recent example would be colonization of Quebec, Canada around 400 years ago by about 8,500 French settlers. Such event allows variants to rapidly reach fixation through pure *genetic drift*, a phenomena called *gene surfing* [96], which mimics *genetic hitchhiking*.

1.5.4 Has a region of interest evolved under positive selection?

One major challenge while assessing whether a region of interest has evolved under the action of positive selection is certainly to circumvent the confounding factors of both past demographic processes and of the ascertainment bias of the data. For that purpose, one can compute the statistic from some method designed to detect footprints of positive selection (see Section 1.4.2) and estimate its significance by comparison to a reference distribution. This reference distribution must reflect the expected score under selectively neutral evolution with the data used. Indeed, values of statistics are relative to the studied population and to the kind of data analyzed, rather than absolute. There are two main approaches to obtain such reference distribution: the *outlier approach* and the use of simulations.

Table 1.2: Some different demographic processes which can leave molecular patterns expected under positive selection.

Process	Description	Molecular pattern
Migration	Individuals move from one population to other(s)	Increased genetic variability within each population. Lower genetic differentiation among populations
Isolation	One population is isolated from the others and drifts on its own	Increased genetic differentiation among populations
Population structure	The studied population is actually structured into several ones	Higher variability than expected
Population expansion	The population increases rapidly in size	Increased number of rare variants and decreased variability
Population bottleneck	The population decreases rapidly in size and retrieves its original size after several generations	Increased number of rare variants and derived alleles at high frequency and decreased variability
Founder effect	A new population is founded by a small number of individuals from a larger population. The new population then increases in size	Gene surfing: mutations that occur on the frontier of a growing population are more likely to expand and get fixed since only a few individuals are founding the population.

Using simulations accounting for demography.

Since the development of coalescence theory [97–99] (described below) and the recent wealth in computational capacity, simulations became a powerful approach in population genetics. It is now possible to generate large independent data sets fitting the real data for several features accurately assessed through simulations of genetic data that mimic population demographics and evolution. Those data sets are, in turn, used to assess the statistical significance of empirical data. Particularly, one can simulate sets of genetic data under a neutral model and appropriate demographic parameters to infer what the empirical data would look like without the action of positive selection, and then a significance threshold at a given false positive rate (FPR) can be estimated using the distribution of the simulated neutral data. In this case, any putative bias from empirical data are eliminated. Furthermore, if one is interested in evaluating the reliability of the estimated threshold, simulations incorporating selective events to the neutral model can be ran in order to infer the power of the approach.

The simulation software that have been implemented so far can be divided into the ones relying on coalescent theory and the ones based on forward simulation. Coalescent simulation is the first widely approach that has been used to simulate genetic data at the sequence level and, as the name suggest, is based on coalescent theory first introduced by John Kingman in 1982 [98]. It relies on a backward model describing the characteristics of the joining of lineages back in time to the most recent common ancestor (MRCA) as showed in Figure 1.14. It represents the theoretical background for most of neutral genetic models, as well as the estimation of many population genetic parameters. If a population has an effective size of N_e, the probability that two lineages, i.e. gene copies, are derived from the same parent in the previous generation, i.e. coalesce, is $1/2N_e$. The coalescence time of the lineages sampled for the study in previous generation follows a geometric distribution with mean $2N_e$. Therefore, if there are p lineages at a given generation, the probability that two coalesce at the previous generation, thus reducing the number of copies to $p-1$ is $p(p-1)/4N_e$, and the expectation time that any

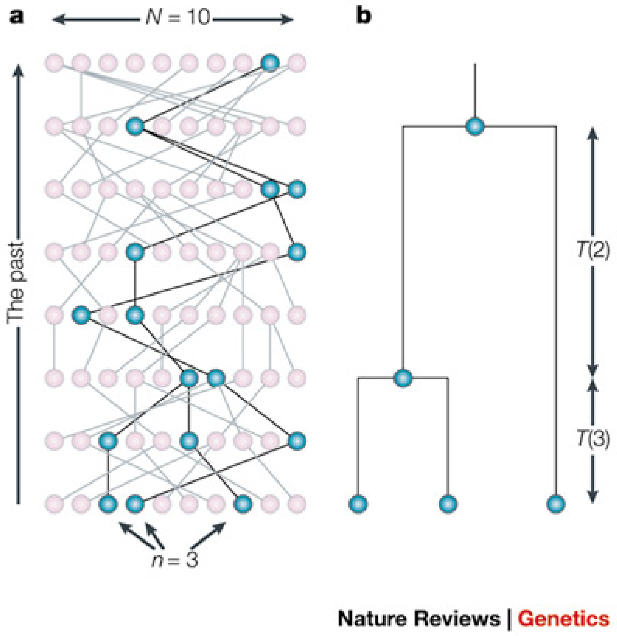


Figure 1.14: Coalescence. a. The complete genealogy for a population of ten haploid individuals is shown. In coalescence theory, diploid individuals are considered as two independent haploid ones. Here a sample size of three haploid individuals is considered, and their ancestries back to a single common ancestor are shown with the black lines. b. The subgenealogy for the three sampled haploid individuals. Coalescence needs to keep track of the times between coalescence events and the topology. From [100].

pair of lineages coalesces is $4N_e/p(p-1)$. From those simple equations, the main conclusions of coalescence theory are : (1) while the rate of coalescence ($p(p-1)/4N_e$) increases with the sample size (p), it decreases with the effective population size (N_e); (2) time to coalescence increases when the process moves towards the MRCA; and (3) the probability of the MRCA of the samples being the same as of the population is $(p-1)/(p+1)$, and thus, even small sample sizes have a high probability of including the MRCA. The coalescence theory provides computational efficiency and several coalescence simulation software have been implemented, such as *FastCoal* [101], *CoaSim* [102], *SelSim* [103], *cosi* [10], *ms* [104] and *msms* [12].

For many of the underlying coalescent models, parameters have been calibrated to fit empirical data in order to retrieve the past demographic history of human populations. For example, Schaffner *et al.* used HapMapIII data to infer the demographic history of three populations (see Figure 1.2) through the calibration of their model to make the simulated data match empirical data for pairwise F_{ST} values, LD decay (how LD for pairwise SNPs decreases with physical distance in the genome) and SFS [10]. Further implementations used more complex empirical data features, such as the joint SFS across populations [105]. Those programs allow to simulate genomic regions spanning few megabases in hundreds of samples without important computational costs in term of time and resource. This is particularly useful to compute large simulated distribution of the statistics of interest and, thereby, estimate the statistical significance for the studied genomic region. However, coalescent simulations present several limitations. The most important one is limited accuracy in simulating the number recombination and gene conversion events, as well as the level of recombination patterns that is possible to include in the model. As a consequence, when a realistic recombination map is incorporated into the model one, only few megabases can be simulated with an increased computational cost. Otherwise, with a simplistic recombination map, the simulated genomic regions can be longer but the model is likely not to be accurate. Another traditional issue with coalescent simulations is the incorporation of selective events. Although efforts have been made to circumvent this limitation, e.g. see [12, 103, 106], usually it is at the cost of over-simplifying other aspects of the model, such as recombination map, population changes, sample size and length of the simulated genomic regions.

To circumvent the limitation faced by coalescent simulations, the forward simulation approach has been explored as an alternative. Genomic data is simulated forward in time from an ancestral status, allowing much more flexibility to the model. Then, complex recombination patterns and other genomic features (gene content, background selection) can be considered, for example see *SFS_CODE* [107]. The demographic processes included in the model can also present a much higher layer of complexity (see [108, 109]). However, such approach requires the simulation of the whole population and, therefore, is very computationally expensive, preventing its use for generating large data sets. If one is interested in obtaining a neutral model for human demography, Excoffier and colleagues implemented a coalescent model, *fastsimcoal2*, which allows a high level of demographic complexity, with serial founder effects, range expansions and admixture among populations [11]. This model overpasses forward simulation models such as *dadi* which is the reference in the field [11].

As mentioned above, the models are calibrated to make the simulated data fitting the empirical data. Therefore, when the empirical data contains ascertainment bias, it is important to either correct for it [110] or to take into account in the estimation procedure [111, 112]. However, it is not an easy task and the calibrated model can inaccurately reflect past demography (but see [11]). In addition, most models also rely on *a priori* assumptions on demographic events and therefore accurate models are available for a reduced numbers of well-studied populations.

Outlier approach.

As mentioned before, constructing a neutral model using simulation is computational expensive and the model does likely not incorporate all the layers of demographic and genomic complexity. One may prefer to use the outlier approach: an empirical distribution of the statistics to detect positive selection is built from a large number of loci across the genome.

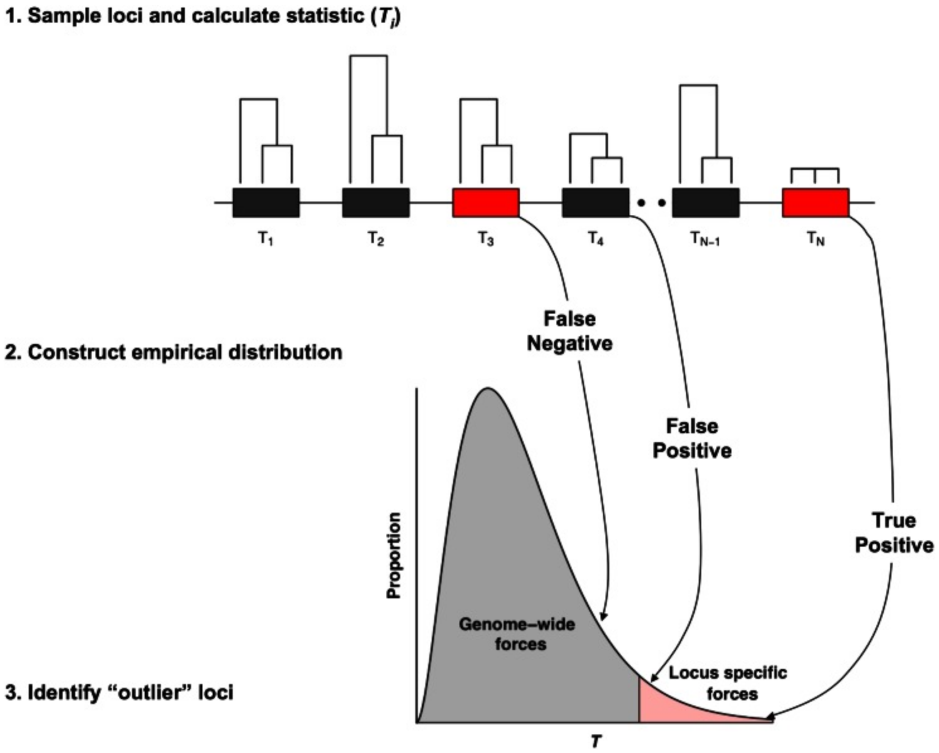


Figure 1.15: Outlier Approach. A typical study design based on outlier approach. From many sampled loci for which a statistic designed to detect positive selection have been calculated (1) are used to build an empirical distribution (2). Then, for the position of some loci of interest within the distribution is observed to infer the empirical significance (3). From [113].

The loci located in the extreme tail(s) of the distribution, i.e. outliers, are then considered as possible targets of positive selection (Figure 1.15). The assumption behind this framework is that demography affects stochastically the whole genome in the same overall way, while positive selection, a deterministic process, affects only a few loci, and thereby, does not distort the distribution. Moreover, such approach allows to correct for ascertainment bias, as well as the confounding effect of background selection if the reference loci are accurately sampled. However, the genome can be seen as a mosaic of several chunks, each with its own history. Although, if the population definition is accurate, the chunk demographies may be very similar, some specific genomic regions may have extreme molecular patterns that mimic positive selection and would be inaccurately identified as having under positive selection, i.e. there would be false positives [114]. This seems to be particularly the case when positive selection targets recessive rather than co-dominant allele, when it acts on a standing variant rather than on a newly arose mutation (more on standing variants below) and when a population bottleneck occurred [115]. Another difficulty of the outlier approach is the arbitrary threshold used to consider a scores as significance. Indeed, setting such threshold requires to define *a priori* which is the proportion of the genome expected to be under positive selection. For example, if the 5% more extreme scores are considered as being under putative positive selection, the underlying assumption is that 5% of the genome is expected to be so. However, there is no accurate estimate of such proportion and it remains the main question assessed by people studying positive selection. Finally, the outlier approach only identifies the most extreme case of positive selection, and many of the selected alleles with a relatively low selective coefficient are likely to be false negatives.

Combination of different tests.

Assessing statistical significance for a given score through either simulations or the outlier approach is a required step to contrast whether a genomic region has been evolving under positive selection. However, it is very delicate to make sure that a significant score is not actually a false positive and, when using only a single method it always remains a substantial doubt when concluding that a locus of interest has been targeted by positive selection. To reduce the risk of false positive, it is wise to use different methods developed to detect the impact of positive selection at molecular level. Particularly, one may use methods based on different kinds of molecular footprints left by a selective sweep (SFS, LD and genetic differentiation). This way, the false discovery rate is likely to be reduced: the false positives from individual methods are unlikely to overlap, since each method are sensitive to different demographic processes. Zeng *et al.* implemented two compound tests, DH [116] and $DHEW$ [117], which combine different SFS-based methods: Fay and Wu's *H* [59] and Tajima's *D* [66] for the former DH [116], and adding the Ewens-Watterson test [118] for the latter. Focusing on DH, the underlying idea is that Fay and Wu's *H* and Tajima's *D* are sensitive to population bottleneck and expansion, respectively [116], but each is insensitive to the other demographic process. Thus, combining the two tests is robust to both demographic processes. The idea is very simple, using neutral simulation, one set a join threshold of significance for both tests for a given FPR. Afterwards, if a region of interest is significant for both tests, it is identified at being targeted by positive (Figure 1.16). However the original method relies on neutral simulations with rather simplistic demography using *ms* [104]. One can use the framework suggested by Zeng *et al.* under an outlier approach as in Luisi *et al.* (2014) (see Chapter 6) by computing Fay and Wu's *H* and Tajima's *D* in a large number of genomic regions to make the reference distribution and estimate the join threshold of significance.

A more simplistic manner is to use any combination test, i.e. a test that combine individual test *P*-values, such as the Fisher combination's test:

$$Z_F = -2 \sum_{i=1}^{K} log(P_i) \tag{1.4}$$

where P_i is the *P*-value associated to the score of the i^{th} test.

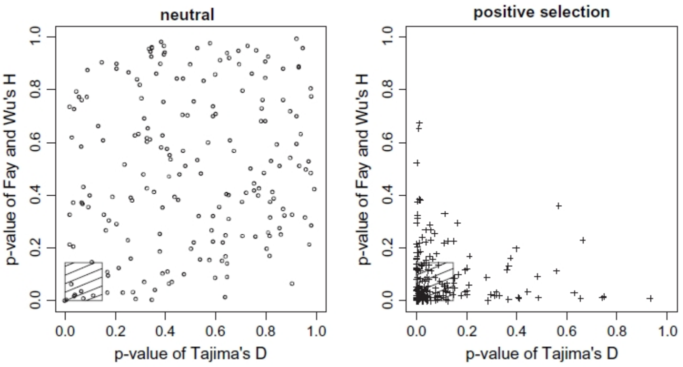


Figure 1.16: The DH test. A neutral distribution is computed for both Tajima's *D* and Fay and Wu's *H*, here from coalescence simulations (left panel). A join threshold is obtained at a given False Positive Rate and represented by the bottom left square. Any region located within this square is considered as having evolved under positive selection. On the right panel, simulations with selective events are used to infer the sensitivity. From [117].

Following this idea, Grossman *et al.* implemented a Composite Multiple Score (CMS) [106] which multiplies *P*-values of five individual tests based on long haplotypes ($XPEHH$, ΔiHH and iHS) and genetic differentiation (F_{ST} and ΔDAF) [106, 119]. The main improvement from rather simplistic combination score is that they computed *P*-values from simulations using the demographic model calibrated by Schaffner *et al.* [10] under a neutral scenario and with selective event. Then, the CMS is obtained as following:

$$CMS = \prod_{i=1}^{K} \frac{P(s_i|selected) \times \pi}{P(s_i|selected) \times \pi + P(s_i|unselected) \times (1-\pi)} \tag{1.5}$$

where s_i is the score of the i^{th} method, the *P*-values are obtained from reference distribution from simulations under either neutral ($unselected$)

or selective ($selected$) scenarios and π is the uniform prior probability of selection.
The combination tests and CMS can not use any kind of tests since they rely on the assumption of the independence among tests. Moreover, they attribute to each test the same contribution to the combined score. In Pybus *et al.* (submitted; see Chapter IV), an alternative framework, *Boosting*, to incorporate the information from different methods. Based on *Boosting* functions [120], this framework allows to detect and classify selective events. *Boosting* is a Support Vector Machine (SVM) [121] which is trained on simulated data to estimate the best regression function of scores from different individual methods to distinguish between two scenario. Making profit of a neutral model with demography [10] to which a selective sweep scenario can be incorporated (as in [106]), thousands of genomic regions have been simulated under a selectively neutral scenario and 45 selective ones (Figure 1 in Chapter IV). Then, two *boosting* functions have been trained to distinguish among those scenarios, namely, (1) evolution under either pure *genetic drift* or with partial selective sweep (where the selected mutation reaches a final allele frequency -FAF- of 0.2 and 0.4); (2) evolution with incomplete selective sweep (FAF = 0.6 or 0.8); and (3) evolution with complete sweep. Two further *boosting* functions have been also built to further classified regions evolving under complete and incomplete sweep as recent or ancient sweep (as defined in Figure 1 in Chapter IV). Those functions are included in a classification tree as shown in Figure 5 in Chapter IV. Such framework allow to combine different tests, although relatively correlated, but more interestingly, to classify the mode of positive selection for for the detected selective events. A look at the standardized coefficients for each coefficient give valuable insight on the methods that contribute the most to distinguish between two given scenarios, and thus, on their performance to detect a given selective event (Figure 6 in Chapter IV). Moreover, the boosting coefficients are quite similar for the three populations considered (namely YRI, CEU and CHB), and thus seems quite robust to demography.

On one hand there are clear evidences of morphological and physiological adaptations in modern human populations, such as pigmentation for solar radiation, body size for thermal condition, blood flow and oxygen delivery for high altitude. On the other hand, there are few examples of fixed, or almost fixed, genetic differences among populations and/or validated cases of adaptive mutations (see Section 1.6 for an overview). This striking inconsistency between the number of known phenotypic and genotypic adaptive examples, may be explained by the reduced vision of the action of positive selection that have been followed until recently. Indeed, until now, most studies of natural selection relied on the *hard sweep model* (see Section 1.4.2) since they make use of methods designed to detect molecular patterns expected to remain in the genome after such a sweep. In order to have a complete vision of adaptation and the genomic processes allowing it, it is also important to consider other modes of positive selection. Those alternative events of positive selection do not leave the same molecular footprints as expected after a *hard sweep*, and theoretical development is still required. Particularly, the two following alternative mode of positive selection have been recently studied after being overlooked for many decades [122].

Soft sweep.

Recently, works at both empirical [123–127] and theoretical [128–132] point to the importance of *soft sweeps* which can occur through two different modes of adaptation:

1. **Selection on standing variants** which, in opposition to a *hard sweep*, does not rely on the appearance of an advantageous mutation to arise in the population, but rather targets a variant already segregating at relatively important frequency when a change of environment occurs. (Figure 1.17).
2. **Selection on recurrent mutations** where, the derived allele which is advantageous arise in the population several time independently, as a result of recurrent mutation or gene flow from another population. All copy of the derived allele increase in frequency until the allele reaches fixation. If, all copies of the derived allele have a

similar selective coefficient (because the genetic background does not have affect it through, for example, intragenic epistatis), none would fix during the selective event (Figure 1.18).

In both cases different copies of the selected allele may belong to different haplotypes: in the former case because it was already segregating on different haplotypes before the selective event, while in the latter, because it arise on different haplotypes. Anyway, in both cases tests based on long haplotypes are not suited to detect such mode of adaptation. However, if the selective pressure is population specific, and thus the selective event occurred in a given population, methods based on genetic differentiation may be able to detect it. In addition, other haplotype patterns, beyond the EHH, are informative (see below).

1.5.5 Selection not only by hard sweep.

Polygenic adaptation.

Recent genome-wide association studies (GWASs) comfort the view of classic quantitative genetics that many phenotypes are encoded by several dozens, hundreds or even thousand, rather than an unique one [133]. This drastically contrasts with the reduced vision of positive selection acting on a single advantageous mutation to drive phenotypic adaptation. Therefore, more focus on polygenic adaptation is required. Such mode of adaptation would drive simultaneously a limited shift in allele frequency at several variants located in different genomic regions and with small effect on fitness (Figure 1.19). Such shift are extremely difficult to distinguish from pure *genetic drift*.

Recent methodological advances in detecting alternative sweep scenarios.

As appearing in Figures 1.19, 1.18 and 1.17, the molecular patterns expected to be left by *soft sweeps* and polygenic adaptation are not as evident as the ones left by *hard sweeps*. Therefore, there is still an evident scarcity in methods designed to detect such selective events at the genetic

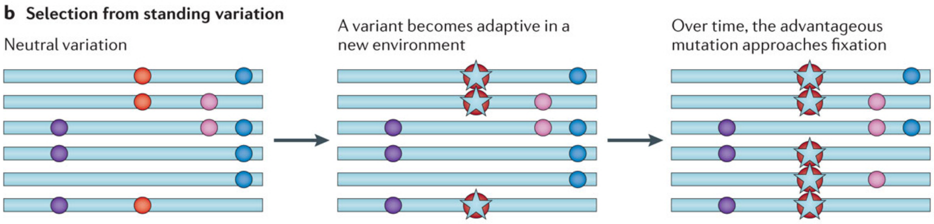

b Selection from standing variation
Neutral variation
A variant becomes adaptive in a new environment
Over time, the advantageous mutation approaches fixation

Figure 1.17: Selection on standing variant. The variant is already segregating in the population and becomes advantageous and increases in frequency until it reaches fixation, or almost fixation.From [134].

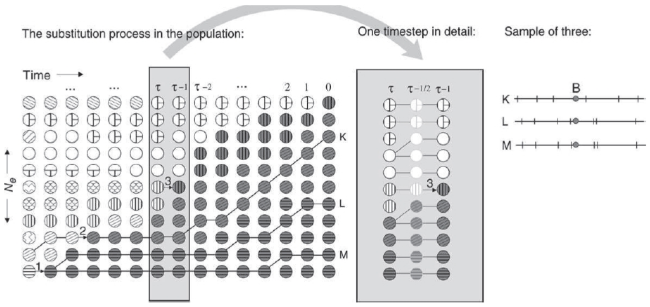


Figure 1.18: Selection on recurrent mutation. Considering the schema of Wright-Fisher model. Circles represent individuals, the different patterns indicate independent ancestral haplotypes. The beneficial allele *B* (dark gray individuals) substitutes the ancestral b allele (white). The *B* allele arises three times by independent mutation; individuals then change their color from white to gray but keep their haplotype pattern. The zoom into a single time step shows how reproduction and mutation are separated. Directly after fixation (time 0), we take a sample of size three (K, L, M) that contains descendants from the first (L, M) and the second (K) mutational origins of B. The right panel shows DNA fragments of the sampled individuals. The vertical ticks represent neutral polymorphisms. Individuals L and M share a recent ancestor and are identical in this region of the genome. Individual K carries a different ancestral haplotype. From [131].

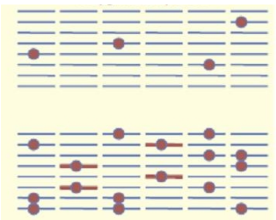


Figure 1.19: Polygenic adaptation. Before the selective event, advantageous alleles located in different genomic regions may already segregate in the population at different frequencies (upper panel). During the selective event they all increase slightly in frequency. The shifts in allele frequency would depend on the variant effect on fitness (lower panel). From [122].

level. However, ongoing methodological development is in progress. First, some already existing methods can be used to detect *soft sweeps*. Indeed, as mentioned above, if the selective pressure is population specific a locus-based statistics of genetic differentiation, e.g. F_{ST} [78], may be powerful provided the variant is segregating at low frequency in the reference populations. Furthermore, iHS [71] shows good sensitivity when positive selection acts on a standing variant that was still segregating at low frequency before the selective event [135]. Two other methods relying on specific haplotypic patterns have been recently developed [135, 136]. First, nS_L [135] is based on the comparison between EHH for derived and ancestral alleles, as for iHS, but also takes into account the length of the segment of haplotype homozygosity between a pair of haplotypes. Besides showing greater power than iHS for scenarios where the advantageous allele was already presents in the population at frequency $> 3\%$, it does not need any genetic map and it is robust to recombination rate and mutation rate. Second, the $H12$ and $H2/H1$ statistics [136] also rely on homozygosity of multiple haplotypes. $H12$ use the combined frequency of the first and second most frequent haplotypes observed in a genomic region as following:

$$H12 = (p_1 + p_2)^2 + \sum_{i>2} {p_i}^2 \tag{1.6}$$

where p_i is the frequency of the i^{th} most common haplotype in the sample.

The $H12$ has power to detect *hard sweeps* and - not so- *soft sweeps*, i.e. when the starting frequency is below 0.1%. In order to distinguish between those two scenario Garud *et al.* further developped the $H2/H1$ statistics [136]:

$$H2/H1 = \frac{\sum_{i>2} {p_i}^2}{\sum_{i=1} {p_i}^2} \tag{1.7}$$

where p_i is the frequency of the i^{th} most common haplotype in the sample.

$H1$ and $H2$ are expected to be higher and lower under *hard sweep* than under *soft sweep* scenario, respectively. Therefore $H2/H1$ increases with the softness of the sweep, i.e. the number of haplotypes on which the advantageous mutation is segregating previous to the selective event.

Those two recent methods demonstrate that accurate theoretical implementation allows to detect *soft sweeps* although the molecular patterns are more difficult to recognize. Further effort is required to increase the power to detect even *softer sweeps*. However, the reduced shift in allele frequency expected under polygenic adaptation leave very weak footprints in the genome. Hence, although it could be argued that increasing the sample size would increase the power, implementing methods using only genetic information seems a loosing battle. For this reason, the few methods proposed until now, include other kind of information. First, the $BayENV$ [137, 138] method uses environmental variables. Indeed, it is based on the correlation between allele frequency and an environmental variable observed at many populations. It provides a Bayes Factor for each studied locus which is the ratio between two Bayesian posterior probabilities:

1. Under the null (neutral) model, the correlation we observe in allele frequencies between different populations is just explained by demographic factors (basically *genetic drift*, migration and population size changes)
2. Under the model where a specific environmental variable has been

a selective pressure in one(s) population(s) and then may have imbalanced the allele frequency spectrum across the populations.

Therefore, this method is detecting variants that shifted similarly in allele frequency in populations facing the same environmental pressure compared to their neighbouring populations (Figure 1.20). Such parallel selection, recently theoretically analysed by Ralph and Coop [139], is more likely to occur on rather ancient variants that are shared among worldwide populations. Note that the signal is driven by the consistency of the shift in allele frequency across populations rather than by its amplitude. Such method corrects for population structure and therefore is less sensitive to demography than a simple correlation analysis. Indeed, the genetic differentiation among populations is directly related to their geographic distance (Figure 1.21) due to the *isolation by distance* phenomena. However, retrieving environmental variables from many populations may be challenging, especially because it relies on representative geo-localization. Another approach was suggested by Mendizabal *et al.* [141] in order to include phenotypic information rather than the selective pressure. More precisely, the authors have analysed the covariance between allele frequencies and height measurement to detect genetic variants allowing Pygmy adaptation to the rainforest climate through reduced size for thermo regulation purposes (the Bergmann's rule) as shown in Figure 1.22. Such approach would requires extensive phenotypical measurements but the authors implemented a permutation procedure allowing to only use the average and variance of the phenotype of interest from literature. This method can be describe as a method to detect advantageous variants only if one acknowledges that the phenotype under study arose from an adaptive process.

H. Allen Orr also suggested a sign test [142], to test whether the observed number of plus (or minus) alleles at *Quantitative Trait Loci* (QTLs) is different in two groups of individuals with different phenotype, instead of being similar as expected under pure *genetic drift*. The Orr's sign test has recently been used for *expression QTLs* (eQTLs). Even if each eQTL has low effect on phenotype, the accumulation of alleles increasing (or decreasing) the phenotype points to polygenic adaptation. Similarly, an

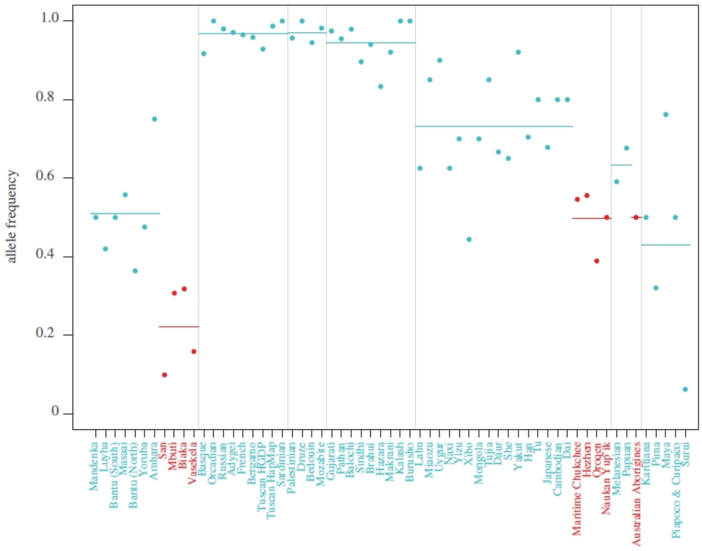


Figure 1.20: Example of SNP detected by $BayENV$. Populations are ordered by main geographical region and shown on the x-axis. The y-axis represents the allele frequency by points for in each individual population or bars for the average. Populations sharing one given mode of subsistence are shown in red. From [140].

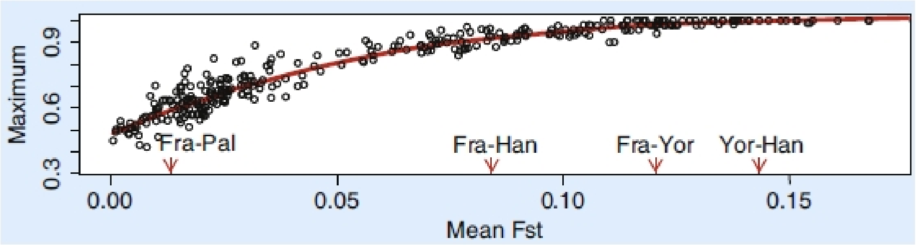


Figure 1.21: There is no strong differentiation between close populations. The mean and the max F_{ST} between population pairs are shown on the x-axis and y-axis respectively. Fra: France; Yor: Yoruba; Han: Han Chinese. From [122].

alternative is to use a set of SNPs associated with a given phenotype, e.g. height in European populations [143], and show a systematic allele frequency differences between populations with different phenotypic values that better fits a model of adaptive evolution than *genetic drift*. Finally, Berg and Coop [144] have implemented a test using the mean additive genetic value, Q_X, estimated from the additive effect size of loci associated to a given phenotype (GWAS loci). The test is an extension of the $BayENV$ method and contrasts whether the genetic value (instead of the allele frequency) covariates with a given environmental variable. They further developed a generalization of the Q_{ST}/F_{ST} comparison [145]. The Q_{ST}/F_{ST} test of neutrality contrasts whether there is an excess of quantitative trait differentiation (as measure by the Q_{ST} index) compared to the genetic differentiation among populations (as measured in a large set of loci by the F_{ST} index), to identity traits that have evolved adaptively. In their implementation Berg and Coop use the estimated Q_X instead of Q_{ST}.

The theoretical development to identify variants with small effect on fitness but at the basis of phenotypical adaptation through polygenic adaptation is progressing. However, most of the methods rely on GWAS loci, and as a consequence, are still limited. Indeed, first they assume that the associated loci act in a strictly additive manner, putting aside the putative dominance or epistatis among them. Second, GWAS loci are unlikely to be the causal ones, but rather tag them; therefore, since the LD patterns are variable among population, the GWAS loci may not be a good proxy of the causal one in all the studied populations. Third, the genetic values are relatively accurate when calculated in a population where the association studies were performed, but the GWAS loci may not be portable to any genetic background, i.e. in any populations.

1.5.6 From putative advantageous mutation to the increased fitness.

Most studies set as a goal the identification of the advantageous mutation. This goal can be reach if, at least, the four following steps are completed

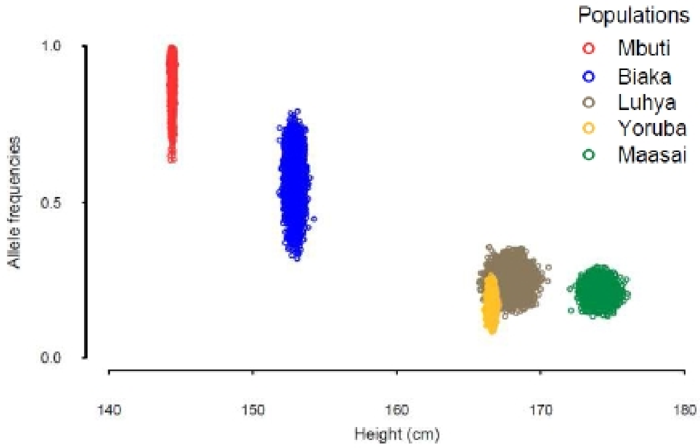


Figure 1.22: Covariance of genotype and phenotype. Covariance of allele frequency and height for a given SNP. From [141].

(Figure 1.23).

1. **Identify candidate adaptive loci**. The main issue is to disentangle whether a statistic for detecting positive selection is extreme at a loci because of the impact of positive selection or alternative processes aforementioned.
2. **Identify the underlying functional variant**. For that purpose, one must get rid of the strong LD within a genomic region which has faced *hitchhiking* in order to pinpoint the variant targeted by positive selection.
3. **Quantify the phenotypic consequences of the candidate adaptive allele** by performing experiments*in vivo* with model organism (mouse, zebrafish, etc...), *in vitro* using call cultures, or genotype-phenotype association studies. An alternative is to use the wealth of functional public database to retrieve such information from the literature.

4. **Clarify the relationship between phenotype and reproductive fitness** in the population and environment where the allele has increased in frequency. This is a complicated task because one must infer what was the relevant environment which acted as a selective pressure on the ancestors of the actual studied population, and whether the phenotypic change encoded by the functional variant is fitter than the ancestral one.

There are very few studies which presented conclusive results for the four steps together (see Section 1.6). Particularly, the fourth step appeals to story-telling and it is for now impossible to formally test such relationship in humans. Therefore, it is important not to dismiss the possibility that a locus has been adaptive when facing the inability to determine the past selective pressures and to demonstrate that the phenotypic change induced an increase in fitness in the past populations.

In the future, the recent wealth in *omics* data will most probably allow to —partially— bridge the gap between genotype and phenotype when studying adaptive evolution. Indeed, thanks to NGS data, functional data has been produced in the past few years in epigenomics, metabolomics, transcriptomics and interactomics, among others. For example, the Encyclopedia of DNA Elements (ENCODE) project has identified functional elements in across the genome, being in coding or non-coding regions [147]. In order to identify the underlying functional variant, one may use this emerging functional data, for example through an integrative genomics approach, along with results from population genetics of positive selection (Figure 1.23).

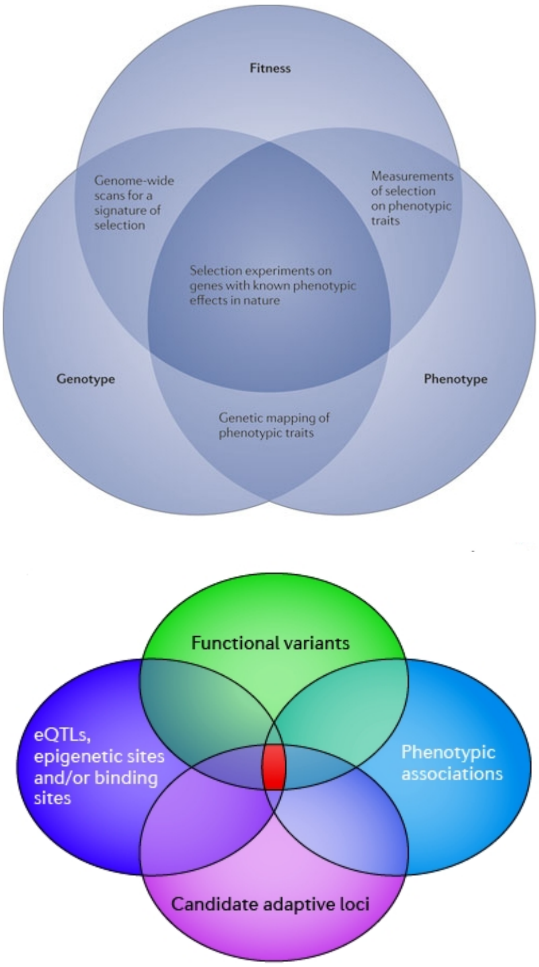

Fitness
Genome-wide scans for a signature of selection
Measurements of selection on phenotypic traits
Selection experiments on genes with known phenotypic effects in nature
Genotype
Phenotype
Genetic mapping of phenotypic traits
Functional variants
eQTLs, epigenetic sites and/or binding sites
Phenotypic associations
Candidate adaptive loci

Figure 1.23: An integrative approach. **Upper Panel**: the overlap among various approach is informative to understand the molecular basis of phenotypic adaptation; from [146]. **Lower Panel**: use of functional data within an integrative genomics approach. One may use several types of data to identify functionally relevant regions of the genome. By focusing on the variants putatively selected, with documented function and associated to a given phenotype, it may be possible to build lists of candidate regions that have adaptively evolved; from [134].

Table 1.3: Examples of positively selected genes supported by functional evidence. Caution: an unique article is cited while for many genes, several studies were required to conclude both on the impact of positive selection and on the function of the putative selective allele.

Gene	**Selected function(s)**	**Adapted population**	**Approach**	**Reference**
ABCC11	ear wax secretion	Asian	Genome-wide scan	[157]
CASP12	sepsis resistance	worldwide	Gene-candidate	[151]
CCR5	bubonic plague or smallpox resistance	European	Gene-candidate	[158]
CD5	pathogen recognition	East Asian	Gene-candidate	[152]
DARC	malaria resistance	African	Gene-candidate	[123]
EDAR	hair/teeth/sweat gland development	Asian	Genome-wide scan	[70]
EGLN1	response to hypoxia	Tibetan and Sherpa	Genome-wide scan	[159, 160]
EPAS1	response to hypoxia	Tibetan and Sherpa	Genome-wide scan	[160, 161]
G6PD	malaria resistance	African	Gene-candidate	[148]
HBB	malaria resistance	African	Gene-candidate	[149]
HERC2	eye pigmentation	European	Gene-candidate	[162]
LCT	lactase persistence	European and African	Gene-candidate	[124, 150]
SLC24A5	skin pigmentation	European	Gene-candidate	[163]
SLC45A2	skin pigmentation	European	Genome-wide scan	[70]
TLR5	bacterial flagellin	African	Genome-wide scan	[119]
TNFSF5	malaria resistance	African	Gene-candidate	[65]
ZIP4	Zinc uptake	West Africa	Gene-candidate	[164]

1.6 Current knowledge on positive selection in the human genome.

The previous section (1.5) makes emphasis on the practical challenges to (1) detect positive selection in the genome, (2) confirm which are the adaptive loci, and (3) to link the genotype to the phenotype. Although, such challenges are numerous and have avoided to reach a complete knowledge of the past human adaptation, there have been several striking successes since the advent of the genomic area one decade ago (Table 1.3). Studies of the impact of positive selection can be divided into the ones focusing on candidate genes and the genome-wide scans.

1.6.1 Candidate gene studies of positive selection.

Candidate gene studies focussing on a gene are driven by an *a priori* hypothesis on the implication of a gene in a putatively adaptive phenotype. Before the recent wealth in genomic data, they used to be the most peformed analyses of positive selection. They allowed to understand the impact of positive selection on specific genomic regions, to identify candidate adaptive loci and provided informative insights into the molecular basis of phenotypic adaptation across human populations. For example, several genes have been identified to have been targeted by positive selection with supporting functional evidence for the candidate adaptive locus and a link to a phenotype change conferring a fitness increase (Table 1.3): *G6PD*, *DARC*, *TNFS5*, *HBB* which provide malaria resistance in Africa [65, 123, 148, 149]; *LCT* proffering lactose resistance in population with herder ancestors in Europe [150] or Africa [124] ; *CASP12* increasing resistance to sepsis [151]; and *CD5* allowing better pathogen recognition [152].

Although the aforementioned successes in detecting variants that have been selected for, the candidate-gene approach suffers from the three following main drawback.

1. An *a priori* hypothesis is required about which genes have been

under positive selection, as well as a knowledge of the relationship between genotype and phenotype. Candidate-gene approach aims to pinpoint the functional variant, but the goal is rarely reached. Furthermore, when the function of the adaptive allele is established, it is difficult to stress how it confers a selective advantage to its carriers.

2. When there is prior insights on the genes that could have been involved in phenotypic adaptation, the adaptive variant can be located far away from the region spanning the gene, rather being within the coding or flanking region. In that case, if there is no previous knowledge on the gene regulatory regions, it would be impossible to detect the adaptive locus within a gene-candidate framework.
3. There is, in general, no sufficient biological knowledge on the molecular basis of adaptative phenotype (or even of diseases) across the genome to make good *a priori* hypothesis of the underlying molecular bases of traits. Thus, candidate-gene approach is reduced to the study of annotated genes encoding relatively simple phenotypes.

For those reasons, with the recent wealth of polymorphism data, an alternative approach have been also used: the genome-wide scan approach.

1.6.2 Genome-wide scans for positive selection.

During the last decade, impressive technological progresses have been made to obtain genomic data from high-throughput genotyping arrays to NGS, leading to a bulk of genotype data to perform population genetics analyses. Now, large catalogs of genetic variability in worldwide human populations are publicly available (see Section 1.2.2) allowing to study the impact of natural selection on our genome. For that reason, a large number of genome-wide scans of positive selection in different populations has been published in the last years (reviewed in [113, 133, 134]). Such *top-down* approach, with no *a priori* hypothesis on the adaptive phenotype, allowed to overcome the limitations of candidate-gene studies. The first genome-wide scan for positive selection in human populations was performed by Akey *et al.* in 2002 [153] and was followed by more than 20 other ones. Since 2002, the number of individuals and markers available increased consequently (see Section 1.2.2), and theoretical development allowed the implementation of several new methods for both *hard sweep* and alternative modes of positive selection (see Sections 1.4.2 and 1.5.5). The multiplication of the data and statistical methods to detect positive selection, obviously engendered a multiplication of the genomic regions that have putatively evolved in at least a population. Already in 2009, more than 5,000 regions in the genome spanning a total of 400Mb and encompassing more than 4,000 protein-coding genes were reported in a review of 21 genome-wide scans published at that time [113]. Those 21 scans used methods designed to detect the molecular patterns left by a *hard sweep*. They also relied on the outlier approach and, therefore, established *a priori* an expected proportion of the genome under positive selection in the studied populations, likely leading to an important false positive rate. Indeed, in his review [113], Joshua Akey looked at the overlap of the genomic regions reported by 10 studies using the same data, but different statistics. Strikingly, only 14.1%, 5.3% and 2.5% of the overall regions were reported in two, three or four studies, respectively (Figure 1.24). Besides the FPR issue, it is clear that those genome-wide scans can also miss some real event of selection as suggested by the fact both *G6PD* and *DARC* have never been reported by such studies.

Although the overlap among individual scans is low, more than 700 regions have been identified encompassing previous candidate adaptive loci and new well-supported ones (Table 1.3). Moreover, it appears that most signals of putative positive selection are not shared among populations from different geographic regions (for example see [71, 154]). This is expected when considering that the scans mostly relied on the *hard sweep* model, and therefore detected advantageous mutation that appears in the population just before being selected for. Indeed, geographically distant population present different genetic background and have to adapt to very heterogeneous environmental conditions.

Genome-wide scans allow to build maps of putative signals of positive selection and will still give great insights on how natural selection has shaped the human genome. They will also keep on helping the discovery of functional elements. However, it remains challenging to extract the relevant information in the bulk of signals of positive selection from genome-wide scans in order to understand how human population really evolved and what is at the molecular basis of phenotypic adaptation. Indeed, although the genome-wide approach circumvents some limitation of the gene-candidate, it presents its own ones.

1. Large scale studies do not allow to extensively control for many layers of complexity. Indeed, in opposition to gene-candidate approach, performing a genome-wide scan it is extremely difficult to build an accurate model including both demographic and genomic processes to describe the evolution of a specific genomic region or to investigate in depth the molecular mechanisms affecting the genetic variability. Therefore, most scans rely on outlier approach, and as mentioned before, only detect the most extreme cases of positive selection as well as suffering a likely important FPR [115]. As already mentionned, one solution to reduce the FPR, is to rely on different scans performed with different methods and/or on different populations.
2. Regions reported by genome-wide scans are usually large, spanning hundreds of kilobases and containing several contiguous genes and regulatory regions. On the other hand, sometimes signals can be located in intergenic regions where no function has been reported yet. Therefore, it is often difficult to follow-up the signals to identify the selected variant and the phenotype putatively increasing the fitness.
3. For most genes, it requires a quite important amount of speculative discussion (*story-telling*) to state which could be the adaptive phenotype.

For those reasons, most genome-wide scans focus on a very reduced signals of putative selection based on biological information for a follow-up analysis. This practice is often referred as *cherry picking*. Hence, most of the signals already reported remain to be explained. The recent scan performed by Grossman *et al.* [119] set up new standards to overcome the aforementioned limitations and represents an important step toward the identification of putative adaptive variants as well as the underlying phenotypes increasing the fitness. This study rely on several progresses: (1) they used the CMS which allows to pinpoint more accurately the selected variant ([106]; see Section 1.5.4); (2) they performed their analysis on the 1,000 Genomes Project Pilot 1 data [155]; and (3) they analysed the putative phenotypic implications of the selective variants by interrogating the ENCODE database [147] as well as the GWAS catalog [156].

1.6.3 Insights from published studies of positive selection in humans.

All the studies aforementioned allowed the identification of putative adaptive loci, but also provide interesting insights on more general questions on the nature of the genomic regions that have been preferentially targeted by positive selection in human populations. Thus, they allow to understand what are the phenotypic differences among populations and species that are induced from adaptation to new environments and which were the underlying biological functions at play.

Functional categories for the selected protein-coding gene.

A functional enrichment analysis is almost always performed after a genome-wide scan for positive selection. Such analysis basically tests whether the set of variants located within the regions with a signal of positive selection is enriched in a given biological process or functional pathway. In other words, it contrasts whether there are more of those variants belonging to a given functional class or pathway than expected by chance from the background list of loci included in the study. To perform a functional enrichment analysis, one may use one of the following databases.

1. **Geno Ontology** [165] groups genes according the the features of the gene product. There are three main domains: (1) cellular com-

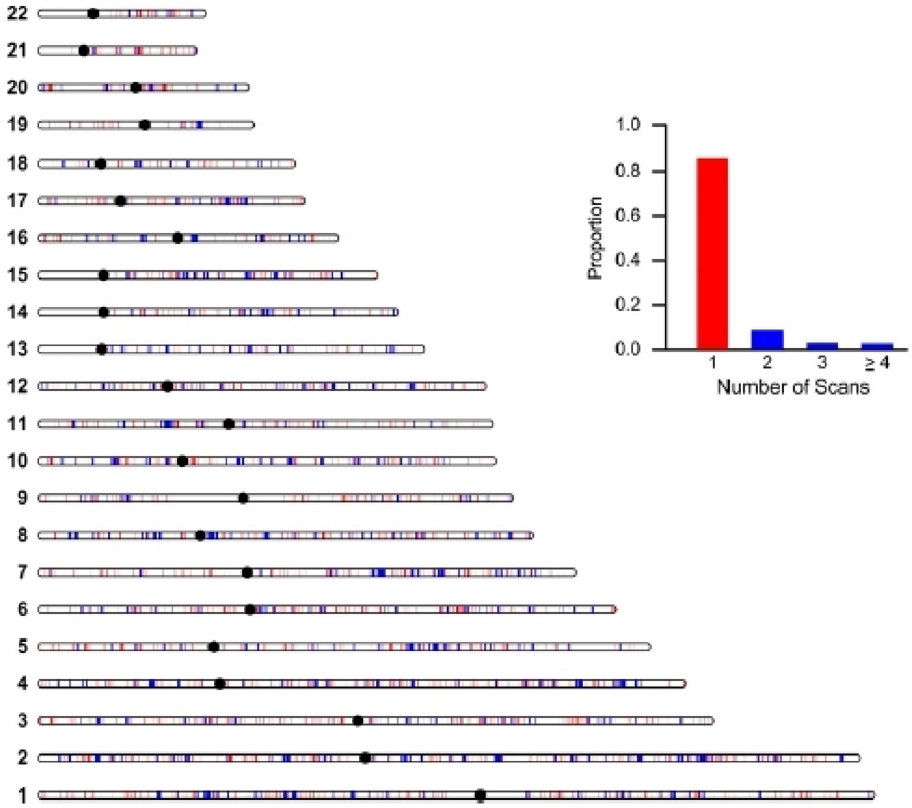


Figure 1.24: A map of signals of positive selection from 10 genome-wide scans. Genomic regions reported in at least one genome-wide scan for positive selection. The histogram shows the number of regions overlap among those scans. From [113].

ponent, i.e. the parts of the cell or its extracellular environment where the gene product is active; (2) molecular function, i.e. the elemental activities of the gene product at the molecular level (e.g. binding, catalysis, etc...); and (3) biological process, i.e. operations and sets of molecular events with a defined beginning and end and pertinent to the functioning of integrated living units.

2. **PANTHER** (Protein Analysis Through Evolutionary Relationships) [166] relies on annotation fron Gene Ontology among others and classifies proteins (and the encoding genes) according to either: (1) family, i.e. groups of evolutionarily related proteins; and subfamily (related proteins that also have the same function); (2) molecular function of the protein by itself or with directly interacting proteins at a biochemical level; (3) biological process, i.e. the function of the protein in the context of a larger network of proteins that interact to accomplish a process at the level of the cell or organism, e.g. mitosis; or (4) pathway which also explicitly specifies the relationships between the interacting molecules.
3. **KEGG** (Kyoto Encyclopedia of Genes and Genomes) [167] is a collection of manually curated databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances.
4. **Reactome Pathway Database** [168] contains curated functional pathway annotations that cover a diverse set of topics in molecular and cellular biology.

Genome-wide scans of positive selection using polymorphism data in human populations pointed to different categories enriched for gene that have evolved under a selective scenario: skin pigmentation, immunity, hair density and sweat gland, etc. [114]. Scans based on comparative genomics have revealed categories such as immunity and pathogen defence or sensory perception [169, 170].

However, functional enrichment analyses using such databases are biased toward protein-coding gene. In addition, they assume that all the genes are independent and that there is no interaction among them and do not attribute any weight according to the gene importance within a pathway or a functional category. Although functional enrichment analysis has shed light on important functions and pathways being preferentially targeted by positive selection, it does not provide a formal test for selection acting on a function. The current approach commonly used for large genome-wide analysis of positive selection is to detect signals at individual genes or regions. However, selected loci are just at the molecular basis of positive selection acting at phenotypic level. Thus, single mutations rarely act in isolation to improve a function or to contribute to the acquisition of new ones. To overcome those limitations, Serra *et al.* created a new method called the Gene Set Selection Analysis (GSSA) to detect significant differences in scores of natural selection over functionally related genes [171]. The method was applied genome-wide to coding regions of five mammals. But it still has never been used to interrogate non-coding elements or for polymorphism data.

Complex adaptive traits.

The studies listed above describe the first intents to move from individual genes to the biological modules they belong to. These studies start from individual genes or loci to then integrate the information on functional modules. The idea behind is that, except for Mendelian traits, many loci will be involved in phenotypic adaptation. This implies that polygenic adaptation is likely to be the main adaptive force acting on the human genome. First, Daub *et al.* used a gene-set enrichment test based on the of F_{ST} statistics ($SUMSTAT$) to functional pathways or gene sets enriched in differentiated loci among populations [172]. They found that most of the pathways enriched in such loci are more or less directly involved in the immune response. This result confirms the general idea that response to pathogens have been a major selective pressure for human populations (for two reviews see [31, 173]). They also observed evidences for epistatic interactions between members of the same pathway. A genome-wide scan although detected several signals of selection for genes involved in the hypoxia-inducible factor 1 (HIF1) pathway which is involved in physiological response to hypoxic conditions [159].

In order to move towards such mode of adaptation, several studies used

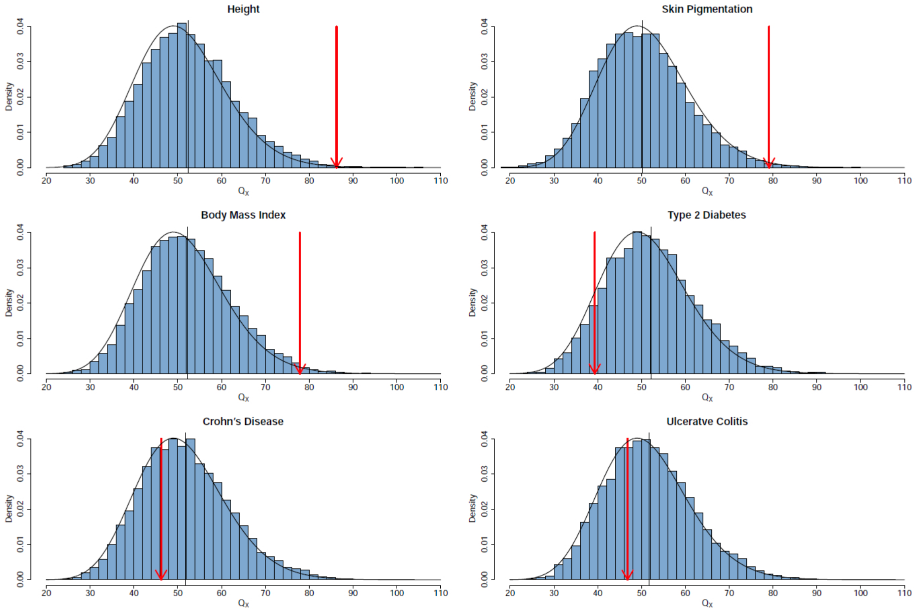


Figure 1.25: Mean genetic value Q_X for several phenotypes. The red arrow shows the Q_X value compared to a reference distribution built from genome-wide resampling of well-matched SNPs. Heigh, Pigmentation, Body Mass Index are good candidate for adaptive phenotype or for being closely related to any adaptive henoype, while Type 2 Diabetes, Chron's Disease and Ulcerative Colitis are not. From [144].

methods better suited to study small shift in allele frequency [140, 174, 175] (see Section 1.5.5). Looking at covariation of diet, subsistence or ecoregion, Hancock *et al.* found that pathways involved in starch and sucrose metabolism are enriched in signals of polygenic adaptation to a diet rich in roots and tubers, as well as an over-representation of signals associated to polar climate in genes involved in energy metabolism pathways [140]. Applying the same method with other environmental variables, they also described an enrichment of signals in gene sets related to UV radiation, infection and immunity. On the other hand , Fumagalli *et al.*, using a similar method, showed that local adaptation has been driven by the diversity of the local pathogenic environment while climate played a relatively minor role [174].

Berg and Coop, using the mean additive genetic value (Q_X, described in Section 1.5.5) described several complex traits likely or not to have evolved through the action of polygenic adaptation [144] as showed in Figure 1.25.

The importance of regulatory elements.

Although the method proposed by Berg and Coop [144] is limited because of relying on GWAS loci and the problem of portability among populations, it is representative of a major shift in the field. Indeed, it is more and more recognized that focusing only on protein-coding elements is not enough to understand adaptive evolution in humans. Indeed, although protein-coding sequence are very well annotated, they only represent around 1.2% of the human genome. Furthermore, the important similarity between humans and chimpanzees in their protein-coding gene sequences can not explain the observed phenotypic differences. In 1975, King and Wilson [176] suggested that differences in gene regulation may largely account for those phenotypic differences among species but also populations. Since 1975, the relative contribution of variants located within protein-coding genes and regulatory regions has been debated. One evidence from the functionality of non protein-coding regions is the amount of conservation among species across the genome. Indeed, 5% of the genome has been estimated to have been largely conserved since the

MRCA of mouse and human through the action of purifying selection. Hence, this conserved proportion of the genome is likely to be somehow functional [177]. Since this proportion is higher than the proportion of protein-coding sequences in the genome, a large fraction of the elements with relevant biological function is non-coding.

Until recently, very few evidence has been provided on the adaptive role of non-coding elements because of technical limitations. On one hand, the annotation outside the genes had been lacking, therefore, it is difficult to distinguish any functional evidence to any putatively adaptive locus. On the other hand, comparative genomics studies, which rely on the comparison of the rate of substitution on functional versus non functional elements struggle to find any equivalent to the non-synonymous and synonymous changes classification. However in the recent years, a group of evidences pointed to the role of regulatory elements in adaptive evolution. Using as reference putatively neutral elements the variants located in ancestral repeats and pseudogenes, Haygood *et al.* found an important amount of promoter regions with signatures of positive selection in the human and chimpanzee lineages [179]. Strikingly, they found an enrichment of signals of selection in promoters related to nervous-system functions. Recent population genetics studies also point to the same direction. First, Kadaravalli *e al.* using a genome-wide set of eQTLs to interrogate for positive selection using iHS [71], found that SNPs showing signals of selection are more likely than random SNPs to be associated with gene expression levels in *cis* [180]. Second, with a similar study design but taking advantage of the recent wealth in eQTL databases and the recently published ENCODE project [147], as well as using $BayENV$ scores for polygenic adaptation ([137, 138]; see Section 1.5.5), Fraser provided the first genome-scale support for the hypothesis that changes in gene expression have driven human adaptation [178] shown in Figure 1.26. Third, Enard *et al.* observed a greater correlation of the observed signatures of positive selection (as inferred by iHS [71], $XPEHH$ [70] and CLR [67] and correcting for background selection) with the presence of regulatory sequences from ENCODE [147] than with the amino acid substitutions (Figure 1.27) [91].

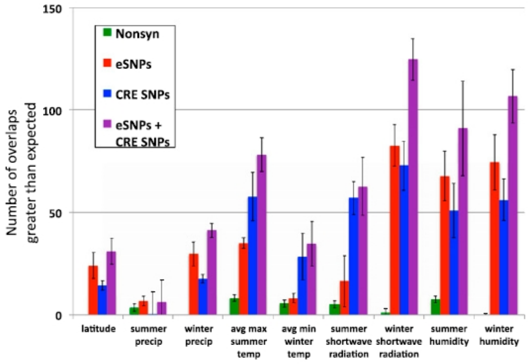


Figure 1.26: Gene expression drives local adaptation in humans. The estimated number of putative local adaptations associated with each of nine climate/geographic variables that are explicable by either a nonsynonymous SNP (green), Gene expression-associated SNPs (eSNP; red), *cis*-regulatory elements SNP (CRE; blue), or combined eSNP/CRE SNP (purple). Error bars indicate the standard deviations when randomly sampling negative control SNPs. From [178].

1.7 The network framework.

1.7.1 Interest of biological networks to understand natural selection.

Under the selectionist view, the community began to address the subtle evolutionary mechanisms contributing to the diversity observed in natural populations. A consequent switch occurred through the work of Richard Dawkins and synthesized in his book “The selfish gene” [181] which put the gene at the centre of evolution. In this theory, instead of following the traditional view, inherited from Darwin’s work, in which organism as a whole was the target of selection, Dawkins stressed that genes themselves were directly targeted. Indeed, he assumed that the own propagation of those “selfish” genes matters more than the success of other genes segregating in the same organism. The organism as a whole was pushed into a

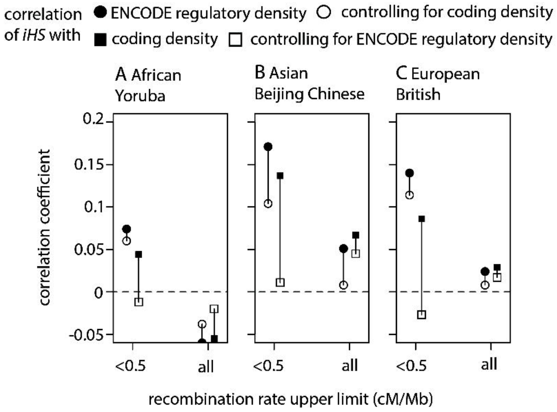


Figure 1.27: Most human recent positive selection occurs in regulatory sequences. The filled circles and squares show the correlation coefficients of the absolute values of iHS with the density of regulatory and coding sequence density, respectively, controlling for recombination and average pairwise diversity (covariate of BGS). The open circles and squares show partial correlations. From [91].

position of secondary importance: its function is merely a "survival machine" for the genes to segregate [181]. Under this view, the biological role of altruism could be explained: alleles promoting altruist behaviour may allow their own survival by helping their fellows to survive, being in the same host organism or in a close related one. Moreover, this theory struck a chord in the genetics community, since it could also justify why some genes extensively replicate even to the detriment of the organism. Although Dawkins' view of natural selection remains wide-spread, it presents different caveats, most formulated by Stephen J. Gould and Ernst Mayr. Especially, genes are not directly exposed to the environment, and only the phenotype not the genotype interfaces with natural selection: the survival of the organism only depends on the viable phenotype resulting from the genes function. The selection at the genome level is only a consequence on the selective pressures endorsed by the phenotype. Moreover, most of the phenotypes result from the joined action of several genes: the phenotypic effect of one gene deeply depends on the genetic background in which it segregates. This mechanism, called epistasis between mutations within and among genes. Inter-genes epistasis, which can be viewed as complex interactions, can not be assessed by the gene-centric view proposed by Dawkins. Epistasis is said to be positive (or synergetic) when a combination of mutations has more effect on the fitness than the additive effect of individual mutations. On the other hand, negative (or antagonistic) epistasis occurs when interacting mutations induce a lower fitness effect than expected by adding the individual effect of each mutation. Finally, sign epistasis describes the inversion in fitness effect of a mutation when in the presence of another mutation. The phenotype is encoded through the effects of several genes in a non-linear way, and it is the prime target of natural selection. Hence, one expects to observed (1) that the interacting partners of a protein would affect the evolution of the gene encoding it; and (2) evolutionary patterns within the biological systems.

There is now a body of evidences showing the prevalence of epistastis emerging from biological systems [182, 183]. Particularly, Dobzhansky-Muller incompatibilities, also known as compensatory mutations have been describe to be a relevant mechanism driving protein evolution [184]. This phenomena is an extreme case of sign epistasis, when two deleterious mutations are beneficial when segregating together (reciprocal sign epistasis). Moreover, genes encoding interacting proteins tend to exhibit correlated evolutionary histories (for review, see [185]). Indeed, such genes tend to duplicate almost simultaneously [186, 187] and to evolve at relatively similar rates [188–193]. Such co-evolution of genes may result from the co-evolutionary dynamics of the proteins they encode, for example when the deleterious effect of mutations in a protein is compensated by other mutations in interacting proteins. Other factors might also drive such coevolution : similar expression levels and/or function of interacting proteins [194, 195]. All together, it is now accepted that interacting partners of a protein affect the evolution of the gene encoding it.

Therefore, the consideration of biological systems in which proteins interact to accomplish a given function may be informative to understand gene evolution, and one would expect some evolutionary patterns within those systems. Indeed, the interacting genes have uneven importance on the function of the biological system, and since natural selection acts on this function, its impact on genes is expected to differ from gene to gene. A straightforward way to study an interacting system of proteins is to represent it through a network. The present thesis focuses on this network framework, and before reviewing the studies trying to relate the action of natural selection -either positive or purifying- on genes to their position and role in the functional networks, a brief description of the methodology adopted is required.

1.7.2 Biological pathways and their representation as networks.

> It is an essential characteristic of experimentation that it is carried out with limited resources, and an essential part of the subject of experimental design to ascertain how these should be best applied; or, in particular, to which causes of disturbance care should be given, and which ought to be deliberately ignored.
>
> *The Design of Experiments*
> Sir Ronald A. Fisher

Three main types of biological networks.

The biological networks are key systems that describe the basic mechanisms that govern life and functioning of the cell. Genes and their products, i.e. proteins, interact in several ways. The global function of the cell, or even the organism, is led by all the different interactions occurring simultaneously. However, each interaction type is driven by different mechanisms and, thus, operate under different constraints. Treating those different types of interactions separately may be an useful simplification.

Indeed, the present knowledge of molecular biology is too limited to allow an ideal integration of all the active elements in the cell into a single biological system, and trying to describe all the complexity of all the cellular interactions is not feasible yet. One would prefer to focus on a specific aspect of the interactions in order to study its own properties.
Therefore, biological networks can be divided into three main families according to the type of interactions at stake:

1. **Gene regulatory networks** govern the expression levels of messenger RNA (mRNA) and proteins in the cell. They are composed of several DNA sequences which interact with each other -and with other molecules- indirectly in the cell. The main elements of those networks are the transcription factors (TFs), which turn on or down the transcription of other genes by binding to their promoter region at the start (in 3') or at the end (in 5') region. In multicellular organisms, different cells perform different functions with the same genomic information thanks to the control of the genes that are turned on and expressed. Gene regulatory networks are the main drivers of such control of the cell function.

2. **Signal transduction networks** are activated when an extracellular signaling molecule binds a specific receptor located either on the cell surface or inside the cell. The activation of the receptor, in turn triggers a chain of biochemical events inside the cell, creating a response to the signal and, thus performing a given biological function. Signal transduction networks are mainly composed by physical protein-protein interactions (PPIs) : phosphorylation which activates the function of the protein by transferring a phosphate group to a protein, dephosphorylation, proteins binding through their binding sites to be functional together, sometimes aggregating as protein complex. The proteins involved in those transduction cascades can be active inside or outside the cell, and can be classified according to their function as defined by their domains.

3. **Metabolic networks** represent a series of chemical reactions, catalyzed by enzymes, to transform an initial molecule to form another

product. During the metabolic reaction chain are produced metabolites, which are the product of a given reaction and the substrate of the following one. Enzymes are the active elements of metabolic networks and are linked by the metabolites shared between the reaction they catalyse.

.

Topological representation of biological networks.
Graph theory provides useful tools to describe the structure of the biological networks. A network is made of nodes (active elements) related one with the others by edges (interactions). Table 1.4 shows the nature of the nodes and edges depending on the three kind of biological networks.

Table 1.4: Elements and their interactions within the different types of biological networks. A simplistic overview of the active elements (nodes) and their interactions (edges) composing three different types of biological networks

	Nodes	**Edges**
Metabolic Network	Enzymes	Shared Metabolites
Signal tranduction Pathways	Proteins	Physical Interactions
Gene Regulatory Network	Transcription Factors	Regulatory Relationships

Graphs theory allows to compute several descriptive statistics representing the topology of a given network, in order to discriminate the elements composing it. Particularly, there are a body of measures to describe the centrality of each element. The three most widely use are :

1. **Degree centrality** (or connectivity) is the number of edges of a node. Within a biological network, it is the number of interactions a given protein participate to.
2. **Betweenness centrality** is the number of shortest paths that pass

through a node [196]:

$$c_B(v) = \sum_{s,t \in \mathbf{N}} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{1.8}$$

, where v is the node of interest, s and t are two other nodes in the network $\mathbf{N}$, $\sigma(s,t)$ is the total number of shortest paths between s and t (minimal sequences of nodes that connect s and t), and $\sigma(s,t|v)$ is the total number of shortest paths between s and t that pass through v. Nodes acting as *information bridges* will be assigned a high betweenness centrality measure.

3. **Closeness centrality** is the reciprocal of the sum of the shortest path distances between a node and all the other nodes in the network [196]:

$$c_C(v) = \frac{n-1}{\sum_{u=1}^{n} d(v,u)} \tag{1.9}$$

where n is the number of nodes in the network and $d(v,u)$ is the shortest path distance between nodes v and u. Notably, high values of closeness should indicate that all other nodes are in proximity to node v. In contrast, low values of closeness should indicate that all other nodes are distant from node v.

An important assumption.

When studying evolution of genes involved in a biological network, one obvious assumption is the fact that network topology is fixed. While it is conceivable to think that network structure might affect and constrain the possibilities of evolution of individual genes having a role in the network, in turn, the evolution of individual genes changes the network itself by adding and removing nodes and edges as a consequence of their own evolution. However, even if the influence presumably goes in both directions, the rate of link dynamics (gain and loss of edges) is estimated to be much slower than the rate of protein sequence evolution [197]. This keeps down the magnitude of the effect of the assumption of fixed topology and makes it of practical use to study the influence of network structure on genes' evolution.

Two different biological scales for evolutionary network analsyses. The studies of the patterns of molecular evolution within biological networks can be divided into two lines of investigation distinguished by the biological scale considered (Figure 1.28):

1. **Small-scale network** which represents a particular biochemical system of interest in order to gain insight into its specific evolutionary histories. This approach focus on the coordinate effect among genes performing a given biological function. In that case, succesive functions are determined on the basis of well established molecular knowledge on the process. Then, one studies the impact of natural selection on the gene evolution integrating information on the players at succesive steps interacting together across a given biological network describing such small and well annotated pathways. Although, such small-scale approach is interesting in the sense that it focuses on a specific biological function, which is, again, the prime target of natural selection, it considers that the biological system is totally isolated from the others in the organism.

2. **Large-scale network** which considers the complete set of interactions of a kind comprised by the organism, in order to detect universal patterns of evolution. Such scale allows to incorporate the information on cross-talks among pathways and gene pleitropy (when a gene influences multiple, seemingly unrelated biological functions). Taking the pleotropic effect of a gene is of significant importance when considering gene evolution, as predicted by Ronald Fisher [33] in its geometric model of adaptation (Figure 1.29). Nevertheless, using such large-scale networks present its own drawbacks: the information on the interactions may be incomplete and contain a relatively important amount of errors since it is mainly generated from high-throughput analyses. The large-scale networks can also be divided according to the type of interactions considered: the metabolome describes all the metabolic interactions among enzymes, the interactome is the whole Physical Protein-protein Interaction map (PPI) and the whole regulatory network. In any case, the retrieved network is fully dependent on how the full set of data

was obtained, and thus, is technology dependent.

Moreover, Khurana *et al.* integrates the diverse modes of gene interactions to create a unified biological network called MULTINET [200]. Despite the extreme differences in scope when using different network scales, the network representation use the same descriptive measures mentioned above.

1.8 Evolutionary patterns within biological networks.

The evolution of biological networks is two-fold. First, the addition, removal and/or change of elements and their connections account for variation of network structure across time. Second, the elements are themselves evolving entities. If the network structure evolves at slow pace, it can be freeze to interrogate what is its impact on the evolutionary history of its elements. Does network structure constrain the evolution (purifying selection) of specific elements; does it incite innovations to arise (positive selection) at particular positions within the network?

Below, an overview of the studies that have investigated how biological network topology and structure may drive natural selection. The results of those studies will be described according to the scale of the biological networks (either small or large scale; see Section 1.7.2), the mode of natural selection studied (either purifying or positive) and the approach used to detect it (using either polymorphism or divergence data).

1.8.1 Evolutionary analysis of small-scale networks.

Inter-specific analysis of small-scale networks.

The anthocyanin biosynthetic pathway was the first biological network studied within the scope considered in this thesis. The authors used divergence data from three plant species to infer the evolutionary rate of the 6 genes involved in this metabolic pathway [201]. In the following

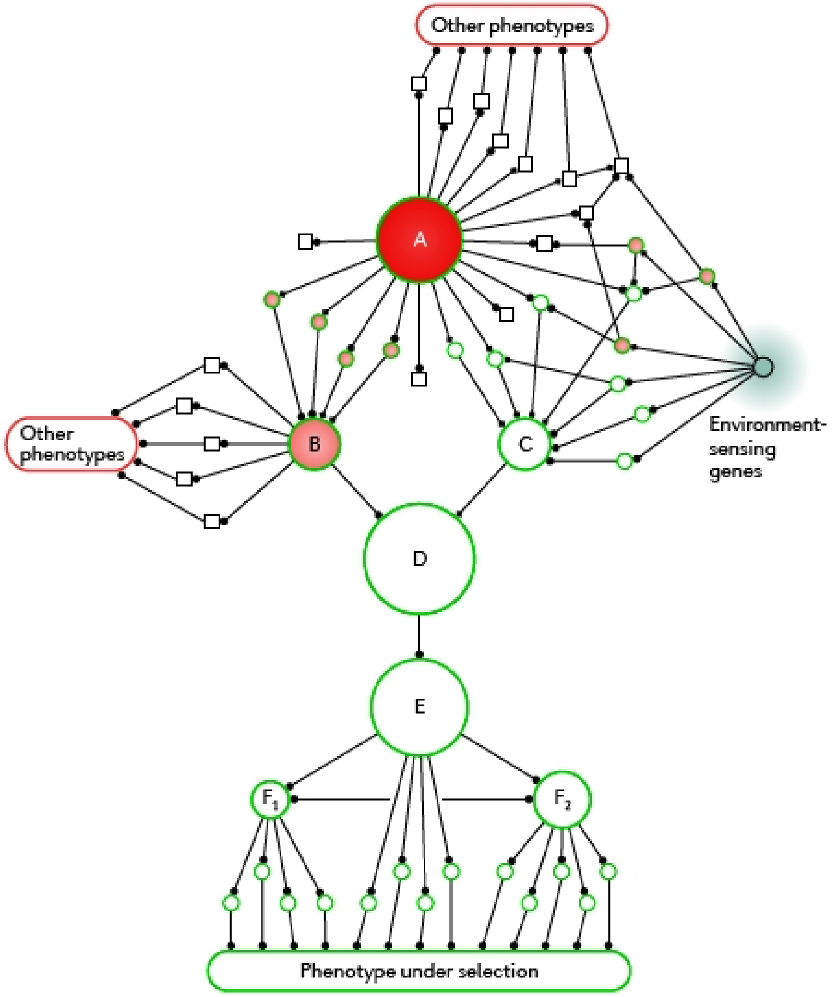


Figure 3 | **Importance of network effects for adaptation.** A gene's position in a

Figure 1.28: Cross-talks among different pathways. A gene's position in a network influences its effects on a target phenotype and on other traits. Circular node sizes are proportional to the gene's effect on the selected phenotype; the intensity of red colouration is proportional to effects on other traits (where no colour indicates no effect on other traits). Square nodes have no effect on the target phenotype owing to the directionality of the network, but they may influence other phenotypes. Small black circles indicate the directions of the interactions in the network; modes of interaction are not specified. From [198].

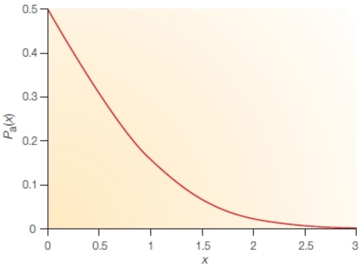


Figure 1.29: Fisher’s geometric model of adaptation. In its geometric model of adaptation [33], Fisher showed that the probability, $P_a(x)$, that a mutation with a phenotypic effect r, is favourable is $1 - \Phi(x)$, where $\Phi(x)$ represents the cumulative distribution function of a standard normal random variable, and $x = r \times \sqrt{\frac{n}{2z}}$, where n is the number of traits -or biological functions- the mutation participates to and z is the distance to the optimum. From [199].

years, other metabolic [202–206] and signaling pathways [207–212] have been analysed using divergence data. All of them are composed of a small number of genes functionally related, with the the role of each within the system well established. Such knowledge retrieved from biochemical research allows to construct accurately the network structure.

In many of these studies, it has been observed that upstream genes in the pathway tend to be more constrained in their evolution -through the action of purifying selection- than downstream genes [201, 203, 205, 207, 213, 214]. A possible reason for this pattern would be that upstream genes are more constraint in their evolution because they are likely to have more pleitropic effect than those downstream. Indeed, since they are more likely to be above some branching points in the pathway, they are involved in the synthesis of more products than downstream genes [201]. Nevertheless, this gradient of decreasing purifying selection along a pathway has not always been found. First, Yang *et al.* did not identify any significant evidence for such relationship in the gibberilin metabolic biosynthetic pathway [204]. Second, the opposite trend has been reported for the insulin/TOR transduction signaling pathway in Drosophila [208] and vertebrates [209], as well as in the N-glycosylation pathway in primates

[206]. Ramsay *et al.* introduced another measure which weights the position of a genes in relation to pathway branch points. This measure, called the "Pathway Pleiotropy Index", counts groups of enzymes between pathway branch points and enzymes between two consecutive branch points are given the same position [205]. In the plant terpenoid biosynthesis pathway, composed of 40 genes, this measures positively correlates with evolutionary rates, as estimated comparing angiosperms sequences of five plant species, better than simple pathway position. This result points the significant effect of pleitropy on evolutionary rates [205]. Two other studies suggest that branch points in metabolic pathways play a relevant role and are critical to evolution: branch points are under stronger purifying selection [204] and more likely to be targeted by positive selection[202]. The impact on metabolic flux of branching points may account for these results [215, 216].

Other studies used centrality measure to assess the importance of a gene within a biological network [210–212], all describing that central genes in the network are more constrained in their evolution, as inferred from divergence data, than genes acting at the periphery.

Intra-specific analysis of small-scale networks.

Until the last three years, very few was known on the relationship between the impact of natural selection on a gene, as inferred using polymorphism data, and the position it occupies in a biological network. Two early studies indicate that genes that act at metabolic pathway branch points are targets of positive selection [217, 218]. An analysis of six genes in the *Arabidopsis* floral developmental pathway suggests that four downstream TF genes have evolved neutrally, while the two earliest-acting genes included in the study present a significant reduction in silent site nucleotide variation consistent with a recent selective sweep [219]. Nevertheless, they did not consider the biological network as a whole, but rather focus on a group of genes with relevant function within it.

Moreover, a study of polymorphism data for the *D. melanogaster* species confirmed the gradient of the levels of purifying selection along the insulin/TOR transduction signaling pathway found using divergence data from either *Drosophila* [208] or vertebrate [209] species. Indeed, at intra-specific level, the downstream genes being the most constrained in the pathway [220]. Finally, Casals *et al.* constructed the innate immunity interaction network to infer the action of both positive and purifying selection using polymorphism data [221]. Although this network does not represent a biological pathway performing a specific function within the cell, it is informative on the relationship between gene centrality and the impact of natural selection. Indeed, they also found that at intra-specific level, selective constraint is greater for gene acting at the core of the network while adaptation (balancing and positive selection) mostly occurred at particular positions at the network edges. When the article presented in this thesis in Chapter 5 was published, the study by Casals *et al.* [221] seems to be the only one on the distribution of adaptive selection across an network of interacting genes using polymorphism data. Since, several articles have been published and are discussed in Chapter 7.

1.8.2 Evolutionary analysis of large-scale networks.

Protein interaction network.
Since the development of the yeast two-hybrid technique, high throughput determinations of physical interactions among proteins allowed to retrieve an important amount of information on the whole map of Physical Protein-protein Interactions (PPIs) occurring in a given organism, also referred to as Interactome, or Protein Interaction Network (PIN). It has been observed that the PIN is organized following a scale-free model [222, 223]: the connectivity (or degree) follows a power-law distribution. Thus, the probability, $P(k)$, of a gene to be involved in k interactions is proportional to $k^{-\gamma}$, where γ is the degree exponent which determines important features of the network. The lower γ, the more important the role of the hubs, i.e. the highly connected elements in the network. The discovery of this PINs feature for many organism, brought to attention the role and importance of node connectivity. Does the topology itself confer some properties to the system? What are its functional implications in case of the PPI networks? Since the scale-free organization has emerged

independently in many biological networks, it suggests that such topology is able to arise from a self-organizing process through the influence or as the result of selective mechanisms. Several studies have studied the biological characteristics of the hubs in order to better understand the putative function(s) of these highly connected elements. Genes encoding proteins acting as hubs in the PIN may have some special features, such as being indispensable for the cell to correctly carry its functions. In other word, if a gene encoding such proteins is removed from the system, could the organism survive? To answer this question the relation between connectivity and other gene features related to fitness have been addressed.

1. **Connectivity and Indispensability**. To estimate the indispensability of a gene, one way is to knock-out it in a model organism and observe whether the gene deletion is lethal. Such experiments in yeast revealed that that highly connected proteins are three times more likely to be indispensable than less-connected ones [224, 225]. Thus, gene indispensability is due to the position within the PIN occupied by the protein it encodes, most likely because when hub proteins are removed the network would be quickly disrupted while it would tolerate the absence of a protein with few interactions.

2. **Connectivity and evolutive constraint**. The patterns described above suggest an obvious expectation: since the removal of highly connected proteins is lethal, any impairing mutation appearing in the underlying gene must be purged by purifying selection. Several studies testing this hypothesis have been published during the last decade. First, Fraser *et al.* reported a negative correlation between connectivity and evolutionary rate using divergence data of *S.cerevisiae* species [188]. Hence, proteins interacting with many others evolve more slowly than less connected ones. Two different explanations can arise from such observations: highly connected protein may (1) have a greater effect on fitness (see the mutational robustness hypothesis introduced by Jeong *et al.* [224]); (2) be more pleitropic since they have a larger proportion of their structure involved in their different functions and thus may be under an

overall higher constraint. Performing a partial correlation analysis, the authors concluded the second choice is more likely : central proteins have a higher proportion of the protein structure involved in its functions. The correlation between evolutionary rates and connectivity is not fully accepted. Indeed, although some alternative studies point to the same direction [225], it has not been fully validated by other studies using different data to describe the PIN as well as other methodology to estimate the rate of gene evolution [226, 227]. The difficulties to confirm such tendency brought light on the accuracy of the networks' reconstruction associated with the technology used [228]. Furthermore, Bloom and Adami used different PINs from different data sets and estimate for each the relationship between connectivity [227]. They deduced that the correlation between connectivity and evolutionary rates might just be a by-product of highly connected genes being highly expressed [227, 229]. Indeed, high protein expression level induces a constraint of the substitution rate in protein sequences [230]. A lesson from this debate would be the importance of considering other genomic determinants that influence the gene evolution, before concluding on a putative relationship between the rate of evolution and gene centrality. For instance, a study of different PINs in yeast yields inconsistent conclusions, even when correcting for the confounding factor of expression level [231]. All together, although the negative correlation between evolutionary rate and gene connectivity seems to be widely accepted, the evidences remain controversial.

3. **Evolutionary rate and other centrality measures**. Beyond connectivity, betweenness and closeness centrality (see Section 1.7.2), other measures have been used to interrogate the putative relationship between gene centrality and three different eukaryotic PINs [232]. The authors reported that d_N negatively correlate with betweenness while d_N/d_S correlate with centrality, independently of the measure used. Although closeness and betweenness are largely correlated to degree (connectivity), they are informative measure to

take into account the global network context by considering proteins beyond the directly interacting partners. Proteins with higher centrality measure likely play a relevant role in the cross-talks between different parts of the network, and thus, have an important pleitropic effect accounting for the selective constraint acting on them. In order to estimate the relative importance of a protein within the network, one can also consider its participation and position in relation to module. Modules are groups of highly connected genes which, together, perform a given function. Under this scope, in *S. cerevisiae*, the evolutionary rate (d_N/d_S) has been described to be lower for local hubs, that is highly connected genes with interactions essentially within modules than for global hubs, i.e. which connect different modules [233]. This observation suggests that modules are quite conserved while their co-action is more likely to be targeted by positive selection.

4. **Positive Selection in Protein-Interaction Networks**. The fraction of genes that shows this signature of positive selection is generally small. However studying the genes that have evolved under putative positive selection is informative because they are the molecular basis of new phenotypic adaptations. A study of the distribution of the events of positive selection within the human PIN [234] using divergence data from human and chimpanzee species, using the d_N/d_S based likelihood ratio test (see Section 1.4.1), demonstrate that they are more likely to occur at the periphery of the PIN (Figure 1.30). The authors claimed that the periphery corresponds to the physical periphery of the cell. This article of the distribution of events of selection within a PIN seems to be the only one preceding the study presented in Chapter 6 which has been carried using both polymorphism data in human populations and divergence data in mammals and analyzing both positive and purifying selection.

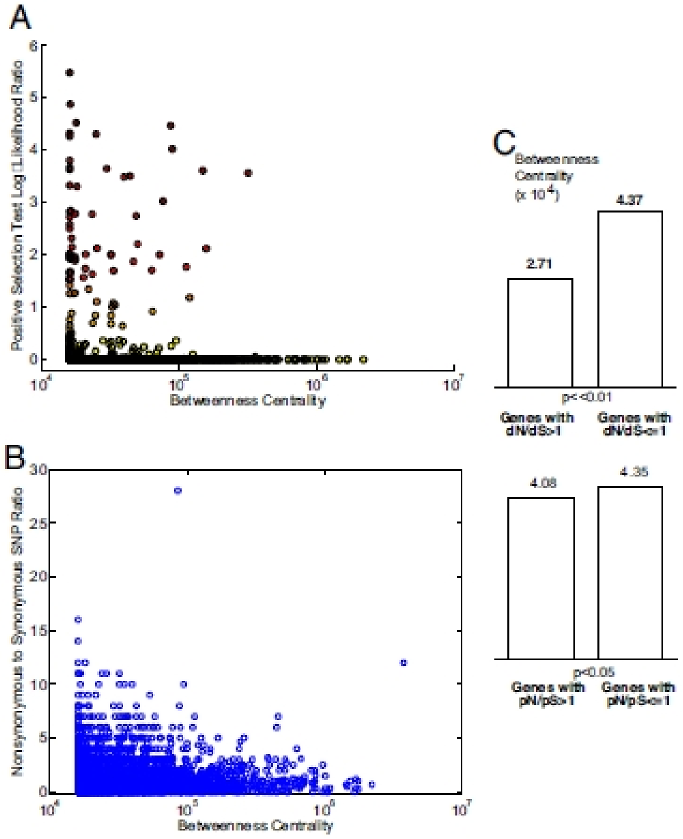


Figure 1.30: Events of positive selection are more likely to occur at the periphery of the Physical Protein-protein Interaction Network. (A) The gene likelihood to be positively selected and betweenness centrality are represented along the *y* and *x*-axes, respectively (Spearman's correlation coefficient, ρ, = 0.06; *P* = 1.2e-06). Dark red and light red points for genes likely to be under positive selection with high and low likelihood, respectively, while yellow points represent genes that has not evolved under positive selection. (B) The gene d_N/d_S score and betweenness centrality are represented along the *y* and *x*-axes, respectively (Spearman's correlation coefficient, ρ, = 0.06; *P* = 1.2e-06) (C Upper) Betweenness centrality of genes under positive selection *vs* all other genes. (C Lower) Betweenness centrality of genes with a high ratio of nonsynonymous to synonymous SNPs (p_N/p_S) vs. genes with a low p_N/p_S. From [234].

Metabolic interaction networks.

Knowledge on metabolic reactions in model organisms has been drawn from decades of biochemical research and its integration with the currently available whole-genome information is allowing the reconstruction of a single organism-scale metabolic network [235–237] which integrates the whole metabolic machinery of an organism. Based on the huge amount of knowledge on the biochemistry of the involved processes, metabolic networks result to present much better annotation of the functions and role within the global process for the elements composing it. Therefore, interpret the phenotype from the the genotype happens to be easier.

In metabolic networks, there is the information about reactions, enzymes and metabolites. Because of this coupled information about the reactions, the enzymes catalyzing them and their substrates and products, different graph representation can be adopted. Each representation captures different aspect of the metabolism machinery and, thus, gives a different focus to the analysis. Two main representations arise [238]: the substrate graph, in which the substrates are the nodes and edges their co-occurrence in the same reaction; and the reaction graph, in which nodes represent reactions and edges indicate shared compounds.

1. **Metabolic network organization**. Using the substrate graph, those networks appear to (1) follow the scale-free organization [224, 239]; (2) be hierarchical [240]; and (3) be organized according the the small-world property [238, 239]. This latter property implies that most pairs of nodes (substrates) can be connected through a relatively short path of reactions, thus, conferring an advantage to the system by enabling it a quick recover of the required concentration of metabolites after perturbations.

2. **Evolutionary Rates and Centrality Measures**. The reaction graph, where the genes encoding the enzymes are the nodes is more informative to study the distribution of the impact of natural selection across the network. This representation has been used for the studies described below. In the *E. coli* metabolic network, no significant relationship has been found between connectivity and the

evolutionary rate, measured with the d_N score [241]. On the other hand, for *Drosophila* and yeast, the d_N/d_S score is negatively correlated with connectivity. This points to the same direction than studies which describe in the same networks that highly connected genes evolve at slower rates, most likely due to the action of purifying selection [242, 243]. Moreover, this relationship happens to be stronger with global centrality measure, such as betweenness [244]. To date, it seems that there is no published analysis carried out with polymorphism data (at intraspecific level) to describe the distribution of natural selection throughout the metabolic network.

Gene regutory networks.

Transcription Factors (TFs) regulate the expression of their targets (which can also encode TFs). The gene regulatory network (or TF network) encodes those regulatory relationships. Although in other networks the interactions are undirected, the gene regulatory network is composed of directed edges (or arcs) connecting each TF to its targets. For directed graphs, connectivity can be divided into two components: the in-degree and out-degree which are the number of incoming edges upon a node and the number of outgoing edges from a node, respectively.

Available network reconstructions may be incomplete since TF-target associations are highly context dependent [245], and it is experimentally possible to only test a few of them. Therefore, one can argue that the gene regulatory networks that have been retrieved to date are largely incomplete. Nevertheless, different large scale compilations of gene regulatory networks are now available and can be used as a proxy of the whole network to study its topological architecture and evolution [246, 247]. In yeast (*S. cerevisiae*), gene regulatory networks are organized into a hierarchical structure and follow the scale-free model [248] in which few TFs are at the top and are not regulated by others. Gerstein *et al.* described a hierarchical human gene regulatory network derived from ENCODE data with a bow-tie structure, with middle level elements having the most information flow bottlenecks [249]. Moreover, Rodriguez-Caso *et al.* also observed that the human gene regulatory network shows the properties of a scall-free and small-world network [247]. The upstream TFs of gene regulatory network are usually activated through a signal transduction pathway in response to an extra-cellular stimuli.

1. **Evolutionary rates and centrality measures**. TF hubs in the yeast TF network do not exhibit lower evolutive rates than other elements [250, 251] suggesting that TF connectivity does not affect its selective constraint. However, another study on a bigger dataset found a significant correlation between the rate of protein evolution and centrality with central TFs evolving faster [252]. The authors also claimed that the higher rate of divergence among central TFs could be due to the action of positive selection because of their role in controlling information flow. On the other hand, the opposite trend was described in the human regulatory network : more connected elements of the network are more constrained (as inferred by the SNP density in the genes encoding it and the d_N/d_S rate) [249]. Hence, the effect of TFs position in the gene regulatory network on the evolution of genes encoding them remains to be elucidated, and further studies are needed.

2. **Gene co-expression networks**. Analysing co-expression networks is also informative to understand gene regulation. In such networks, two genes are linked if they are co-expressed in the same tissues and conditions. The human co-expression network, retrieved from tissue-specific expression profiles, exhibits scale-free properties [253]. This implies evolutionary self-organization through preferential node attachment. The authors also observed that genes with many co-expressed partners, i.e. the hubs, evolve more slowly on average than genes than others, as well as similar evolutive rates for co-expressed genes.

Integration of different types of large-scale networks within a meta-network.

Khurana *et al.* integrated the diverse modes of gene interactions (regulatory, genetic, phosphorylation, signaling, metabolic and physical protein-

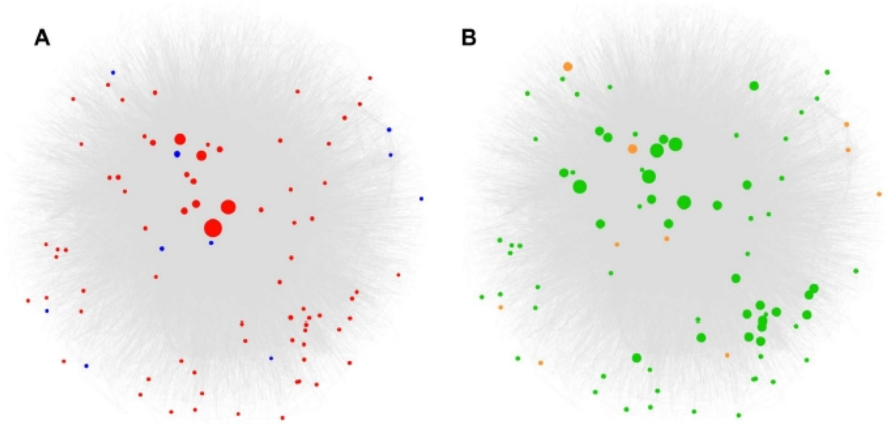


Figure 1.31: MULTINET: a meta-network integrating different types of large-scale networks. The edges of the MULTINET are shown in grey (only the interactions of genes that are involved in more than one network are shown). (A) Nodes corresponding to loss-of-function tolerant and essential genes are shown in blue and red respectively. Size of the nodes reflects the degree centrality of the gene within MULTINET. Essential and loss-of-function tolerant genes tend to be at the centre and the periphery of the network, respectively. (B) Nodes corresponding to loss-of-function tolerant and essential genes are shown in orange and green respectively. Size of the nodes reflects the number of networks the gene is involved in. Essential and loss-of-function tolerant genes genes tend to be involved in more and less networks, respectively. Moreover, most loss-of-function tolerant genes are not involved in any network and are not present in the MILTINET. From [200].

protein interactions) in order to create an unified biological network in humans called MULTINET (Figure 1.31) [200]. Although such meta-network is likely to suffer a large number of errors, it is a step forwards to the integration of many layers of complexity in the functioning of the organism. Indeed, different pathways and networks interact one with the others in a complex dynamical way. For instance, a signal transduction pathway can respond to a stimuli to active a gene regulatory network which to turn on the expression of enzymes involved in some metabolic reactions. The authors then studied how natural selection acts within each individual network as well as in the MULTINET. First, when evolutive constraint was estimated by the d_N/d_S ratio using human and chimpanzee sequences, they found that d_N/d_S values are negatively correlated with their degree centralities in all networks. This shows that highly connected genes tend to be under stronger purifying selection constraints over long evolutionary time-scale. Then, they analysed polymorphism data from three human populations [155], and calculated the average heterozygosity for missense SNPs for each gene. They observed a significant negative correlation between MULTINET connectivity and heterozygosity, suggesting that more variation has been allowed to arise at the periphery of the network. Such trend was not observed when analysing synonymous sites. The higher selective constraint in human populations for gene acting at the core of the MULTINET may account for those results.

Based on this MULTINET and other features (such as selective constraint, participation to the diferrent networks, etc...) of loss-of-function tolerant genes (as defined in [254]) and essential genes (as defined in [255]), they implemented a computational model to predict global perturbation caused by deleterious mutations in all genes with good accuracy (Figure 1.31) [200].

Chapter 2

OBJECTIVES

This work aimed at broadening our knowledge on selective events at the molecular level, using mostly polymorphsim data. The action of positive selection on advantageous genetic variants is at the molecular basis of phenotypic adaptation of a population to a given environment. However, genes and proteins rarely act in isolation, and, therefore they have to be considered within networks describing the interactions occurring among them. Thus, working within a network framework allows to reduce the gap between genotype and phenotype but has been overlooked while studying adaptation. First, positive selection has been detected at gene-level using either candidate-gene (Section 2.1) or genome-scan (Section 2.2) approaches. Then, an analysis of a small-scale gene-network representing a specific transduction signalling pathway has been performed in order to illustrate how the information on protein-protein interactions can be useful to understand how adaptation occurs at biological pathway level (Section 2.3). Finally, a study was carried out to describe the distribution of positive selection throughout the whole protein-protein interaction map to implicitly consider a much larger biological scale taking into account cross-talks among pathways and gene pleitropy (Section 2.4).

2.1 Study the impact of positive selection on a candidate gene.

VKORC1 enzyme is the direct pharmacologic target of oral anticoagulants of antivitamin K type (AVK), such as wafarin and acenocoumarol. These drugs are widely prescribed and the dose recommended vary substantially among individuals according to their genetic composition (e.g. see [256] for warfarin). The enzyme is encoded by *VKORC1* gene. Several genome-wide scans suggest that positive selection was the main force driving the evolution of the 450 Kb region encompassing this gene [257–260]. Furthermore, while interrogating the patterns of genetic differentiation of four loci associated with warfarin dose requirement in HGDP [19], Ross *et al.* showed an extensive geographic differentiation at the site located within *VKORC1* [261]. They revealed that the high frequency of the derived allele in East Asians population accounted for such high genetic differentiation. Moreover, they observed molecular patterns suggesting the past action of positive selection in CHB+JPT population as measured by several methods applied to Hapmap data [20]. Chapter 3 describes a follow-up study on HGDP with extra genotyped SNPs to increase the variant density to interrogate in *VKORC1*. This study allowed to conclude that the signal of selection is restricted to East Asia and shared among all East Asian populations. Moreover, the gene targeted by selection could be either *VKORC1* or another gene located in the 45 kb region covered by the selective sweep signal identified in East Asia.

This study provides an explicit example of the difficulties to pinpoint the target of positive selection. Some *a priori* information on *VKORC1* function and on the selective pressure allowed to start with an *a priori* hypothesis on the adaptive phenotype: large geographic differences in dietary vitamin K intake, especially in vitamin K2, is well-documented; and the highest plasma levels of vitamin K are found in Asian populations. Although the differences in AVK sensibility could be a direct consequence of adaptation of East Asian populations to their diet, it remains impossible to conclude which variant drove the signal without performing functional analysis for the several genes hitchhiked by the selective sweep.

2.2 Scan the genome for positive selection.

While Chapter 3 focuses on a gene-candidate study, Chapter 4 describes a genome-wide scan. For this project, specific populations with known recent history were selected to detect signals of positive selection shared by populations with different ancestry but that have lived within the same environment. Namely, three populations were analysed: (1) one with European ancestry from Romania; (2) a Rroma/Gypsie one which has also lived in Romania for the last thousand years; and (3) a Northwest Indian one. The Rroma/Gypsie population migrated from Northwest India to Europe one thousand years ago [141]. Therefore, Rroma/Gypsie and Romanian populations may have adapted to the European environment with their different genetic background. Particularly, the last thousand year European history has been affected by severe epidemic events (plague, influenza, smallpox, etc.) which may have been severe selective pressures for the immune system. Using the Illumina Immunochip array [262], several tests for positive selection in the three populations have been performed and signals from the same genomic region in Rroma/Gypsies and Romanians, but absent in Northindians were identified.
Chapter 4 gives an interesting example of the genome-wide scan approach. Several signals of putative interest were identified, and a striking one was picked for follow-up studies. Indeed, the molecular patterns observed at Toll-like receptor 1 (TLR1/TLR6/TLR10) suggested that positive selection was the main evolutive force acting on it in Rroma/Gypsies and Romanians. Functional analyses pointed to its role as a pattern recognition pathway of *Yersinia pestis*, the vector of plague.

2.3 Distribution of selective events within a small-scale protein-protein interaction map.

In order to understand the evolution of genes, it is informative to bridge the gap between genotype and phenotype. For that purpose, one may take into account the interaction networks they are involved in. A network-level and population genetics analysis of the Insulin/TOR transduction pathway (IT pathway) is presented in Chapter 5. This study was the first one to describe how recent positive selection, as inferred using polymorphism data, distributes within an interaction network. Indeed previous works focused either on the distribution of purifying selection or on positive selection at much larger evolutionary time-scale using a comparative genomics approach.

It was observed that the most central elements in the pathway have been more targeted by positive selection than other genes. This result contrasts to previous observations in the whole human interactome where positive selection was inferred since the divergence between human and chimpanzee [234]. This indicates that the IT pathway structure affects the impact of positive selection on genes composing it. Therefore, considering the topology of a network representing a given pathway seems to provide important insights on how local adaptation to new environments occurs to efficiently tune a biological function. Further analyses of different pathways are needed to contrast whether this is a general pattern shared among many pathways, or whether it is specific to the IT pathway.

2.4 Distribution of selective events within a large-scale protein-protein interaction map.

As mentioned in Section 1.7, studies of interaction networks can be performed at either small or large scale. Small-scale analyses present the caveat of not taking into account the cross-talks among different pathways and of the artificial delimitation of a specific pathway. The pleitropic effect of the genes is very likely to be an important feature affecting gene evolution, and particularly its likelihood to be targeted by positive selection, as formulated by Ronald A. Fisher [33]. Chapter 6 illustrates a study of the distribution of positive selection within the human protein–protein interaction network (PIN), also referred as Interactome. This large-scale network is a fair representation of all the physical interactions occurring among proteins in the human organism. In a previous study, Kim *et al.*

estimated the likelihood of a gene to be targeted by positive selection through a comparative genomics approach of Human and Chimpanzee sequences [234]. The authors observed that positive selection concentrates at the PIN periphery. The study described in Chapter 6 used both polymorphism data in human populations and divergence data from 10 mammal species to estimate the impact of positive selection at small and large evolutionary time-scales, respectively. It was observed that signatures of recent positive selection are more prone to target genes with high number of interactions while at large evolutionary scale the results were in the opposite direction, with more selective events detected at low connected nodes.

These results show that signals of positive selection at different evolutionary time-scales are distributed in different parts of the interactome. It seems that innovations have a molecular basis with variants in genes with more pleiotropic effects, more indispensable and in general are responsible for strong changes as a result of a mutation. More layers of complexity must be considered to fully understand those results and will be discussed in Chapter 7.

Part II

Results

Chapter 3

STUDY THE IMPACT OF POSITIVE SELECTION ON A CANDIDATE GENE.

Blandine Patillon, Pierre Luisi, Hélène Blanché, Etienne Patin, Howard M. Cann, Emmanunelle Génin, Audrey Sabbagh

Published [263]

Patillon B, Luisi P, Blanché H, Patin E, Cann HM, Génin E, Sabbagh A. Positive selection in the chromosome 16 VKORC1 genomic region has contributed to the variability of anticoagulant response in humans. PLoS One. 2012;7(12):e53049. doi: 10.1371/journal.pone.0053049.

OPEN ACCESS Freely available online

PLOS ONE

Positive Selection in the Chromosome 16 *VKORC1* Genomic Region Has Contributed to the Variability of Anticoagulant Response in Humans

Blandine Patillon[1,2]*❡, Pierre Luisi[3]❡, Hélène Blanché[4], Etienne Patin[5], Howard M. Cann[4], Emmanuelle Génin[1]¶, Audrey Sabbagh[6]¶

[1] Inserm UMRS-946, Genetic Variability and Human Diseases, Institut Universitaire d'Hématologie, Université Paris Diderot, Paris, France, [2] Université Paris Sud, Kremlin-Bicêtre, France, [3] Institute of Evolutionary Biology, CEXS-UPF-PRBB, Catalonia, Barcelona, Spain, [4] Fondation Jean-Dausset-CEPH, Paris, France, [5] Human Evolutionary Genetics, CNRS URA3012, Institut Pasteur, Paris, France, [6] UMR IRD 216, Université Paris Descartes, Paris, France

Abstract

VKORC1 (vitamin K epoxide reductase complex subunit 1, 16p11.2) is the main genetic determinant of human response to oral anticoagulants of antivitamin K type (AVK). This gene was recently suggested to be a putative target of positive selection in East Asian populations. In this study, we genotyped the HGDP-CEPH Panel for six *VKORC1* SNPs and downloaded chromosome 16 genotypes from the HGDP-CEPH database in order to characterize the geographic distribution of footprints of positive selection within and around this locus. A unique *VKORC1* haplotype carrying the promoter mutation associated with AVK sensitivity showed especially high frequencies in all the 17 HGDP-CEPH East Asian population samples. *VKORC1* and 24 neighboring genes were found to lie in a 505 kb region of strong linkage disequilibrium in these populations. Patterns of allele frequency differentiation and haplotype structure suggest that this genomic region has been submitted to a near complete selective sweep in all East Asian populations and only in this geographic area. The most extreme scores of the different selection tests are found within a smaller 45 kb region that contains *VKORC1* and three other genes (*BCKDK, MYST1 (KAT8), and PRSS8)* with different functions. Because of the strong linkage disequilibrium, it is not possible to determine if *VKORC1* or one of the three other genes is the target of this strong positive selection that could explain present-day differences among human populations in AVK dose requirement. Our results show that the extended region surrounding a presumable single target of positive selection should be analyzed for genetic variation in a wide range of genetically diverse populations in order to account for other neighboring and confounding selective events and the hitchhiking effect.

Citation: Patillon B, Luisi P, Blanché H, Patin E, Cann HM, et al. (2012) Positive Selection in the Chromosome 16 *VKORC1* Genomic Region Has Contributed to the Variability of Anticoagulant Response in Humans. PLoS ONE 7(12): e53049. doi:10.1371/journal.pone.0053049

Editor: Yury E. Khudyakov, Centers for Disease Control and Prevention, United States of America

Received September 20, 2012; **Accepted** November 23, 2012; **Published** December 28, 2012

Funding: This work was supported by the Spanish National Institute for Bioinformatics (www.inab.org). PL is supported by a PhD fellowship from "Acción Estratégica de Salud, en el Marco del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2008–2011" from Instituto de Salud Carlos III. BP is supported by a PhD fellowship from the doctoral program in Public Health (ED420: www.ed-sante-publique.u-psud.fr) from Paris Sud University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: blandine.patillon@inserm.fr

❡ These authors contributed equally to this work.

¶ These authors also contributed equally to this work.

Introduction

Oral anticoagulants of antivitamin K type (AVK) – such as warfarin and acenocoumarol – are widely prescribed drugs for the prevention and treatment of arterial and venous thromboembolic disorders [1,2]. They exert their anticoagulant effect by inhibiting the vitamin K 2,3-epoxide reductase complex 1 (VKORC1). Besides well-known physiopathological and environmental factors, including age, sex, body mass index, disease states, co-medications and diet, genetic factors have been identified as major determinants of AVK dose variability [3]. Candidate-gene and genome-wide association studies have identified four main genes – *CYP2C9*, *CYP4F2*, *CYP2C18* and *VKORC1*– which explain together between 28.2% and 43.5% of the AVK dose variance [3,4,5,6,7]. *CYP2C9*, *CYP4F2* and *CYP2C18* encode proteins involved in the hepatic metabolism of AVK [8,9,10]. *VKORC1* encodes the VKORC1 enzyme, which is the direct pharmacologic target of AVK [11,12]. Differences in the worldwide distribution of the most important polymorphisms influencing AVK dosing are likely to underlie the wide interethnic variability in AVK dose requirements: current population-based trends in warfarin dosing, as reported by the International Warfarin Pharmacogenetics Consortium, indicate a mean weekly dose of 21 mg in Asians, 31.5 mg in Europeans and 40 mg in individuals of African ancestry [13].

Recently, Ross *et al.* [14] documented the distribution of four functional variants located in the three main genes known to influence AVK dose requirement – rs9923231 (*VKORC1*), rs1799853 and rs1057910 (*CYP2C9*), and rs2108622 (*CYP4F2*) – in a large set of samples from the Human Genome Diversity Project - Centre d'Etude du Polymorphisme Humain (HGDP-

CEPH) Panel, representing 52 world populations [15]. They observed a pattern of genetic differentiation among human populations for the *VKORC1* single nucleotide polymorphism (SNP) rs9923231. They applied three formal tests of positive selection to the *VKORC1* gene − the locus-specific branch length (LSBL) test [16], the log of the ratio of heterozygosities (ln*RH*) test [17], and Tajima's *D* [18] − using genome-wide data available for the West African, European and East Asian HapMap samples [19]. The tests yielded significant results in the East Asian sample. Interestingly, the rs9923231 SNP (g.-1639G>A), which was found to be a putative target of positive selection [14], is the main genetic determinant of AVK dose requirement and can alone explain between 25% to 30% of the dose variance among patients [4,5,6,7]. This SNP, located in the promoter region, alters a *VKORC1* transcription factor binding site, leading to lower protein expression [20]. By decreasing VKORC1 activity, the derived -1639A allele thus confers an increased AVK sensitivity phenotype and patients carrying one and two -1639A alleles require on average respectively 25% and 50% lower daily warfarin doses than -1639G homozygous carriers to obtain the same anticoagulant effect [21,22]. Understanding the processes of local adaption that may result in high levels of population differentiation and important interethnic differences in the required AVK dose is thus of particular relevance.

During these last few years, newer methods than those proposed by Ross *et al.* have been developed to detect the molecular footprints of positive selection. These methods are particularly well suited to detect classical signatures of selective sweeps, *i.e.* when a new advantageous mutation spreads rapidly to fixation in particular populations (the so-called 'hard sweep' model) [23]. Such a selective sweep occurs too quickly to leave enough time for recombination events to break down the linkage disequilibrium (LD), leading to a similar increase in frequency of alleles at nearby variants. Therefore, the pattern of genetic variation in the genomic region surrounding the selected allele may differ among populations [24], and the selected allele is expected to be carried by a long and frequent haplotype only in those populations that experienced the local adaptive event [25]. Signals of positive selection can thus be detected by looking for an increased genetic differentiation among populations (using methods such as F_{ST} [26] and the Cross-Population Composite Likelihood Ratio (XP-CLR) test [24]), and an extended haplotype homozygosity (EHH) at the putatively selected locus (using methods such as the Cross-Population Extended Haplotype Homozygosity (XP-EHH) test [27] and the integrated Haplotype Score (iHS) [28]). These methods have proved to be powerful and largely complementary to detect and localize a selective sweep, and are more robust to ascertainment bias in SNP discovery than methods based on the allele frequency spectrum such as the Tajima's *D* used by Ross *et al.* [14,29].

In this study, we investigated whether and how positive selection has acted on the *VKORC1* gene locus using these complementary analytic methods. Our first objective was to determine (1) if the selective sweep is restricted to East Asia or if it is detected in other geographic regions, in particular Central South Asia and America, which are geographically close to East Asia, and (2) if it occurred in all East Asian populations or only in a few of them. Thus, we genotyped six *VKORC1* SNPs in the HGDP-CEPH Panel [30] which covers a much wider range of world populations – including 17 populations from East Asia – than the HapMap Panel in which positive selection at the *VKORC1* locus was initially evidenced. Furthermore, by expanding the analysis to a 2 Mb region encompassing the *VKORC1* gene, we sought to determine if the selective sweep identified around *VKORC1* was due to positive selection directly acting on this gene, or if it was caused by positive selection at a nearby linked gene resulting in genetic hitchhiking [23]. Finally, we discuss combining different methods for uncovering distinct selection signatures, in order to both increase power to detect a selective signal and precisely define its genomic location. We address the difficulty, even with such detailed analyses, in identifying the specific target of selection.

Results

VKORC1 Haplotype Study

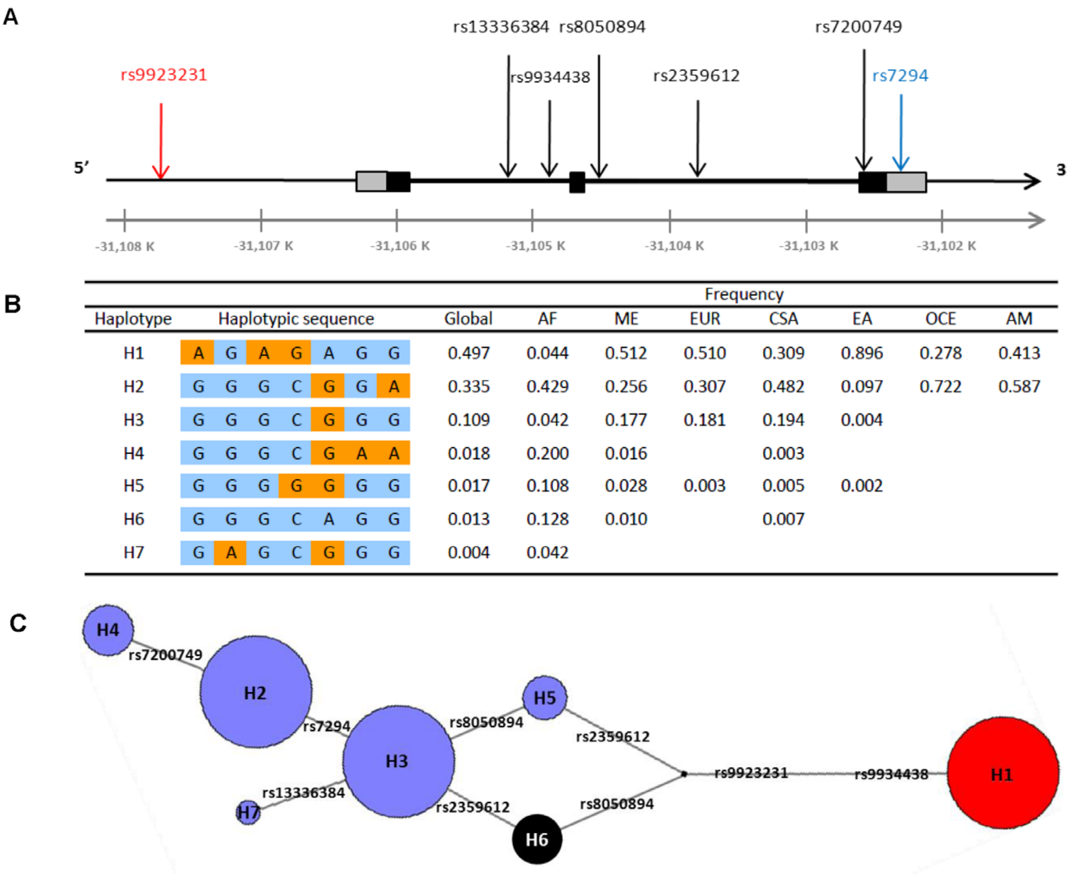
A haplotype study of the 4.1 kb *VKORC1* gene was carried out with seven *VKORC1* SNPs genotyped in the 52 HGDP-CEPH population samples (Figure 1A). Haplotypes were reconstructed from these SNPs. Seven of these haplotypes had a frequency above 1% in at least one geographic region and were labeled H1 to H7 according to their frequency at the global level (Figure 1B). Four haplotypes are found in at least five geographic regions and only two are shared among all regions. The highest and lowest haplotype diversity values are observed in Sub-Saharan Africa (0.75±0.02) and East Asia (0.19±0.02), respectively. Most individuals carrying the ancestral haplotype (H6), *i.e.* the haplotype carrying the ancestral allele at each SNP, are from Sub-Saharan Africa (Figure 1B and Figure S1). Interestingly, the -1639A allele (rs9923231) conferring the increased sensitivity to AVK is carried by a unique haplotype (H1). This haplotype associated with AVK sensitivity is the most frequent at the worldwide level (49.7%) and shows an extremely high differentiation among geographic regions (Figure 1B). While rare in Sub-Saharan Africa (4.4%), it is found at intermediate frequencies in the Middle East, Europe, Central South Asia, Oceania and America (from 27.8% to 51.2%), and is largely predominant in East Asia (89.6%). The prevalence of H1 tends to be high in all of the 17 East Asian population samples investigated, ranging from 75% in She to 100% in Oroqen (Figure S1). However, the sample size is small for most of them, with 10 or less individuals.

The median-joining haplotype network describes the mutational relationships between the different *VKORC1* haplotypes inferred (Figure 1C). Haplotype H1 differs from the others by two nucleotide substitutions at the functional rs9923231 SNP and at the rs9934438 SNP, which are found in complete LD in all geographic regions ($D' = 1$ and $r^2 = 1$, Figure S2).

Detection of Signatures of Positive Selection

To support the hypothesis that positive selection has played a role in shaping patterns of genetic variation at *VKORC1*, four complementary methods were applied to detect signatures of selective sweeps in the genome. F_{ST} and XP-CLR are both based on allele frequency differentiation, whereas XP-EHH and iHS are based on haplotype structure. Scores for the four test statistics were computed at both the regional and population levels for the seven *VKORC1* SNPs and for some other available SNPs [15] representing the expected neutral genomic background. For each score, a *p*-value was derived from the empirical distribution obtained from the genomic background (*cf.* Material and Methods). We considered as significant any *p*-value below 0.05. The results of the four tests are presented in Table 1 and Table 2.

At the global level, when we evaluated the level of genetic differentiation among the seven HGDP-CEPH Panel geographic regions, an atypical pattern of genetic differentiation was detected for four *VKORC1* SNPs: rs2359612, rs8050894, rs9934438 and rs9923231 ($p<0.05$). The functional rs9923231 polymorphism and the rs9934438 SNP, in complete LD with each other, displayed F_{ST} values falling above the 99th percentile of the empirical



Haplotype	Haplotypic sequence	Global	AF	ME	EUR	CSA	EA	OCE	AM
		Frequency							
H1	A G A G A G G	0.497	0.044	0.512	0.510	0.309	0.896	0.278	0.413
H2	G G G C G G A	0.335	0.429	0.256	0.307	0.482	0.097	0.722	0.587
H3	G G G C G G G	0.109	0.042	0.177	0.181	0.194	0.004		
H4	G G G C G A A	0.018	0.200	0.016		0.003			
H5	G G G G G G G	0.017	0.108	0.028	0.003	0.005	0.002		
H6	G G G C A G G	0.013	0.128	0.010		0.007			
H7	G A G C G G G	0.004	0.042						

Figure 1. Results of the *VKORC1* haplotype study. (**A**) **Position of the seven SNPs along the *VKORC1* gene.** *VKORC1* is a 4.1 kb gene (GenBank accession number AY587020) located at 16p11.2. The three exons of the gene are represented as boxes, with 5'UTR and 3'UTR regions colored in grey and coding regions in black. Flanking and intronic regions are represented as thin and thick lines, respectively. The seven studied SNPs are shown in their sequential order along the *VKORC1* gene. The functional polymorphism rs9923231 located in the promoter, is highlighted in red and the SNP already present in the Illumina 650K chip in blue. Physical position along chromosome 16 is indicated in kb below. (**B**) **Distribution of *VKORC1* haplotypes at the global and regional level.** For each haplotype, SNPs are listed in the same sequential order than in Figure 1A. Ancestral and derived alleles are shown in blue and orange, respectively. Haplotype labels H1 to H7 were given according to the global haplotype frequency. AF, sub-Saharan Africa; ME, Middle East; EUR, Europe; CSA, Central South Asia; EA, East Asia; OCE, Oceania; AM, America. (**C**) **Median-joining network of the inferred *VKORC1* haplotypes at the global level.** Circles areas are proportional to the global haplotype frequency and branch lengths to the number of mutations separating haplotypes. Labels of haplotypes are indicated in corresponding circles, and labels of mutations on the network branches. The haplotype carrying the -1639A allele conferring the AVK sensitivity phenotype (H1) is shown in red and the ancestral haplotype (H6) in black.
doi:10.1371/journal.pone.0053049.g001

genome-wide distribution ($F_{ST}=0.32$, $p=0.008$) (Figure 2A). When global F_{ST} values were computed among the 52 world populations, very similar results were obtained (Table S1). At the inter-regional level, *i.e.* between a given geographic region and the remaining ones, the same four *VKORC1* SNPs showed highly significant F_{ST} values ($p<0.01$) when comparing Central South Asia and East Asia to the rest of the world (Table 1, Figures 2B and 2C). Regarding East Asia, the highest F_{ST} values ($F_{ST}=0.41$, $p=0.003$) were also observed for the two SNPs, rs9923231 and rs9934438. For the other geographic regions, no *VKORC1* SNP displayed an inter-regional F_{ST} value as much significant as the ones observed for Central South Asia and East Asia (Table 1 and Figure S3). At the intra-regional level, *i.e.* among populations within a region, no extreme pattern of genetic differentiation ($p<0.01$) was observed for any *VKORC1* SNP in any geographic region (Table 1 and Figure S4).

The XP-CLR test applied to each geographic region also provided evidence of an atypical pattern of genetic differentiation at the *VKORC1* gene locus, with XP-CLR scores in East Asia ranging from 16.53 ($p=0.050$) to 43.44 ($p=0.012$) in the 16 kb genomic region centered on *VKORC1* (Table 2). For each of the other six geographic regions, the XP-CLR scores were very low, supporting the existence of a selective sweep restricted to East Asia. In this geographic region, when the XP-CLR test was

Table 1. Results of the inter-regional F_{ST}, intra-regional F_{ST}, XP-EHH and iHS tests in the seven geographic regions.

Region	SNP	DAF[a]	Inter-regional F_{ST}[b]	Inter-regional F_{ST} *p*-value[c]	Intra-regional F_{ST}[d]	Intra- regional F_{ST} *p*-value[c]	XP-EHH score	XP-EHH *p*-value[e]	iHS score	iHS *p*-value[e]
Africa	rs7294	0.63	0.18	0.215	0.13	0.074	−0.94	0.833	−1.24	0.183
	rs7200749	0.20	0.48	0.217	0.02	0.643	−1.50	0.923	−0.31	0.748
	rs2359612	0.82	0.23	0.173	0.09	0.123	−1.15	0.875	−0.96	0.305
	rs8050894	0.16	0.25	0.143	0.09	0.125	−1.14	0.872	0.11	0.909
	rs9934438	0.04	0.36	0.029 *	0.10	0.058	−1.05	0.855	−0.03	0.974
	rs13336384	0.04	0.16	0.329	0.02	0.448	−1.06	0.858	0.14	0.883
	rs9923231	0.04	0.36	0.029 *	0.10	0.058	−1.01	0.845	−0.01	0.989
Middle East	rs7294	0.27	0.02	0.411	0.00	0.906	0.94	0.171	1.55	0.103
	rs7200749	0.02	0.00	0.670	0.02	0.310	1.58	0.069	0.65	0.492
	rs2359612	0.48	0.00	1.000	0.006	0.595	1.17	0.127	2.69	0.009 **
	rs8050894	0.54	0.00	0.849	0.002	0.667	1.15	0.132	−1.40	0.141
	rs9934438	0.51	0.00	0.946	0.01	0.481	1.05	0.149	−1.76	0.066
	rs13336384	0.00	0.005	0.044 *	0.00	1.000	1.07	0.145	NA	NA
	rs9923231	0.51	0.00	0.946	0.01	0.481	1.01	0.156	−1.76	0.066
Europe	rs7294	0.30	0.005	0.670	0.004	0.570	0.94	0.167	0.67	0.474
	rs7200749	0.00	0.02	0.477	0.00	1.000	1.50	0.077	NA	NA
	rs2359612	0.49	0.00	1.000	0.02	0.304	1.15	0.125	2.00	0.039 *
	rs8050894	0.51	0.00	1.000	0.02	0.286	1.14	0.128	−0.98	0.298
	rs9934438	0.51	0.00	0.993	0.02	0.304	1.05	0.145	−1.02	0.281
	rs13336384	0.00	0.005	0.071	0.00	1.000	1.06	0.142	NA	NA
	rs9923231	0.51	0.00	0.993	0.02	0.304	1.01	0.155	−1.02	0.281
Central South Asia	rs7294	0.49	0.06	0.042 *	0.07	0.026*	0.62	0.260	−0.54	0.550
	rs7200749	0.003	0.02	0.261	0.02	0.041*	1.25	0.116	NA	NA
	rs2359612	0.69	0.12	0.002 **	0.07	0.025*	0.89	0.187	1.00	0.280
	rs8050894	0.32	0.12	0.003 **	0.08	0.015*	0.87	0.191	−0.13	0.893
	rs9934438	0.31	0.10	0.006 **	0.07	0.020*	0.78	0.215	−0.20	0.834
	rs13336384	0.00	0.005	0.088	0.00	1.000	0.79	0.210	NA	NA
	rs9923231	0.31	0.10	0.006 **	0.07	0.020*	0.73	0.227	−0.20	0.834
East Asia	rs7294	0.10	0.21	0.063	0.02	0.311	2.68	0.011 *	1.99	0.040 *
	rs7200749	0.00	0.02	0.576	0.00	1.000	3.10	0.005 **	NA	NA
	rs2359612	0.10	0.39	0.005 **	0.02	0.317	2.89	0.008 **	1.92	0.047 *
	rs8050894	0.90	0.38	0.005 **	0.02	0.331	2.88	0.008 **	−1.20	0.200
	rs9934438	0.90	0.41	0.003 **	0.02	0.300	2.81	0.009 **	−1.27	0.174
	rs13336384	0.00	0.005	0.252	0.00	1.000	2.81	0.009 **	NA	NA
	rs9923231	0.90	0.41	0.003 **	0.02	0.300	2.773	0.010 *	−1.274	0.174

Table 1. Cont.

Region	SNP	DAF[a]	Inter-regional F_{ST}[b]	Inter-regional F_{ST} *p*-value[c]	Intra-regional F_{ST}[d]	Intra- regional F_{ST} *p*-value[c]	XP-EHH score	XP-EHH *p*-value[e]	iHS score	iHS *p*-value[e]
Oceania	rs7294	0.72	0.23	0.090	0.00	0.771	0.03	0.456	0.05	0.961
	rs7200749	0.00	0.005	0.401	0.00	1.000	0.51	0.285	NA	NA
	rs2359612	0.72	0.09	0.404	0.00	0.771	0.32	0.346	0.05	0.961
	rs8050894	0.25	0.12	0.327	0.02	0.530	0.29	0.355	0.50	0.584
	rs9934438	0.28	0.08	0.438	0.00	0.749	0.20	0.388	0.50	0.584
	rs13336384	0.00	0.009	0.014 *	0.00	1.000	0.21	0.384	NA	NA
	rs9923231	0.28	0.08	0.438	0.00	0.749	0.16	0.404	0.50	0.584
America	rs7294	0.58	0.11	0.320	0.17	0.195	0.76	0.207	NA	NA
	rs7200749	0.00	0.01	0.553	0.00	1.000	1.15	0.125	NA	NA
	rs2359612	0.59	0.02	0.674	0.17	0.190	0.96	0.162	NA	NA
	rs8050894	0.41	0.02	0.667	0.17	0.190	0.95	0.163	NA	NA
	rs9934438	0.41	0.01	0.743	0.17	0.190	0.88	0.178	NA	NA
	rs13336384	0.00	0.006	0.013 *	0.00	1.000	0.89	0.176	NA	NA
	rs9923231	0.41	0.01	0.743	0.17	0.190	0.85	0.185	NA	NA

[a] Derived allele frequency estimated at the global level.
[b] F_{ST} estimated at the inter-regional level, *i.e.* between a given geographic region and the remaining ones.
[c] *P*-values are derived from the genome-wide empirical distribution of F_{ST} values.
[d] F_{ST} estimated at the intra-regional level, *i.e.* among populations within a region.
[e] *P*-values are derived from the empirical distribution of the iHS and XP-EHH scores along the chromosome 16.
*$p<0.05$; ** $p<0.01$; *** $p<0.005$.
NA: Not Applicable (for iHS: when a gap>200 kb between successive SNPs is found in the region in the region delimited by the SNPs where the EHH value drops below 0.05 around the core SNP).
doi:10.1371/journal.pone.0053049.t001

Table 2. Results of the XP-CLR test in a 16 kb region centered on *VKORC1* in the seven geographic regions.

Region	Physical position	XP-CLR score	XP-CLR *p*-value[a]
Africa	31005354	0.00	1.000
	31009354	0.00	1.000
	31013354	0.96	0.289
	31017354	0.58	0.348
	31021354	0.09	0.470
Middle East	31005354	4.00	0.138
	31009354	0.85	0.306
	31013354	3.28	0.158
	31017354	0.27	0.403
	31021354	6.25	0.092
Europe	31005354	0.54	0.351
	31009354	0.00	1.000
	31013354	2.63	0.186
	31017354	0.15	0.427
	31021354	2.40	0.198
Central South Asia	31005354	0.03	0.464
	31009354	0.00	1.000
	31013354	0.00	1.000
	31017354	0.01	0.476
	31021354	0.00	0.490
East Asia	31005354	24.08	0.032 *
	31009354	16.53	0.050 *
	31013354	30.49	0.023 *
	31017354	26.82	0.028 *
	31021354	43.44	0.012 *
Oceania	31005263	0.00	1.000
	31009263	0.00	1.000
	31013263	0.00	1.000
	31017263	0.00	1.000
	31021263	0.00	1.000
America	31005354	0.00	1.000
	31009354	0.00	1.000
	31013354	0.00	1.000
	31017354	0.01	0.587
	31021354	0.00	0.597

[a] *P*-values are derived from the empirical distribution of the XP-CLR scores along the chromosome 16.

*$p<0.05$; ** $p<0.01$; *** $p<0.005$.

doi:10.1371/journal.pone.0053049.t002

performed for each population, all of the 17 population samples, except Oroqen, showed this extreme pattern of genetic differentiation, with at least three significant XP-CLR scores out of the five scores computed in the 16 kb genomic region surrounding *VKORC1* (Table S2). As most of the SNPs in the *VKORC1* genomic region have reached fixation in the Oroqen sample, XP-CLR scores could be calculated for only very few SNPs on either side of *VKORC1*, making difficult the interpretation of XP-CLR results in this sample.

Regional results obtained with the extended haplotype-based XP-EHH test indicated that the unusual pattern of genetic differentiation observed at the *VKORC1* gene locus resulted from a selective sweep in East Asia. Significant XP-EHH scores, ranging from 2.68 ($p = 0.011$) to 3.10 ($p = 0.005$), were observed for the seven *VKORC1* SNPs in East Asia, while no significant values were observed for any other geographic region (Table 1). For East Asian populations, evidence for a selective sweep was detected in all 17 population samples with significant XP-EHH scores for each of the seven *VKORC1* SNPs, ranging from 1.84 ($p = 0.$ 049) in the Dai sample for rs7294, to 3.78 ($p = 0.004$) in the Tujia sample for rs8050894 (Table S3).

With the iHS test, only two *VKORC1* SNPs (rs7294 and rs2359612) exhibited significant iHS scores in East Asia ($p = 0.040$ and 0.047, respectively; Table 1). Two other significant scores were observed for the rs2359612 SNP in the Middle East (2.69, $p = 0.009$) and Europe (2.00, $p = 0.039$). At the population level in East Asia, only three samples (Hezhen, Lahu, and Yakut) displayed significant iHS scores for two, three and four SNPs, respectively (Table S3).

The four selection tests consistently evidenced the signature of a selective sweep involving the *VKORC1* genomic region in East Asia. However, this result did not allow us to determine with certainty that *VKORC1* is the direct target of positive selection. A linked gene could be the target instead, resulting in genetic hitchhiking of *VKORC1* [23]. In an attempt to seek the true target of positive selection, we probed the downloaded chromosome 16 genotypes [15] with the four tests for selection and examined the results over an extended 2 Mb genomic region centered on *VKORC1*. We focused on clusters of selection test scores with highly significant *p*-values ($p<0.01$) for East Asia only. Three clusters were observed (Figure 3): (i) ~ 570 kb downstream of *VKORC1*, the first cluster was found with partially overlapping clusters of extreme XP-CLR and XP-EHH scores over a region of 64 and 39 kb, respectively, involving the genes *ITGAL*, *ZNF768*, and *ZNF747*; (ii) at or close to *VKORC1* genomic position, the second cluster was determined by overlapping clusters of extreme F_{ST} values when comparing East Asia to the rest of the world (with the lowest *p*-values observed for the same two *VKORC1* SNPs evidenced before, rs9923231 and rs9934438) and extreme XP-CLR and XP-EHH scores. These clusters ranged in size from 45 to 244 kb; (iii) ~ 230 kb upstream of *VKORC1*, the third cluster of 32 kb was found with XP-EHH and concerned the genes *ITGAM* and *ITGAX*. If SNPs within clusters are in high LD ($D' \geq 0.97$, except for one SNP in the third cluster), only limited LD exists between the SNPs located in the different clusters (Figure 4 and Figure S5) and several recombination hotspots are present between these clusters (Figure 4). This suggests that each of the three clusters represents a different adaptive event.

Examination of the second cluster showed that *VKORC1* is contained in a block of strong LD spanning ~ 505 kb in East Asia (Figure 4 and Figure S5). Similar LD blocks were observed for Central South Asia and Europe, and to a lesser extent, for the Middle East (Figure S5). This LD block encompasses 25 genes (Figure 4). We used the most extreme F_{ST}, XP-CLR and XP-EHH scores in order to spatially localize a target of selection within the LD block. Significant XP-CLR scores ($p<0.05$) were found in a 350 kb region encompassing 19 genes including *VKORC1* (Table S4). XP-EHH scores were almost all significant at the 0.05 threshold but four adjacent genes *VKORC1*, *BCKDK*, *MYST1* (*KAT8*) and *PRSS8* displayed most extreme XP-EHH scores ($p<0.01$). Clusters of highly significant F_{ST} values when comparing East Asia to the rest of the world ($p<0.01$) and significant global F_{ST} values ($p<0.05$) were also found for these four genes (Table S5). It is thus probable that the selective pressure has targeted one of these genes.

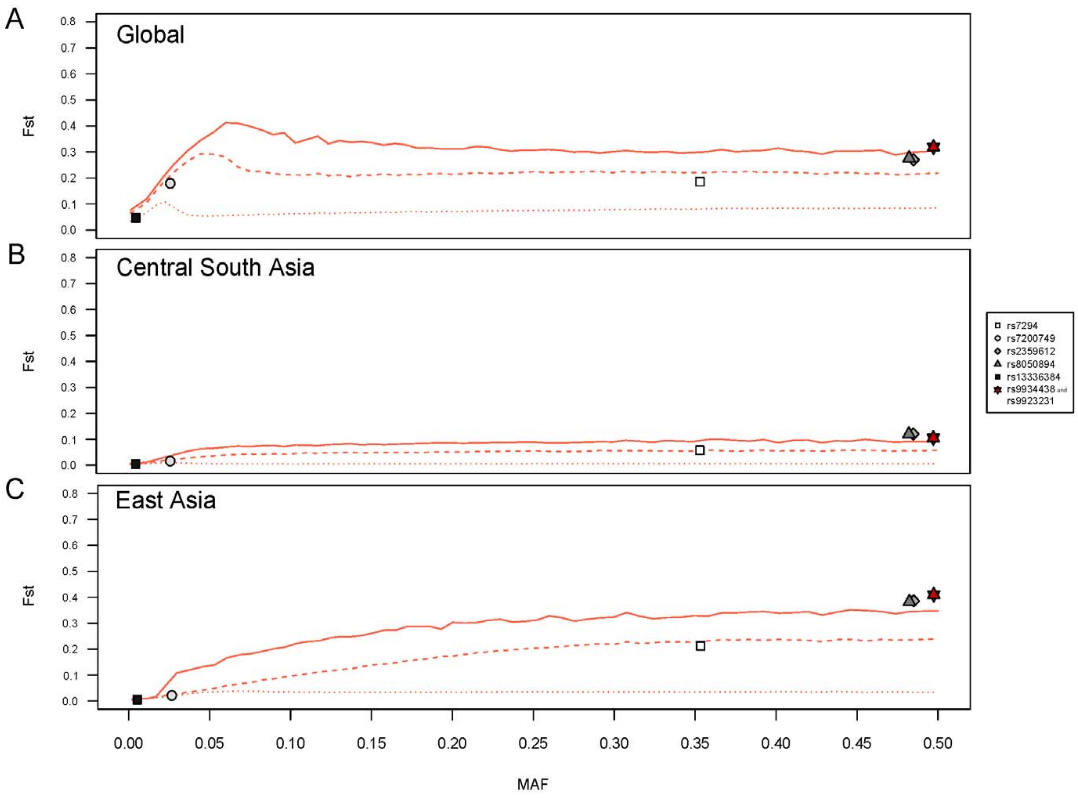


Figure 2. Atypical patterns of genetic differentiation observed for *VKORC1* SNPs. Genome-wide empirical distributions of F_{ST} values were constructed from 644,143 SNPs having a MAF ≥0.001 at the global level. Individual values of F_{ST} calculated for each of the seven *VKORC1* SNPs are plotted against their global MAF. The functional rs9923231 SNP is shown in red. The 50th, 95th and 99th percentiles are indicated as dotted, dashed and full red lines, respectively.
doi:10.1371/journal.pone.0053049.g002

When did the -1639A VKORC1 Allele begin to Increase in East Asia?

The time at which the frequency of the -1639A allele started to increase in East Asia was estimated by using a maximum-likelihood method [31] with the 17 East Asian HGDP-CEPH sample data. Our analysis yielded an age estimate of 181 generations (95% CI: 128–256 generations). Assuming a generation time of 25 years, the expansion therefore occurred about 4,525 years ago (95% CI: 3,200–6,400 years).

Discussion

Numerous genes involved in absorption, distribution, metabolism and excretion (ADME) of drugs, exhibit evidence of recent positive selection and/or high population differentiation levels [32]. However, there are fewer examples of the action of natural selection on genes involved in the pharmacodynamics of drugs, such as *VKORC1*. Although numerous surveys have examined the genetic polymorphism of *VKORC1* in samples from diverse ethnic origins [13,20,33,34,35,36,37], these studies provided an incomplete picture of haplotype diversity because different sets of SNPs were used and worldwide coverage was incomplete. In this study, we took advantage of the worldwide coverage of the HGDP-CEPH Panel to provide the first detailed analysis of *VKORC1* population diversity using the same set of SNPs. Haplotype analysis revealed that the -1639A derived allele that confers AVK sensitivity is carried by a unique haplotype in all 52 population samples investigated. This haplotype associated with AVK sensitivity is predominant in East Asia, rare in Sub-Saharan Africa and occurs at intermediate frequencies in other geographic regions. Because it is found in Sub-Saharan Africa and other world populations, this haplotype is probably rather old. Its geographic distribution leads to striking differences between East Asian and non East Asian samples for genetic susceptibility to AVK sensitivity.

One explanation for worldwide diversity of this haplotype could be positive selection. This hypothesis was supported by five genome-wide scans that found atypical patterns of the allele frequency spectrum [38], extended LD [39,40], and unusual genetic differentiation [40,41,42] in a 450 kb genomic region encompassing *VKORC1*. When specified, the target population was Asian [38,40]. Ross *et al.* [14] found evidence of positive selection at *VKORC1* in the East Asian HapMap sample, based on the level of genetic diversity (ln*RH* test [17]), genetic differentiation (LSBL test [16]) and allele frequency spectrum (Tajima’s *D* [18]).

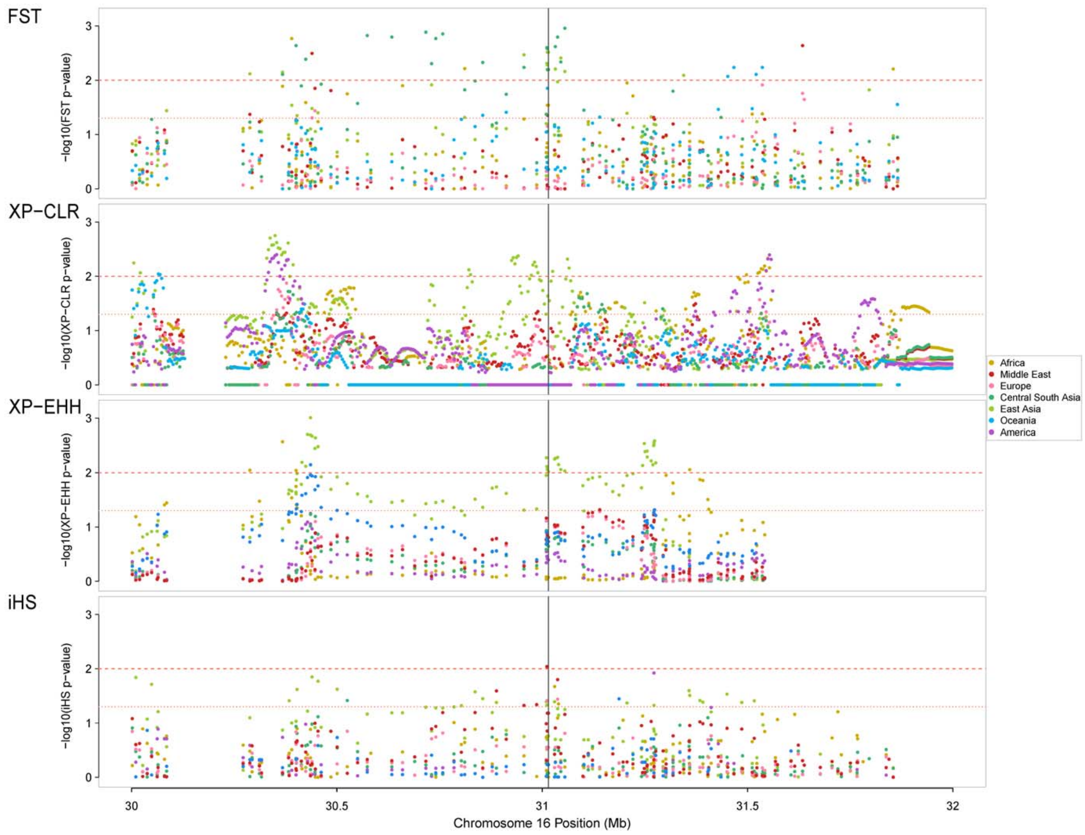


Figure 3. Distribution of –log$_{10}$ (*p*-values) for four selection tests across a 2 Mb region centered on *VKORC1*. A black vertical line indicates the physical position of *VKORC1* on chromosome 16. Horizontal red dotted and dashed lines show 0.05 and 0.01 chromosome-wide significance levels, respectively. The selection tests (inter-regional F_{ST}, XP-CLR, XP-EHH and iHS, respectively) were separately applied in each of the seven geographic regions.
doi:10.1371/journal.pone.0053049.g003

In this study, we provided compelling evidence of positive selection at the *VKORC1* gene locus in East Asia and only in this geographic region. A footprint of natural selection was found in each of the widely distributed 17 HGDP-CEPH East Asian population samples. By using four different tests of positive selection and by assessing significance at a given locus on the basis of an empirical distribution derived from the genomic background, we believe we can be confident that positive selection, rather than demographic forces, accounts for the data presented here. Indeed, it is well known that large allele frequency differences between populations are not infallible proofs of positive selection: these can also result from genetic drift, migration and other neutral demographic processes [43,44]. This might be the explanation for the significant inter-regional F_{ST} values observed in Central South Asia (Table 1 and Figure 2B).

Because the XP-EHH test is designed to detect fixation events that are relatively young (~ 30,000 years) [27], the selective event we have detected is likely to be rather recent. This is indeed supported by an age estimate of 4,525 years (95% CI: 3,200–6,400 years) for the time at which the *VKORC1* -1639A allele started to increase in frequency in East Asia. The poor performance of the iHS test that detected only very few signals of positive selection in this study could have been predicted since its power to detect selective sweeps involving alleles near fixation is known to be low [28,45]. By contrast, XP-EHH and XP-CLR perform better when the allele targeted by selection is near fixation and indeed showed strong evidence of a selective sweep in this study [24,27].

In an attempt to determine if the *VKORC1* gene has been the direct target of positive selection or if it reflects genetic hitchhiking [23], we extended our analysis to a 2 Mb region surrounding the *VKORC1* gene (Figure 3). Apart from the highly significant footprint of positive selection localized in the *VKORC1* region, two other significant signals, at ~ 570 kb downstream and ~ 230 kb upstream of *VKORC1*, were detected with XP-CLR and/or XP-EHH in East Asia. These two regions contains genes that belong to the same integrin family – specifically to the CD11 gene cluster: *ITGAL* downstream, and adjoining genes *ITGAM* and *ITGAX* upstream – involved in immune functions and being thus good candidates for positive selection [46,47,48]. However, since SNPs located in these integrin genes show limited LD with those of *VKORC1*, a single adaptive event is unlikely. Apart from East Asia, the *ITGAL* region showed signals of positive selection in other

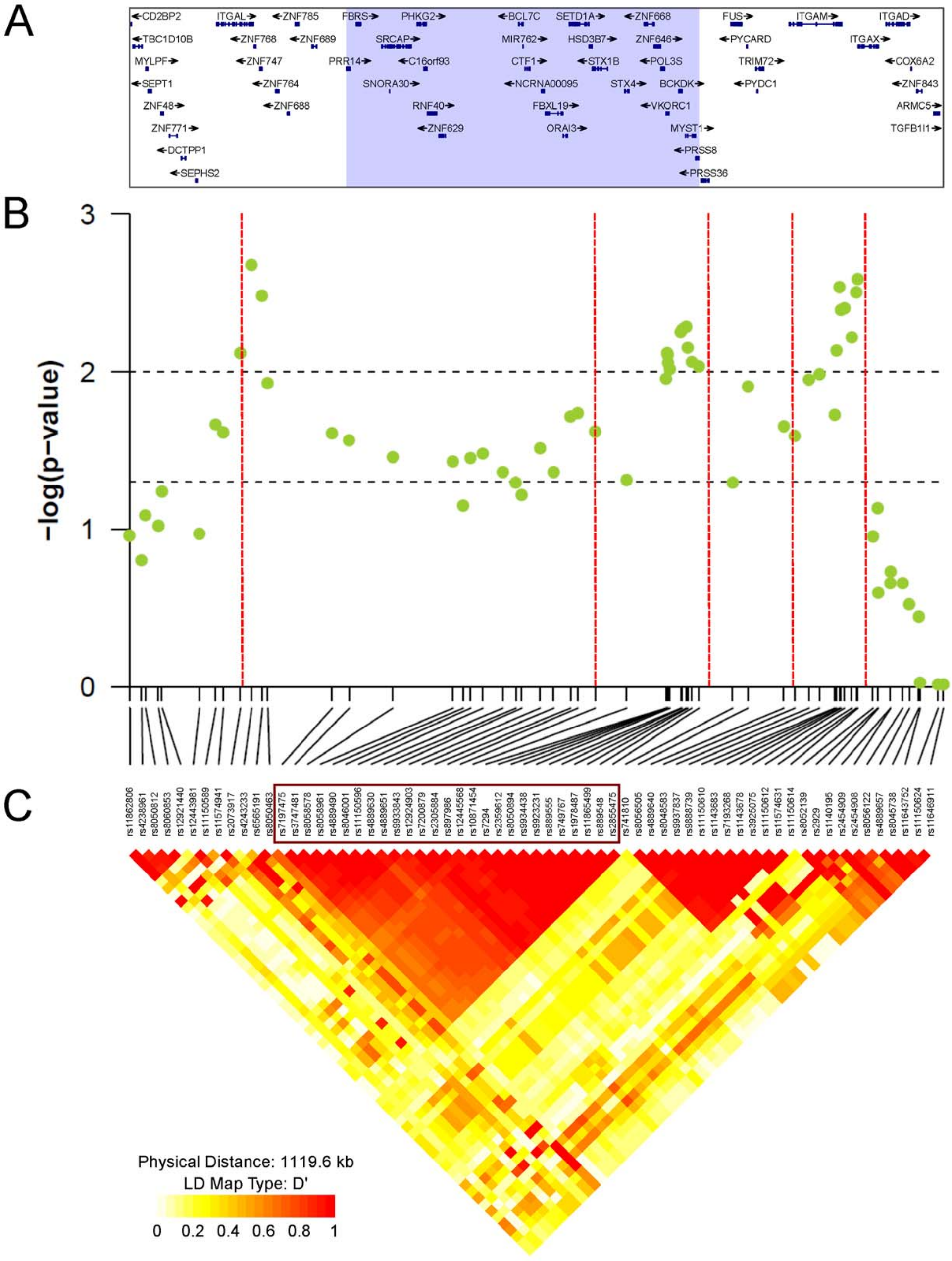


Figure 4. Detailed analysis of a 1.1 Mb genomic region surrounding the *VKORC1* gene locus in East Asia. The boundaries of the region displayed (chr16:30,271,572-31,391,123; UCSC human genome build hg18) were chosen so as to include the three clusters of significant scores detected in East Asia by the selection tests in the 2 Mb region centered on *VKORC1* (Figure 3). (**A**) **Name and location of genes.** Exons are displayed as blue boxes and the transcribed strand is indicated with an arrow. Genes located in the block of strong LD encompassing *VKORC1* and including the SNPs in the red box shown in Figure 4C, are highlighted in the grey area. (**B**) **XP-EHH results in East Asia.** The significance of the XP-EHH scores ($-\log_{10}$ empirical *p*-value) are shown for individual SNPs with a MAF ≥0.01 in East Asia. Horizontal dashed lines indicate 0.05 and 0.01 chromosome-wide significance levels. Recombination hotspots detected in HapMap Phase II data are indicated by red vertical dotted lines. The data

and methods used to derive these hotspots are available from the HapMap website (http://www.hapmap.org/) [83,84]. (**C**) **LD plot.** Pairwise LD values, depicted as *D′*, are shown for SNPs with a MAF ≥0.01 in East Asia. *D′* values are displayed in different colors from yellow to red for *D′* = 0 to *D′* = 1, respectively. The red box highlights SNPs included in the LD block encompassing *VKORC1*. The plot was produced using the snp.plotter R package [74].
doi:10.1371/journal.pone.0053049.g004

geographic regions (America with XP-CLR, and Sub-Saharan Africa and Oceania with XP-EHH), arguing for a different evolutionary history from that of *VKORC1*, which was only found in East Asia. This observation emphasizes the need for studying the geographic distribution of a selective event in a wide range of genetically diverse populations, as per Scheinfeldt *et al.* [49] who, after performing a detailed analysis of a 3 Mb region surrounding a gene showing strong footprints of positive selection, discovered patterns of genetic variation consistent with the presence of a cluster of three independent selective events occurring in different populations. By extending their analysis to the entire genome, they identified several other genomic regions exhibiting evidence for the presence of multiple and independent selective targets, suggesting that clusters of adaptive evolution, such as the one detected herein, are widespread in the human genome.

After delimitating the selective signal for *VKORC1* by analyzing selective events identified in the 2 Mb region just described, we aimed at precisely mapping the gene targeted by positive selection. *VKORC1* is located in a ~ 505 kb LD block in East Asia containing 25 genes (Figure 4), and the selective pressure could have targeted any gene in this LD block. We used F_{ST}, XP-CLR and XP-EHH scores to spatially localize possible targets of positive selection within the LD region. A block of four adjacent genes – *VKORC1, BCKDK, MYST1,* and *PRSS8*– was found to be the most likely selective target (Table S4).

BCKDK codes for the mitochondrial branched chain ketoacid dehydrogenase kinase. *MYST1* and *PRSS8* are two immunity-related genes, listed as candidates for positive selection in several databases [40,42,50]. If, indeed, one of these three genes is the target of the selective sweep detected here, it should contain a functional variant of high frequency in East Asia and we did not find such a variant in HapMap data.

Assuming that selection has directly targeted the *VKORC1* gene, the advantage would then probably be related to vitamin K metabolism,vitamin K being the only known substrate of VKORC1. This vitamin plays a crucial role in the synthesis of vitamin K-dependent (VKD) proteins, especially blood coagulation factors, which requires VKORC1 activity [51,52]. Large geographic differences in dietary vitamin K intake, especially in vitamin K2, exist between human populations, with the highest plasma levels found in Asian populations, as compared to Europeans and Africans [53,54]. These differences could be explained by the wide consumption of fermented soybean food (*natto*) - a major source of vitamin K2 - in East Asia [55,56]. It is then possible that, at some points in the history of East Asian populations, these high levels of vitamin K intakecould have been deleterious and created a selective pressure against *VKORC1* gene expression and coagulant activity. There is, however, no report so far of a deleterious effect associated with a high consumption of vitamin K and it is more the low dietary vitamin K intake that is problematic, hampering the adequate synthesis of VKD proteins in extrahepatic tissues notably bone and arterial vessels [57]. An alternative hypothesis could be that a naturally occurring environmental molecule of AVK type - such as a coumarin derivative - specifically found in East Asia, exerted a selective pressure on the *VKORC1* gene in populations of this region during their recent history. Such molecules are present in the nature, as illustrated by the example of the sweet clover disease that affected cattle in Canada and North America in the 1920's. Sweet clover hay, used to feed cattle, contains a natural coumarin that is oxidized in mouldy hay to form dicoumarol, a hemorrhagic agent. Its discovery led to the synthesis of coumarin derivatives used in clinical application as oral anticoagulants since the 1940's [58,59]. Evidence of an effect of warfarin in shaping *VKORC1* genetic diversity could be found in rats and mices. Indeed, since the introduction in the 1950's of this molecule as rodenticide, mutations in the *VKORC1* gene conferring warfarin resistance have spread in rodent populations but the mechanisms by which they lead to warfarin resistance are still not elucidated [60,61,62,63].

In conclusion, we found that the *VKORC1* genomic region exhibits diversity patterns consistent with the action of positive selection in East Asia. Nearly complete selective sweeps, such as the one described herein, are believed to be rare in recent human adaptive history [64,65,66,67]. This selective event is probably responsible for the spread of the derived -1639A allele conferring the increased AVK-sensitive phenotype in East Asian populations and contributes to present-day differences among human populations in the genetic sensitivity to AVK. A detailed analysis of the extended *VKORC1* genomic region revealed selective signals at several independent genetic loci, indicating a complex evolutionary history for this chromosome 16 region. Our evolutionary analysis emphasizes the importance of considering the surrounding genomic region of a candidate gene for selection in order to avoid erroneous conclusions about the true target of selection. We show here that the gene targeted by selection could be either *VKORC1* or another gene located in the 45 kb region covered by selective sweep detected in East Asia. Our ability to identify the target of selection may be limited by the number of genetic polymorphisms investigated. Examining the selective signal with more genetic variation using whole-genome sequences from the 1000 Genomes Project [68] may well improve the mapping of the gene targeted by selection. Furthermore, allele frequency spectrum bias tends to be minimized with whole genome sequences, which may allow the use of tests for natural selection based on this spectrum.

Materials and Methods

The HGDP-CEPH Panel

We used the HGDP-CEPH Panel that presently includes 1,064 individuals from 52 populations worldwide [69]. For the analysis presented here, the standardized subset panel H952 containing no first nor second degree relative pairs, was used [70]. This subpanel includes 952 individuals grouped into seven broad geographic regions as defined by Li *et al.* [15]: Sub-Saharan Africa (N = 105), the Middle East and Mozabites from north Africa (N = 163), Europe (N = 158), Central South Asia (N = 202), East Asia (N = 232), Oceania (N = 28) and America (N = 64). A full description of the 52 samples included in the HGDP-CEPH Panel is provided in Table S6.

SNP Genotyping

A total of 940 individuals from the original H952 subpanel were previously genotyped by Li *et al.* [15] with the Illumina HumanHap 650 K platform and their genotypes at 644,258 autosomal SNPs were downloaded from the public HGDP-CEPH database (http://www.cephb.fr/en/hgdp/). Only one SNP (rs7294) from this dataset is located in the *VKORC1* gene. We additionally genotyped six SNPs in *VKORC1* in the 940 individuals, using the TaqMan® SNP Genotyping Assay-by-Design method in 5 µl reaction volumes according to the manufacturer's protocol (Applied Biosystems, Foster City, CA): rs9923231 (g.-1639G>A) located in the promoter region, rs13336384 and rs9934438 in the first intron, rs2359612 and rs8050894 in the second intron, and rs7200749 in the third exon (Figure 1A). Missing genotype rates varied from 0.5% to 2.2% for SNPs rs7200749 and rs9934438, respectively. Since the two SNPs rs9923231 and rs9934438 were found in complete LD in the seven geographic regions (Figure S2), we were able to impute the missing genotypes of a given SNP using available information from the other, leading to a total of 0.96% missing genotypes for these two SNPs. No significant deviations from the Hardy-Weinberg proportions were observed for any *VKORC1* SNP in any of the 52 population samples at the 0.01 significance level (data not shown). Allele frequency distributions of the seven *VKORC1* SNPs in the 52 population samples are shown in Figure S6.

Statistical Analysis

***VKORC1* haplotype study.** To investigate the worldwide diversity of the *VKORC1* gene, we conducted a haplotype study using the seven genotyped SNPs. A total of 931 individuals with less than three missing genotypes were included in the haplotype reconstruction. For each geographic region, haplotype frequencies were estimated with the Bayesian statistical method implemented in Phase v2.1 [71] using defaults parameters. To avoid the convergence of the algorithm to a local maximum, we ran it 10 times with different random seeds and kept the output from the run with the best average value. The worldwide haplotype frequencies were then calculated as the weighted average of the frequencies estimated in each of the seven geographic regions. Similar results were obtained when a single pooled sample of all individuals was considered in the haplotype frequency estimation (data not shown). Since information on ancestral allele state is required to distinguish between ancestral and derived haplotypes, we used the snp131OrthoPt2Pa2Rm2.txt file downloaded from the UCSC genome browser (http://genome.ucsc.edu/) which provides the orthologous alleles in chimpanzee, orangutan and rhesus macaque. For each SNP, the allele shared by the three species was identified as the ancestral allele. Haplotype networks were drawn with the Network v4.5.1.6 software (http://www.fluxus-engineering.com/), using the median-joining algorithm which builds the minimum spanning network from the given haplotypes by favoring short connections [72]. LD analyses were performed with Haploview v4.1 [73] and the snp.plotter R package [74], using Lewontin's disequilibrium coefficient *D*' [75] and the correlation coefficient r^2 [76].

Detection of Signatures of Positive Selection

To explore whether *VKORC1* has evolved under positive selection in humans, we looked for two distinct genetic patterns of a selective sweep that are expected to remain detectable in the genome over different time scales after the action of natural selection: (i) an important genetic differentiation among populations nearby the locus of interest, and (ii) the presence of unusually frequent and long haplotypes in the surrounding genomic region. For each method, we used an outlier approach to calculate the *p*-values of the computed scores. Under this approach, an empirical distribution is constructed using other SNPs in the genome that are assumed to be neutral and to represent the genomic background under neutrality. An empirical *p*-value is computed that corresponds to the proportion of values from the empirical distribution that are higher than the value observed at the locus of interest. If the value obtained for the SNP of interest is greater than the 95th percentile ($p<0.05$) of the empirical distribution, positive selection is invoked. For that purpose, we used the empirical distributions obtained from the scores calculated either on a genome-wide (all autosomal chromosomes) or chromosome-wide (chromosome 16, where *VKORC1* is located) basis.

First, we used two statistics, F_{ST} and XP-CLR, which measure the genetic differentiation among human populations [24,26]. These methods are able to detect selective sweeps that have occurred up to 75,000 years ago [77]. The fixation index F_{ST} [78] quantifies the proportion of genetic variance explained by allele frequency differences among populations. F_{ST} ranges from 0 (for genetically identical populations) to 1 (for completely differentiated populations). We calculated F_{ST} values using the BioPerl module PopGen [79] for each autosomal SNP with a minor allele frequency (MAF) $\geq 10^{-3}$ (644,143 SNPs) at three different levels: (i) global level (either among the seven HGDP-CEPH Panel geographic regions or among the 52 Panel populations), (ii) inter-regional level (each geographic region versus the remaining ones), and (iii) intra-regional level (among populations within a region). Since F_{ST} strongly correlates with heterozygosity [41,80,81], empirical *p*-values were calculated within bins of 10,000 SNPs grouped according to MAF. The resulting distributions represent the average genetic differentiation of human populations corrected for heterozygosity.

We next applied the XP-CLR test [24] which identifies selective sweeps in a population by detecting significant genetic differentiation in an extended genomic region of interest as compared to a reference population. This method presents both the advantages of being robust to ascertainment bias and of not requiring any information on haplotypes, thus avoiding errors of haplotype estimation from genotype data. XP-CLR scores were computed at regularly spaced grid points (every 4 kb) across chromosome 16 using the genotypes from SNPs within overlapping windows of 0.1 cM around each grid point. To account for different SNP densities among genomic regions, we restricted to 200 the maximal number of SNPs used to compute a XP-CLR score within the 0.1 cM genomic region, by removing excess SNPs at random. We applied this method by considering all SNPs with a MAF $\geq 10^{-3}$ on chromosome 16 at both the regional and population levels (17,729 SNPs). *P*-values were calculated from the empirical distribution of the collected scores obtained with these SNPs. XP-CLR requires the definition of a reference population: the Sub-Saharan African samples were used as a reference for non Sub-Saharan African regions, and the European samples as a reference for Sub-Saharan Africa. For the analyses performed at the population level, we defined the Yoruba as the reference for non Sub-Saharan African samples, and the French for Sub-Saharan African samples.

The second class of methods that we used is based on EHH, *i.e.* the sharing of identical alleles across relatively long distances by most haplotypes in population samples [25]. In brief, the EHH is computed for a given SNP (the core SNP) of a sequence being interrogated for a selective sweep. In the absence of a selective sweep, recombination events break down haplotypes relatively rapidly with time and with increasing distance from the core SNP. In the case of a selective sweep, LD tends to maintain the haplotype carrying the selected allele, and the relative frequency of this (favored) haplotype will increase with time leading to so-called EHH. Integration of genetic distance in both directions from the core SNP can be used to discriminate between selected and non-selected alleles, and be applied to ancestral and derived alleles.

Analytic methods based on EHH are able to detect recent selective sweeps (*i.e.* those occurring less than 30,000 years ago [77]). Such analyses require haplotype data. We used fastPHASE v1.3.0 EM algorithm [82] to infer haplotypes with chromosome 16 SNPs for individuals from each geographic region. For each region, the *K*-selection procedure was first run several times in order to define the optimal number of clusters of similar haplotypes by minimizing chance error rates. Ultimately, phase was determined with $K = 6$ for Oceania, $K = 14$ for Europe and Central-South Asia and $K = 12$ for the remaining regions. Using these values, the EM algorithm was then run with 20 random starts and 25 iterations.

Once haplotypes were reconstructed, we computed the XP-EHH statistic [27] that compares the integrated EHH computed in a test population versus that of a reference population. Therefore, this method detects a sweep in which the selected allele has risen to near fixation in one population but remains polymorphic in the other. XP-EHH scores were computed using the same parameters as those described in Sabeti *et al.* (2007). Reference populations were defined as for XP-CLR.

We finally applied the iHS [28] that compares the rate of EHH decay observed for both the derived and ancestral allele at the core SNP. An extremely positive or negative value at the core SNP provides evidence of positive selection with unusually long haplotypes carrying the ancestral or the derived allele, respectively. The raw iHS scores were computed using the iHS option implemented in the WHAMM software developed by Voight *et al.* (2006). The scores were standardized to have null mean and unit variance in 5% bins of the derived allele frequency at the core SNP. Information on ancestral allele state was obtained from the snp131OrthoPt2Pa2Rm2.txt file downloaded from the UCSC website. We were unable to determine with certainty the ancestral allele status of 111 SNPs on chromosome 16 and we removed them from the analysis.

XP-EHH and iHS scores were calculated for all available SNPs on chromosome 16 (19,733 and 19,622, respectively) at both the regional and population levels. The resulting distributions were used to calculate empirical *p*-values.

The genetic map used for applying XP-CLR, XP-EHH and iHS was retrieved from release 22, build 36 of HapMap (www.hapmap.org).

Age of the Expansion of the -1639A *VKORC1* Allele in East Asia

We inferred the age at which the -1639A allele started to increase in frequency in East Asia by estimating the age of the most recent common ancestor carrying this allele in East Asia using the likelihood-based method implemented in the Estiage program [31]. This method assumes that all individuals derive from a common ancestor who introduced the mutation *n* generations ago. Estimation of *n* is based on the length of the haplotype shared by the individuals, which is estimated through the identification of recombination events on the ancestral haplotype by taking into account allele frequencies and recombination rates. We estimated *n* using only one haplotype per East Asian population sample (*i.e.*, 17 haplotypes). For each population, this one haplotype was constructed by taking at each locus over a 6 Mb region the allele the most frequently seen in individuals from the population carrying the -1639A allele. A mutation rate of 10^{-6} per individual and per generation, and a 25-year generation time were assumed.

Supporting Information

Figure S1 Distribution of *VKORC1* haplotypes in the 52 HGDP-CEPH samples. The haplotype carrying the -1639A allele (H1) is represented in red and the ancestral haplotype (H6) in black.
(TIF)

Figure S2 Pairwise LD between the seven *VKORC1* SNPs at the regional and global level. Red squares indicate statistically significant (logarithm of odds >2) LD between the pair of SNPs, as measured by the *D*' statistic [75] with the Haploview software [73]; darker colors of red indicate higher values of *D*', up to a maximum of 1. White squares indicate pairwise *D*' values of <1 with no statistically significant evidence of LD. Blue squares indicate pairwise *D*' values of 1 but without statistical significance.
(TIF)

Figure S3 Genome-wide empirical distributions of inter-regional F_{ST} values against MAF in the seven geographic regions. Empirical distributions of F_{ST} were constructed by calculating an F_{ST} value for 644,413 SNPs having a MAF ≥0.001 at the global level. Individual values of F_{ST} calculated for each of the seven *VKORC1* SNPs are plotted against their global MAF. The functional rs9923231 SNP is shown in red. The 50th, 95th and 99th percentiles are indicated as dotted, dashed and full red lines, respectively.
(TIFF)

Figure S4 Genome-wide empirical distributions of intra-regional F_{ST} values against MAF in the seven geographic regions. Empirical distributions of F_{ST} were constructed by calculating an F_{ST} value for all SNPs having a MAF ≥0.001 at the intra-regional level. Individual values of F_{ST} calculated for each of the seven *VKORC1* SNPs are plotted against the regional MAF. The functional rs9923231 SNP is shown in red. The 50th, 95th and 99th percentiles are indicated as dotted, dashed and full red lines, respectively.
(TIFF)

Figure S5 LD patterns over a 2 Mb region centered on *VKORC1* in the seven geographic regions. Pairwise LD, depicted as *D*', is shown for SNPs with a MAF ≥0.05 at the global level. *D*' values are displayed in different colors from yellow to red for *D*' = 0 to *D*' = 1, respectively. The plot was produced using the snp.plotter R package [74]. The vertical dashed lines delineate *VKORC1* gene position.
(TIF)

Figure S6 Allele frequency distribution of the seven *VKORC1* SNPs in the 52 HGDP-CEPH samples: rs9923231 **(A)**, rs13336384, **(B)** rs9934438 **(C)**, rs8050894 **(D)**, rs2359612 **(E)**, rs7200749 **(F)** and rs7294 **(G)**. The derived and ancestral alleles are represented in orange and blue, respectively.
(TIF)

Table S1 Global F_{ST} values among populations and among regions for the seven *VKORC1* SNPs.
(XLS)

Table S2 Results of the XP-CLR test in a 16 kb region centered on *VKORC1* in the 52 HGDP-CEPH samples.
(XLS)

Table S3 Results of the XP-EHH and iHS tests in the 52 HGDP-CEPH samples.
(XLS)

Table S4 Results of the XP-CLR test in the ~ 500 kb genomic region of the LD block encompassing *VKORC1* in East Asia.
(XLS)

Table S5 Results of the XP-EHH, iHS tests, inter-regional F_{ST} and global F_{ST} for all SNPs located in the linkage disequilibrium block encompassing *VKORC1* in East Asia.
(XLS)

Table S6 Description of the 52 HGDP-CEPH samples grouped into seven main geographic regions.
(XLS)

Acknowledgments

PL thanks Txema Heredia, Angel Carreno and Jordi Rambla for computational support. BP thanks Steven Gazal for his valuable assistance in writing scripts, Quentin Vincent for his advices on the datation protocol, and Marie-Claude Babron and Remi Kazma for their helpful comments on the manuscript.

Author Contributions

Conceived and designed the experiments: EG AS. Performed the experiments: BP PL HB. Analyzed the data: BP PL EG AS. Contributed reagents/materials/analysis tools: HB EP HMC. Wrote the paper: BP PL EP HMC EG AS.

References

1. Hirsh J, Dalen JE, Anderson DR, Poller L, Bussey H, et al. (1998) Oral anticoagulants: mechanism of action, clinical effectiveness, and optimal therapeutic range. Chest 114: 445S–469S.
2. Hyers TM, Agnelli G, Hull RD, Weg JG, Morris TA, et al. (1998) Antithrombotic therapy for venous thromboembolic disease. Chest 114: 561S–578S.
3. D'Andrea G, D'Ambrosio R, Margaglione M (2008) Oral anticoagulants: Pharmacogenetics Relationship between genetic and non-genetic factors. Blood Rev 22: 127–140.
4. Cooper GM, Johnson JA, Langaee TY, Feng H, Stanaway IB, et al. (2008) A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. Blood 112: 1022–1027.
5. Takeuchi F, McGinnis R, Bourgeois S, Barnes C, Eriksson N, et al. (2009) A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. PLoS Genet 5: e1000433.
6. Teichert M, Eijgelsheim M, Rivadeneira F, Uitterlinden AG, van Schaik RH, et al. (2009) A genome-wide association study of acenocoumarol maintenance dosage. Hum Mol Genet 18: 3758–3768.
7. Cha PC, Mushiroda T, Takahashi A, Kubo M, Minami S, et al. (2010) Genome-wide association study identifies genetic determinants of warfarin responsiveness for Japanese. Hum Mol Genet 19: 4735–4744.
8. Goldstein JA, de Morais SM (1994) Biochemistry and molecular biology of the human CYP2C subfamily. Pharmacogenetics 4: 285–299.
9. Stec DE, Roman RJ, Flasch A, Rieder MJ (2007) Functional polymorphism in human CYP4F2 decreases 20-HETE production. Physiol Genomics 30: 74–81.
10. Bardowell SA, Stec DE, Parker RS (2010) Common variants of cytochrome P450 4F2 exhibit altered vitamin E-{omega}-hydroxylase specific activity. J Nutr 140: 1901–1906.
11. Li T, Chang CY, Jin DY, Lin PJ, Khvorova A, et al. (2004) Identification of the gene for vitamin K epoxide reductase. Nature 427: 541–544.
12. Rost S, Fregin A, Ivaskevicius V, Conzelmann E, Hortnagel K, et al. (2004) Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. Nature 427: 537–541.
13. Limdi NA, Wadelius M, Cavallari L, Eriksson N, Crawford DC, et al. (2010) Warfarin pharmacogenetics: a single VKORC1 polymorphism is predictive of dose across 3 racial groups. Blood 115: 3827–3834.
14. Ross KA, Bigham AW, Edwards M, Gozdzik A, Suarez-Kurtz G, et al. (2010) Worldwide allele frequency distribution of four polymorphisms associated with warfarin dose requirements. J Hum Genet 55: 582–589.
15. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–1104.
16. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, et al. (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics 1: 274–286.
17. Storz JF, Payseur BA, Nachman MW (2004) Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Mol Biol Evol 21: 1800–1811.
18. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
19. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.
20. Rieder MJ, Reiner AP, Gage BF, Nickerson DA, Eby CS, et al. (2005) Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. N Engl J Med 352: 2285–2293.
21. Wu AH, Wang P, Smith A, Haller C, Drake K, et al. (2008) Dosing algorithm for warfarin using CYP2C9 and VKORC1 genotyping from a multi-ethnic population: comparison with other equations. Pharmacogenomics 9: 169–178.
22. Yang L, Ge W, Yu F, Zhu H (2010) Impact of VKORC1 gene polymorphism on interindividual and interethnic warfarin dosage requirement–a systematic review and meta analysis. Thromb Res 125: e159–166.
23. Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23: 23–35.
24. Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Res 20: 393–402.
25. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.
26. Weir BS, Hill WG (2002) Estimating F-statistics. Annu Rev Genet 36: 721–750.
27. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.
28. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.
29. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15: 1496–1502.
30. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, et al. (2002) A human genome diversity cell line panel. Science 296: 261–262.
31. Genin E, Tullio-Pelet A, Begeot F, Lyonnet S, Abel L (2004) Estimating the age of rare disease mutations: the example of Triple-A syndrome. J Med Genet 41: 445–449.
32. Li J, Zhang L, Zhou H, Stoneking M, Tang K (2011) Global patterns of genetic diversity and signals of natural selection for human ADME genes. Hum Mol Genet 20: 528–540.
33. Marsh S, King CR, Porche-Sorbet RM, Scott-Horton TJ, Eby CS (2006) Population variation in VKORC1 haplotype structure. J Thromb Haemost 4: 473–474.
34. Limdi NA, Beasley TM, Crowley MR, Goldstein JA, Rieder MJ, et al. (2008) VKORC1 polymorphisms, haplotypes and haplotype groups on warfarin dose among African-Americans and European-Americans. Pharmacogenomics 9: 1445–1458.
35. Schwarz UI, Ritchie MD, Bradford Y, Li C, Dudek SM, et al. (2008) Genetic determinants of response to warfarin during initial anticoagulation. N Engl J Med 358: 999–1008.
36. Geisen C, Watzka M, Sittinger K, Steffens M, Daugela L, et al. (2005) VKORC1 haplotypes and their impact on the inter-individual and inter-ethnical variability of oral anticoagulation. Thromb Haemost 94: 773–779.
37. Bodin L, Verstuyft C, Tregouet DA, Robert A, Dubert L, et al. (2005) Cytochrome P450 2C9 (CYP2C9) and vitamin K epoxide reductase (VKORC1) genotypes as determinants of acenocoumarol sensitivity. Blood 106: 135–140.
38. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, et al. (2005) Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res 15: 1553–1565.
39. Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci U S A 103: 135–140.
40. Teo YY, Sim X, Ong RT, Tan AK, Chen J, et al. (2009) Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res 19: 2154–2162.
41. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40: 340–345.
42. Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19: 711–722.
43. Xue Y, Zhang X, Huang N, Daly A, Gillson CJ, et al. (2009) Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. Genetics 183: 1065–1077.
44. Hofer T, Ray N, Wegmann D, Excoffier L (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. Ann Hum Genet 73: 95–108.
45. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19: 826–837.
46. Nath SK, Han S, Kim-Howard X, Kelly JA, Viswanathan P, et al. (2008) A nonsynonymous functional variant in integrin-alpha(M) (encoded by ITGAM) is associated with systemic lupus erythematosus. Nat Genet 40: 152–154.

47. Hom G, Graham RR, Modrek B, Taylor KE, Ortmann W, et al. (2008) Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. N Engl J Med 358: 900–909.
48. Jarvinen TM, Hellquist A, Koskenmies S, Einarsdottir E, Panelius J, et al. (2010) Polymorphisms of the ITGAM gene confer higher risk of discoid cutaneous than of systemic lupus erythematosus. PLoS One 5: e14212.
49. Scheinfeldt LB, Biswas S, Madeoy J, Connelly CF, Akey JM (2011) Clusters of adaptive evolution in the human genome. Front Genet 2: 50.
50. Barreiro LB, Quintana-Murci L (2010) From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11: 17–30.
51. Suttie JW (1985) Vitamin K-dependent carboxylase. Annu Rev Biochem 54: 459–477.
52. Oldenburg J, Marinova M, Muller-Reible C, Watzka M (2008) The vitamin K cycle. Vitam Horm 78: 35–62.
53. Yan L, Zhou B, Greenberg D, Wang L, Nigdikar S, et al. (2004) Vitamin K status of older individuals in northern China is superior to that of older individuals in the UK. Br J Nutr 92: 939–945.
54. Beavan SR, Prentice A, Stirling DM, Dibba B, Yan L, et al. (2005) Ethnic differences in osteocalcin gamma-carboxylation, plasma phylloquinone (vitamin K1) and apolipoprotein E genotype. Eur J Clin Nutr 59: 72–81.
55. Kaneki M, Hodges SJ, Hosoi T, Fujiwara S, Lyons A, et al. (2001) Japanese fermented soybean food as the major determinant of the large geographic difference in circulating levels of vitamin K2: possible implications for hip-fracture risk. Nutrition 17: 315–321.
56. Fujita Y, Iki M, Tamaki J, Kouda K, Yura A, et al. (2011) Association between vitamin K intake from fermented soybeans, natto, and bone mineral density in elderly Japanese men: the Fujiwara-kyo Osteoporosis Risk in Men (FORMEN) study. Osteoporos Int.
57. Vermeer C, Shearer MJ, Zittermann A, Bolton-Smith C, Szulc P, et al. (2004) Beyond deficiency: potential benefits of increased intakes of vitamin K for bone and vascular health. Eur J Nutr 43: 325–335.
58. Wardrop D, Keeling D (2008) The story of the discovery of heparin and warfarin. Br J Haematol 141: 757–763.
59. Mueller RL, Scheidt S (1994) History of drugs for thrombotic disease. Discovery, development, and directions for the future. Circulation 89: 432–449.
60. Kohn MH, Pelz HJ, Wayne RK (2000) Natural selection mapping of the warfarin-resistance gene. Proc Natl Acad Sci U S A 97: 7911–7915.
61. Kohn MH, Pelz HJ, Wayne RK (2003) Locus-specific genetic differentiation at Rw among warfarin-resistant rat (Rattus norvegicus) populations. Genetics 164: 1055–1070.
62. Diaz JC, Song Y, Moore A, Borchert JN, Kohn MH (2010) Analysis of vkorc1 polymorphisms in Norway rats using the roof rat as outgroup. BMC Genet 11: 43.
63. Song Y, Endepols S, Klemann N, Richter D, Matuschka FR, et al. (2011) Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. Curr Biol 21: 1296–1301.
64. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, et al. (2011) Classic selective sweeps were rare in recent human evolution. Science 331: 920–924.
65. Pritchard JK, Di Rienzo A (2010) Adaptation - not by sweeps alone. Nat Rev Genet 11: 665–667.
66. Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20: R208–215.
67. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, et al. (2009) The role of geography in human adaptation. PLoS Genet 5: e1000500.
68. Consortium TGP (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
69. Cann HM (1998) Human genome diversity. C R Acad Sci III 321: 443–446.
70. Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70: 841–847.
71. Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73: 1162–1169.
72. Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16: 37–48.
73. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.
74. Luna A, Nicodemus KK (2007) snp.plotter: an R-based SNP/haplotype association and linkage disequilibrium plotting package. Bioinformatics 23: 774–776.
75. Lewontin RC (1964) The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics 49: 49–67.
76. Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations Theoretical and Applied Genetics 38: 226–231.
77. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. (2006) Positive natural selection in the human lineage. Science 312: 1614–1620.
78. Wright S (1951) The genetical structure of populations. Annals of Eugenics 15 323–354.
79. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12: 1611–1618.
80. Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. The Royal Society 263: 1619–1626.
81. Gardner M, Bertranpetit J, Comas D (2008) Worldwide genetic variation in dopamine and serotonin pathway genes: implications for association studies. Am J Med Genet B Neuropsychiatr Genet 147B: 1070–1075.
82. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629–644.
83. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, et al. (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304: 581–584.
84. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, et al. (2005) Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308: 107–111.

Chapter 4

SCAN THE GENOME FOR POSITIVE SELECTION.

Hafid Laayouni, Marije Oosting, Pierre Luisi, Mihai Ioana Santos Alonso, Isis Ricaño-Ponce, Gosia Trynka, Alexandra Zhernakova, Theo S. Plantinga, Shih-Chin Cheng, Jos W. M. van der Meer Radu Popp, Ajit Sood, B. K. Thelma, Cisca Wijmenga, Leo A. B. Joosten, Jaume Bertranpetit, and Mihai G. Netea

Published [264]

Laayouni H, Oosting M, Luisi P, Ioana M, Alonso S, Ricaño-Ponce I, Trynka G, Zhernakova A, Plantinga TS, Cheng SC, van der Meer JW, Popp R, Sood A, Thelma BK, Wijmenga C, Joosten LA, Bertranpetit J, Netea MG. Convergent evolution in European and Rroma populations reveals pressure exerted by plague on Toll-like receptors. Proc Natl Acad Sci U S A. 2014 Feb 18;111(7):2668-73. doi: 10.1073/pnas.1317723111.

Chapter 5

DISTRIBUTION OF SELECTIVE EVENTS WITHIN A SMALL-SCALE PROTEIN-PROTEIN INTERACTION MAP.

Pierre Luisi, David Alvarez-Ponce, Giovanni Marco Dall'Olio, Martin Sikora, Jaume Bertranpetit, and Hafid Laayouni

Published [265]

Luisi P, Alvarez-Ponce D, Dall'Olio GM, Sikora M, Bertranpetit J, Laayouni H. Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations. Mol Biol Evol. 2012 ;29(5): 1379-92. doi: 10.1093/molbev/msr298.

Chapter 6

DISTRIBUTION OF SELECTIVE EVENTS WITHIN A LARGE-SCALE PROTEIN-PROTEIN INTERACTION MAP.

Pierre Luisi, David Alvarez-Ponce, Marc Pybus, Mario A. Fares, Jaume Bertranpetit, and Hafid Laayouni

Submitted

RESEARCH ARTICLE

Recent Positive Selection Has Acted On Genes Encoding Proteins With More Interactions Within the Whole Human Interactome

Pierre Luisi[1]†, David Alvarez-Ponce[2,3]†, Marc Pybus[1], Mario A. Fares[2,4], Jaume Bertranpetit[1]* and Hafid Laayouni[1,5]*.

(† These authors contributed equally to this work)

[1]Institute of Evolutionary Biology (Universitat Pompeu Fabra-CSIC), CEXS-UPF-PRBB, Barcelona, Catalonia, Spain.

[2]Integrative Systems Biology Group, Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas (CSIC)-Universidad Politécnica de Valencia (UPV), Valencia, Spain.

[3]Current address: Biology Department, University of Nevada, Reno, NV, USA.

[4]Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin, Ireland.

[5]Departament de Genètica i de Microbiologia, Grup de Biologia Evolutiva (GBE), Universitat Autonòma de Barcelona, Bellaterra (Barcelona), Spain

* Authors for correspondence.

Pierre Luisi:

IBE, Institut de Biologia Evolutiva (UPF-CSIC)
CEXS-Universitat Pompeu Fabra-PRBB
Doctor Aiguader 88
08003 Barcelona, Catalonia, Spain

email: pierre.luisi@upf.edu

David Alvarez-Ponce:

Instituto de Biología Molecular y Celular de Plantas (Consejo Superior de Investigaciones Científicas-Universidad Politécnica de Valencia)
Ingeniero Fausto Elio s/n
46022 Valencia, Spain

Biology Department,
University of Nevada, Reno

Max Fleischmann Agriculture Building
1664 N. Virginia Street
Reno, NV 89557-0314

email: david.alvarez@csic.es

Marc Pybus:

IBE, Institut de Biologia Evolutiva (UPF-CSIC)
CEXS-Universitat Pompeu Fabra-PRBB
Doctor Aiguader 88
08003 Barcelona, Catalonia, Spain

email: marc.pybus@upf.edu

Mario A. Fares:

Instituto de Biología Molecular y Celular de Plantas (Consejo Superior de Investigaciones Científicas-Universidad Politécnica de Valencia)
Ingeniero Fausto Elio s/n
46022 Valencia, Spain

Smurfit Institute of Genetics,
University of Dublin,
Trinity College,
2 College Green
Dublin 2, Ireland

email: faresm@tcd.ie

Hafid Laayouni:

IBE, Institut de Biologia Evolutiva (UPF-CSIC)
CEXS-Universitat Pompeu Fabra-PRBB
Doctor Aiguader 88
08003 Barcelona, Catalonia, Spain

Departament de Genètica i de Microbiologia
Facultat de Biociències, Edifici Cn
Universitat Autònoma de Barcelona
08193 Bellaterra, Catalonia, Spain

email: hafid.laayouni@upf.edu

Jaume Bertranpetit:

IBE, Institut de Biologia Evolutiva (UPF-CSIC)
CEXS-Universitat Pompeu Fabra-PRBB
Doctor Aiguader 88
08003 Barcelona, Catalonia, Spain

email: jaume.bertranpetit@upf.edu

ABSTRACT

Genes vary in their likelihood to undergo adaptive evolution. The genomic factors that determine adaptability, however, remain poorly understood. Genes function in the context of molecular networks, with some occupying more important positions than others and thus being likely to be under stronger selective pressures. However, how positive selection distributes across the different parts of molecular networks is still not fully understood. Here, we inferred positive selection using comparative genomics and population genetics approaches through the comparison of 10 mammalian and 270 human genomes, respectively. In agreement with previous results, we found that genes with lower network centralities are more likely to evolve under positive selection (as inferred from divergence data). Surprisingly, polymorphism data yields results in the opposite direction than divergence data: genes with higher centralities are more likely to have been targeted by recent positive selection during recent human evolution. Our results indicate that the relationship between centrality and the impact of adaptive evolution highly depends on the mode of positive selection and/or the evolutionary time-scale. Most likely, network adaptation occurs through intra-specific adaptive leaps affecting key network genes, followed by fine-tuning adaptations in less important network regions.

Keywords

Physical protein interaction, Protein interaction network, Natural selection, Positive selection, Mammals, Humans

BACKGROUND

In recent years, the availability of large-scale network and genomic datasets has allowed researchers to study the relationship between the position of proteins within molecular networks and their patterns of molecular evolution (Cork & Purugganan 2004; Wagner 2012). These studies have shown that the strength of purifying selection acting on individual genes is affected by the position that their encoded products occupy in molecular networks. Indeed, genes acting at the centre of protein-protein interaction networks (PINs) and metabolic networks (i.e., genes coding for proteins with many interactions or connections) evolve under higher levels of purifying selection than those acting at the network periphery (Alvarez-Ponce & Fares 2012; Alvarez-Ponce 2012; Fraser et al. 2002; Hahn & Kern 2005; Vitkup et al. 2006) (but see (Hahn et al. 2004; Jordan et al. 2003))**.** Furthermore, interacting proteins evolve at similar rates, probably as a result of molecular coevolution (Agrafioti et al. 2005; Codoñer & Fares 2008; Cui et al. 2009; Fraser et al. 2002; Lovell & Robertson 2010; Pérez-Bercoff et al. 2013).

Less well understood, however, is how adaptive events distribute across molecular pathways and networks. Some evidence supports that adaptive events tend to occur in less centrally located regions of gene networks. In an early study using two genomes, the human and chimpanzee genomes, Kim *et al.* found that positive selection often targeted genes acting at the periphery of the PIN (Kim et al. 2007). Powerful detection of positive selection requires, nevertheless, comparing many genomes (Anisimova et al. 2002; Kosiol et al. 2008), making it appropriate to re-evaluate this trend in light of the currently available mammalian genomes.

In addition, some recent population genetics studies appear to contradict the notion that positive selection targets preferentially the periphery of molecular networks. Indeed, positive selection often targets genes acting at the most "influential" positions of these pathways, including the most entral genes in the human insulin/mTOR pathway (Luisi et al. 2012), genes acting at bifurcation points of the human N-glycosylation pathway (Dall'Olio et al. 2012) and the Drosophila pathways involved in glucose metabolism (Flowers et al. 2007), and the gene encoding the first enzyme of the Arabidopsis glucosinolate pathway (Olson-Manning et al. 2013). Simulation studies also indicate that adaptation preferentially targets genes acting at the upstream and branch-point parts of pathways, at least when the system is far from the fitness optimum (Rausher 2012; Wright & Rausher 2010). Proteins occupying these key network positions are expected to exert strong influence over the pathway function, and thus on the associated phenotypes and organism's fitness (Olson-Manning et al. 2013;

Rausher 2012; Wright & Rausher 2010). Therefore, positive selection on genes encoding such proteins may lead to rapid adaptation.

Here, we make use of the unprecedented wealth of genomic (Kersey et al. 2012; The 1000 Genomes Project Consortium 2012) and interactomic data (Stark et al. 2011), to ascertain what parts of the human protein–protein interaction network were affected by positive selection, using both comparative genomics and population genetics approaches. We found that positive selection, as inferred from divergence data, preferentially targets genes acting at more peripheral positions in the network, in agreement with previous observations (Kim et al. 2007). Conversely, genes with signatures of recent positive selection, identified considering polymorphism data, occupy more central parts of the network. We discuss on the apparently contradictory results from divergence and polymorphism data and propose, for the first time, an evolutionary scenario reconciling both patterns.

MATERIAL AND METHODS

Reconstructing the Human Protein–Protein Interaction Network

The human protein–protein interaction network (PIN) was reconstructed from the interactions available from the BioGRID database version 3.1.81 (Stark et al. 2011). Only non-redundant physical interactions were considered to calculate centrality measures. We removed from our analysis proteins without an Ensembl ID as well as Ubiquitin C (encoded by the gene with Ensembl ID ENSG00000150991), which has an outlier degree centrality.

Detecting Natural Selection Events from 10 Mammalian Genomes

In order to infer events of positive selection that have occurred during mammals evolution we used sequence data for a set of mammals, enriched in primates. The analysis was restricted to 10 high-coverage genomes: human, chimpanzee, gorilla, orang-utan, macaque, mouse, rat, cow, dog, and opossum. The platypus genome was not included in the analysis, as the currently available assembly is highly fragmented, making gene annotation difficult. Also excluded were non-mammalian genomes, in order to avoid the problem of saturation of synonymous sites (Smith & Smith 1996), and to maximize the number of genes with 1:1 orthologs in all studied genomes.

All protein and coding (CDS) sequences for the selected genomes were obtained from Ensembl release 62 (Kersey et al. 2012). For each of the 9,041 human protein-coding genes represented in the PIN, we searched the 9 non-human genomes for 1:1 orthologs using the best reciprocal BLAST approach. First, we selected the longest protein (or, in the case of multiple proteins sharing the maximal length, that classified as the canonical isoform), and used it as query in a BLASTP search against each of the non-human proteomes. Second, for the best hits in each proteome, we performed a BLASTP search against the human proteome. If the hit obtained in the second search was the original human protein, then it was considered to be a 1:1 ortholog. Only human genes with 1:1 orthologs in all 9 non-human genomes were used in subsequent analyses (in total, 5,916 genes met this criterion).

Each group of orthologous proteins was aligned using ProbCons 1.12 (Do et al. 2005). Because tests of positive selection are sensitive to sequencing, annotation and alignment errors (Scheinfeldt et al. 2009; Talavera & Castresana 2007), we used highly stringent criteria to filter our alignments. First, unreliably aligned regions were removed using Gblocks version

0.91b (Talavera & Castresana 2007), with default parameters. Additionally, we used an *ad-hoc* filtering procedure in order to remove annotation errors, including the following steps: (1) identification of unique amino acid replacement (i.e., amino acids that are unique to a given species in a certain alignment column); (2) identification of alignment regions with a very high incidence of unique substitutions in the same species; in particular, we used a sliding window approach to identify regions of 15 amino acids containing 10 or more unique substitutions in the same sequence, as well as regions of 5 amino acids containing 5 unique substitutions in the same sequence; these patterns are unlikely to represent true divergence between species, provided that the species included in the current analysis are relatively closely related; and (3) removal of these alignment regions. These procedures resulted in the removal of 35.5% of amino acid positions. The resulting filtered protein alignments were used to guide the alignment of the corresponding CDSs.

We evaluated the impact of both purifying and positive selection on each orthologous group using the program codeml from the package PAML 4.4 (Yang et al. 2005). For each CDS alignment, three different evolutionary models (M0, M7 and M8) were fitted. First, for each gene, an overall ω estimate was obtained from the M0 model, which assumes a homogeneous ω for all branches in the tree and all codons in the alignment. This score was used as a proxy of the impact of purifying selection. Second, in order to infer the action of positive selection, we applied the M7 vs. M8 test (Nielsen & Yang 1998). The M7 model assumes that codons' ω values follow a beta distribution, limited to the interval (0, 1), whereas model M8 allows for an additional class of codons with $\omega > 1$. The likelihood ratio test was used to contrast whether model M8 fits the data significantly better than model M7. Twice the difference between the log-likelihoods of both nested models, [$2\Delta\ell = 2 \times (\ell M8 - \ell M7)$, where ℓii is the log-likelihood of the observed data under model i], is assumed to follow a χ2 distribution with two degrees of freedom. In order to avoid the problem of local optima, for each gene each model was fitted three times, using different starting ω values (0.04, 0.4 and 4), and the computation with the highest likelihood was retained. The commonly accepted tree topology was used.

In order to discard potential alignment errors, not detected by our stringent filtering, the alignments corresponding to genes with $P < 0.1$ in the likelihood ratio test for positive selection were inspected visually. Alignment regions containing evident errors were manually removed using BioEdit v7.0.5.2 (Hall 1999), and analyses of positive selection were re-run. This process was iterated until no further errors were detected in the alignments corresponding to genes with putative signatures of positive selection. We obtained a total of 554 genes with putative signatures of positive selection (divPSGs; $P < 0.05$).

Detecting Natural Selection from 1000 Human Genomes

We obtained phased genotypes from low-coverage data of the phase I of the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012), which makes available data for over 36 millions Single Nucleotide Variants (SNVs) for 1,092 individuals sampled from 14 populations worldwide. We used a subset of 270 individuals from YRI, CEU and CHB populations.

For each of the 9,041 genes contained in the PIN, we analysed the genomic region corresponding to the transcript spanning the longest chromosome region. Gene coordinates were obtained from release 37 of the human genome at NCBI (Flicek et al. 2010). We removed 365 genes located at sex chromosomes because some of the methods used to detect signals of positive selection have been devised for autosomal regions, or provide results that cannot be compared between autosomal and sex chromosomes. In order to increase the statistical power in the detection of positive selection, we removed from the analyses 96 genes with less than 10 SNVs annotated in the 1000 genomes.

We used the genetic map provided by the 1000 Genomes Consortium. Ancestral states inferred from comparison with orthologous sequences in the chimpanzee and rhesus macaque genomes were obtained from the UCSC Genome Bioinformatics Site (Karolchik et al. 2009) (http://genome.ucsc.edu/; table “snp128OrthoPanTro2RheMac2”).

Retained genes (a total of 8,580), have a length ranging from 0.414 to 2,305 Kb (mean = 61.70 Kb; median = 25.95 Kb) and are covered by a total of 6,815,879 SNVs. The number of SNVs located in a gene ranges from 10 (28 genes) to 45,577 with a mean of 794.4 and a median of 312.

To identify the genes belonging to the PIN that have evolved under positive selection during human evolution, we applied three different tests: (i) the Cross-Population Composite Likelihood Ratio method (Chen et al. 2010) (XP-CLR), based on the multi-locus allele frequency differentiation between two populations, (ii) the integrated Haplotype Score

(Voight et al. 2006) (iHS), which aims to detect extended haplotype homozygosity from the local haplotype structure, and (iii) *DH* (Zeng et al. 2007), based on the excess of rare variants, which combines Tajima's *D* (Tajima 1989) and Fay and Wu's *H* (Fay & Wu 2000). Those tests are designed assuming the *hard sweep* model which states that a new advantageous mutation arises in the population and rapidly increases in frequency hitchhiking the surrounding neutral variants located on the same haplotype.

We computed a raw iHS for each SNV with ancestral state information following the method proposed by Voight et al. *(2006)*. We used the script available at http://hgdp.uchicago.edu/Software/, which we slightly modified in order to speed up computation times; thresholds for Extended Haplotype Homozygosity (EHH) decay were modified from 0.25 to 0.15 and we used a size for the analysed region of 0.2 Mb (original size: 2.5 Mb). Using coalescent simulations (COSI) (Schaffner et al. 2005) we validated that these changes do not affect sensitivity and specificity of the method (data not shown). Standardized iHS scores were obtained by grouping SNVs into 20 bins separated by a derived allele frequency of 0.05, subtracting the mean, and dividing by the standard deviation for all SNVs in the same bin as in Voight et al. (2006). Extreme positive or negative values indicate high extended haplotype homozygosity of haplotypes carrying the ancestral or derived allele, respectively. Hence, we consider both extreme positive or negative iHS as potential signatures of positive selection. We integrated the |iHS| scores observed at each gene of interest into a gene-level summary statistic using the mean.

The XP-CLR method aims at detecting important genetic differentiation in an extended genomic region in comparison with a reference population. This method provides a good localization of the position of the selected variant (Chen et al. 2010). XP-CLR scores were computed at regularly spaced grid points (every 2 Kb) using the information from SNVs within a flanking window of 0.2 cM. To account for different SNV densities among genomic regions, we restricted to 200 the maximal number of SNVs used to calculate a XP-CLR score within each window, by randomly removing SNVs in excess. We integrated the XP-CLR scores observed at each gene of interest into a gene-level summary statistic using the mean.

Extreme iHS and XP-CLR scores could also be attributable to the action of non-selective events such as demographic changes and genetic drift. However, these selectively neutral events act randomly on the genome, in contrast with positive selection, which targets specific genes. Therefore, we adopted an outlier approach to infer the action of positive selection on PIN genes (Kelley et al. 2006; Teshima et al. 2006): we evaluated the significance of the scores for each gene by taking into account the whole genome context. For that purpose, we used a genomic gene-level background containing all annotated genes that were distant one from each other and from the 8,580 genes included in the analysis, by at least 5 Kb and contained at least 10 SNVs. The complete background gene set obtained thus includes 13,388 genomic regions and 8,431,716 SNVs. For each of these background genomic regions, we computed the mean summary statistics based on iHS and XP-CLR and then obtained gene-level empirical distributions. Empirical *P*-values associated to iHS and XP-CLR for PIN genes were obtained using these distributions.

For each gene, using the SNVs with ancestral state information, we also computed Tajima's *D*, Fay and Wu's *H* and *DH*, using a program kindly provided by Kai Zeng. For each gene, the *DH* *P*-value was obtained as in (Zeng et al. 2007) from Tajima's *D* and Fay and Wu's *H* by a bivariate comparison to their neutral distributions. However, instead of using 10,000 replicates of coalescent simulations to build these neutral distributions as in the original article, we used the 13,388 genomic regions described before in order to better take into account demographic forces that acted on the studied populations.

In order to summarize the results of the three different tests, we combined the gene-level empirical *P*-values obtained as described above using the Fisher combination test:

$$Z_F = -2\log\sum_{i=1}^{i=3} P_i \quad ,$$

where P_i are the empirical *P*-values obtained from the three tests. Thus, for each gene we obtained a unique Z_F score, which follows a χ^2 distribution with 6 degrees of freedom. This combination requires independence of the three combined *P*-values. We confirmed that this assumption is appropriate to our data (Supplementary Figure 1). We invoked positive selection if the *P*-value associated to the Z_F score was below 5%. Therefore, we obtained 4 lists of genes with putative signatures of positive selection inferred from polymorphism data (polyPSGs): 3 populations + global level.

The major limitation of the methods implemented to detect positive selection using polymorphism data is that demographic events, such as population growth, bottleneck, and/or subdivision, can mimic patterns similar to those produced by selection. However, the outlier approach framework that we implemented and which combines three tests that consider three different molecular patterns (namely genetic differentiation, site frequency spectrum and linkage disequilibrium) is very likely to overcome this issue.

In order to estimate the strength of purifying selection acting on the genes involved in the PIN we calculated the average Derived Allele Frequency (DAF) among the 270 individuals belonging to YRI, CEU and CHB populations (The 1000 Genomes Project Consortium 2012).

Determining Fitness Effect of Genes

Using data from the Mouse Genome Database “MRK_Ensembl_Pheno.rpt” (Bult et al. 2008) (file downloaded on 7 October 2010), we classified genes as essential and non-essential when described to be lethal and viable when knocked out in mice, respectively. We retrieved such information for 3,994 genes represented in the PIN.

We also used the functional indispensability score (Khurana et al. 2013) estimated from functional and evolutionary properties. This score shows great performance to distinguish between essential genes (those showing clinical features of death before puberty or infertility when Loss-of-Function –LoF– mutations occur (Liao & Zhang 2008)) and LoF-tolerant genes (those observed to contain homozygous LoF mutations in at least one individual in the 1000 Genomes Pilot Data (MacArthur et al. 2012)). We obtained the functional indispensability score for 8,816 genes involved in the PIN.

RESULTS

Positive selection inferred from divergence data and gene centrality in the human protein-protein interaction network

We used 10 mammalian genomes (Kersey et al. 2012) to infer events of positive selection that took place within the last ~165 million years. The test used in this study looks for a non-synonymous to synonymous divergence ratio ($\omega = d_N/d_S$) higher than 1 at a subset of codons (Nielsen & Yang 1998). It provides a positive selection likelihood score, termed $2\Delta\ell$ (see Methods), that is proportional to the likelihood of positive selection. We identified a total of 554 putative positively selected genes (divPSGs; those with $P < 0.05$).

We measured the difference in the mean degree (number of protein-protein interactions, or number of proteins with which a protein interacts) between divPSGs and the other genes in the network (non-divPSGs), and tested whether this difference was expected at random through 10,000 random permutations of the two groups containing divPSGs and non-divPSGs. We observed that divPSGs encode proteins with a lower significantly lower degree than non-divPSGs (permutation test: $P = 0.0067$; Figure 1A; Supplementary Table 1). Indeed, divPSGs and non-divPSGs encode proteins with, on average, 7.578 and 9.122 interactions, respectively, i.e. the degree for divPSGs is 17% lower than the one observed for non-divPSGs. The magnitude of this difference is similar to previous observations (Kim et al. 2007).

We next observed that log-likelihood increments ($2\Delta\ell$ scores) from the positive selection test exhibit a significant negative correlation with proteins’ degrees (Spearman's rank correlation coefficient, $\rho = -0.0490$; $P = 0.0002$; Table 1), indicating that central genes are less likely to be under positive selection. Finally, when proteins were binned into four degree classes (low, medium-low, medium-high and high degree), we observed a continuous decrease in their positive selection likelihood scores ($2\Delta\ell$) (Figure 2D; Table 1). Indeed, the non-parametric Analysis Of Variance (ANOVA) *F*-test is significant ($P = 0.0101$), and there is a trend towards higher $2\Delta\ell$ scores in the lower degree groups (linear trend test on ranks; $P = 0.0014$). Taken together, our observations indicate that adaptation (as inferred from divergence data) more frequently occurs at the less connected proteins of the human interactome, consistent with previous observations (Kim et al. 2007).

Positive selection inferred from polymorphism data and gene centrality in the human protein-protein interaction network

We inferred recent events of positive selection in humans using genomic data from three different populations of West African, Northern European and East Asian ancestry (YRI, CEU and CHB, respectively). We used a Fisher's combination (Z_F score) of three tests of positive selection assuming the *hard sweep* model: XP-CLR (Chen et al. 2010), iHS (Voight et al. 2006) and *DH* (Zeng et al. 2007) (see Material and Methods). Assuming that Z_F follows a χ^2 distribution with 6 degrees of freedom, we identified putative positive selection genes (polyPSGs).

We measured the difference in the mean degree between these genes and genes without evidences of having evolved under positive selection (non-polyPSGs) (Figure 1A; Supplementary Table 1). When all populations were analysed together (global analysis), we observed a statistically significant higher degree for genes with signatures of positive selection (permutation test: P = 0.0254). Indeed, polyPSGs and non-polyPSGs encode proteins with, on average, 9.637 and 8.107 interactions, respectively, i.e. the degree for polyPSGs is 19% higher than the one observed for non-polyPSGs. The magnitude of this difference is similar to what has been observed at inter-specific level yet in the other direction. When the three populations were considered separately, polyPSGs were always more connected than non-polyPSGs, although the test was significant only for YRI (Supplementary Table 1).

Z_F scores and network degrees exhibit a significant positive correlation for all three populations (Table 1). Finally, comparison of Z_F scores for the four degree groups (low, medium-low, medium-high and high degree) using a non-parametric ANOVA showed significant differences in all three populations, as a result of higher Z_F scores at the highest degree groups, according to a linear trend test on ranks (Figure 2A–C; Table 1). These results were reproduced using the three positive selection statistics separately (*DH*, iHS and XP-CLR in all populations, except XP-CLR in CEU and CHB), and also using the Composite of Multiple Signals method (CMS) (Grossman et al. 2013, 2010) (Supplementary Note; Supplementary Figure 2; Supplementary Table 2). Furthermore, the observed trends remain significant when removing the putative effect of linkage disequilibrium among genes by using a subset of unlinked genes (see Supplementary Note; Supplementary Figure 3; Supplementary Table 3).

These analyses indicate that genes encoding proteins with a greater number of interactions in the human PIN are more likely to present signals of recent selective sweeps than those acting at more peripheral positions.

Correcting for several putative confounding factors and validations

A number of factors correlate with both network centrality and the likelihood of observing positive selection, and might thus be confounding our observations. In order to discard this possibility, we conducted a number of validations.

In agreement with previous results (Alvarez-Ponce & Fares 2012; Alvarez-Ponce 2012; Fraser et al. 2002; Hahn & Kern 2005; Vitkup et al. 2006), we observed that purifying selection is stronger in genes acting at the centre of the human PIN than at those acting at the periphery, regardless of whether it was measured from the ω ratio or the derived allele frequency (Figure 2E–F, Table 1, Supplementary Table). Purifying selection, through background selection (BGS), can produce signatures that can be confounded with positive selection by tests based on DNA polymorphism (Charlesworth et al. 1993), thus raising the possibility that our results could be a by-product of the distribution of purifying selection across the network. This effect, however, is unlikely to have affected our network-level analyses, given that we combined the results of different positive selection tests. Indeed, multivariate analyses confirmed that the relationship between network degree and positive selection was independent of purifying selection (Supplementary Note; Supplementary Figures 9–10; Supplementary Tables 6–7).

Factors such as gene expression level and breadth (tissue specificity), and the length of the encoded proteins, correlate with both network centralities and the likelihood of detecting positive selection (Duret & Mouchiroud 2000; Pál et al. 2006; Subramanian & Kumar 2004) and thus could also represent confounding factors. However, the relationship between network degree and all metrics of positive selection ($2\Delta l$ and Z_F) and purifying selection (ω and DAF) considered in this study remains unaltered when controlling for these parameters (Table 1; Supplementary Note; Supplementary Figure 4).

Our results might also be biased by the incompleteness and low quality of available interactomic data. However, similar results were obtained when a high-quality sub-network of BioGRID (Stark et al. 2011), or the Human Protein Reference Database (Keshava Prasad et al. 2009), were analysed (see Supplementary Note; Supplementary Figures 5–6; Supplementary Table 4), indicating that our observations are not a by-product of the quality of network data.

In addition to degree, which is a local measure of network centrality, we used two additional centrality measures that take into account the global position of proteins within the network: betweenness (the number of shortest paths between other proteins passing through a protein), and closeness (the inverse of the average distance to all other proteins in the network). Similar trends to those observed when using degree were observed in both cases (see Supplementary Note; Supplementary Figures 7–8; Supplementary Table 5).

Gene essentiality and impact of positive selection

To explore whether genes putatively under recent positive selection in our data set (i.e. affected by a *hard sweep* during recent human evolution) have important fitness effects, we classified the genes under study as viable or lethal using information from The Mouse Genome Database (Bult et al. 2008). Lethal genes present a significantly higher degree than viable genes (Mann-Whitney test; $P < 0.0001$; Table 2), in agreement with previous results (Fraser et al. 2002; Iyer et al. 2013; Jeong & Albert 2000). This demonstrates that, as expected, the phenotypic effect of a gene is highly associated with its position within the PIN (for a review, see (Olson-Manning et al. 2012)). We next compared the scores of positive selection on the PIN genes between the two groups (Table 2; Figure 3). As expected, lethal genes have significantly lower DAF and ω scores (Mann-Whitney test, $P < 0.0001$), indicating that they evolve under higher selective constraints. Moreover, they are more likely to be targeted by recent positive selection, since they exhibit significantly higher positive selection scores in the three human populations (Mann-Whitney test; $P = 0.0047$ in YRI, $P = 0.0009$ in CEU and $P = 0.0248$ in CHB). This indicates that recent positive selection targets genes with the highest effects on fitness. However, during mammal evolution, positive selection is more likely to act on viable genes: $2\Delta\ell$ scores are significantly higher for viable than for lethal genes (Mann-Whitney test; $P < 0.0001$). Similar results were obtained when using the “functional indispensability” score attributed to a specific gene according to its functional and evolutionary properties (Khurana et al. 2013) (Table 2).

DISCUSSION

The results presented here indicate that signatures of positive selection identified following two different methodological frameworks concentrate on different parts of the human PIN: when interrogating mammal divergence data, we observe that positive selection had a greater impact on genes with a lower network centrality, whereas recent, human-specific positive selection (as inferred from polymorphism data) has targeted preferentially genes occupying more central positions in the network. These patterns are independent of several potentially confounding factors.

The signatures of adaptation detected in this study through either a comparative genomics or population genetics approach might correspond to different kinds of changes at the sequence level, a problem with no obvious solution. The maximum-likelihood test used to detect positive selection using divergence data is powerful only in situations in which the gene has experienced recurrent selection events at the coding sequence; adaptation at regulatory sites, however, cannot be detected using this method. Therefore, positive selection during mammal evolution, as inferred here, should be viewed as sequence adaptations that alter the function of proteins recurrently across the mammalian phylogeny. Signatures detected in a genomic region using re-sequencing data, on the contrary, can correspond to unique selective sweeps (not necessarily recurrent) that occurred recently, either at the studied region or at a linked one (e.g., promoters and other regulatory regions). Thus, the putative signals of recent positive selection can be the result of variants that alter protein sequence, but are perhaps more likely to correspond to *cis*-regulatory variants, whose role in recent human evolution seems to have been pivotal (Enard et al. 2014; Fraser 2013). Since protein-coding genes are particularly constrained at the core of the interactome, their regulatory regions may provide the necessary pool of variation for adaptation. Therefore, recent positive selection events detected using polymorphism data are likely to correspond to adaptation through changes in expression patterns (gene expression level or regulation), while selective events detected through divergence analysis may mostly correspond to changes in protein function.

The higher centrality of essential genes suggests that the centre of the network may roughly correspond to the most important, influential and pleiotropic genes of the system. Certain evolutionary mechanisms may promote a higher adaptability at the centre of the network where the effects of genes on fitness are important, whereas others may promote a higher incidence of positive selection at the periphery. On the one hand, in the 1930s, Ronald Fisher formulated the hypothesis that mutations with large effects on phenotype, such as those with highly pleiotropic effects, should often be deleterious (Fisher 1930; Orr 2005). In agreement with this hypothesis, purifying selection is stronger on genes acting at the centre of molecular networks (Fraser et al. 2002; Hahn & Kern 2005; Vitkup et al. 2006; Alvarez-Ponce & Fares 2012; Alvarez-Ponce 2012) (but see (Hahn & Kern 2005; Jordan et al. 2003)), a pattern that we have confirmed analysing both divergence and polymorphism data. Since purifying selection quickly removes a high fraction of new mutations at these genes, one would expect positive selection to rarely act on them because of their reduced variability (Olson-Manning et al. 2012). Therefore, we may expect positive selection to target more frequently the periphery of the network. On the other hand, the action of positive selection at genes occupying the centre of the network is not to be discarded. Indeed, signatures of positive selection are frequent at genes occupying relatively important positions in a number of metabolic and signal transduction pathways (Dall'Olio et al. 2012; Luisi et al. 2012; Flowers et al. 2007; Olson-Manning et al. 2013).

Simulation analyses of hypothetical metabolic pathways have shown that, when pathways are far from the fitness optimum, positive selection first targets enzymes lying at the upstream part, and at the branch points of the pathway, which exert greater control over metabolic flux. In turn, when the system approaches its optimum, positive selection tends to concentrate on enzymes with less flux control, and purifying selection constrains the evolution of upstream and branch-point enzymes (Rausher 2012; Wright & Rausher 2010). These observations match the expected pattern of diminishing returns, first proposed by Ronald Fisher in his Geometric Model of Adaptation (Fisher 1930) (FGM) which states that selection tends to act progressively more often on mutations with smaller phenotypic effects as populations approach a peak in the adaptive landscape. A mutation's effect is measured as a function of both its effect on a given trait and the numbers of phenotypes that are jointly modified by the mutation (pleiotropic effect) (Fisher 1930; Orr 2005), and theoretical development is undergoing in order to relate the FGM to information on protein-protein interaction networks (e.g. see (Martin 2014)). According to the FGM, events of selection are more likely to be observed on mutations with small phenotypic effects (following a geometric distribution), whereas positive selection on mutations with large effects is most likely to occur during the first steps of adaptation.

The results described in the present study can be understood according to both the FGM and the different kinds of advantageous changes detected at the sequence level. Indeed, when focusing at large evolutionary time-scale, i.e. during mammal evolution, we are studying the whole process of adaptation acting exclusively on protein-coding genes that made the species overall fit. Therefore, according to the geometric distribution of the probability of a mutation to be favourable, it is more likely to detect events of adaptation acting on genes with lower effect on fitness, that is genes encoding proteins with less interacting partners. On the other hand, when focussing at much shorter evolutionary time-scale, i.e. during recent human evolution, we are studying the recent adaptation of human populations to a wide range of new environments (e.g. the Mesolithic-Neolithic transition, the human diaspora across the world, etc.). We speculate here that events of strong recent positive selection, as inferred from polymorphism data assuming the *hard sweep* model, mainly targeted *cis*-regulatory regions of genes with important effects on fitness in order to efficiently tune some specific phenotypes without affecting the whole protein interaction map.

In summary, even though the interactome is a raw simplification of the processes that take place within the cell, it contains valuable information on the relative role of the many gene products that interact to sustain life. The position occupied by a protein within an interaction network provides useful information –albeit incomplete– on the phenotypic effects of mutations arising at the encoding gene. Interestingly, we have shown that using this information can also help to better understand the impact of positive selection acting on protein-coding genes and their *cis*-regulatory region. Although network centrality used alone remains a modest predictor of the impact of positive selection, it could be included in an integrative biology approach to shed light on adaptive processes acting on the genome. The present study also underscores the fact that the relationship between positive selection and network position is more complex than previously recognised, when positive selection was suggested to mostly act at the network periphery. Indeed, the discovery of the rules governing network evolution may shed light on the dynamics of the evolutionary processes driven by selection. Notably, the distribution of selective events in a large-scale protein-protein interaction network described in the present study, which relies on extensive sequence data, can be understood in the light of the Fisher's Geometric Model of Adaptation. Particularly, results presented here show that the prime matter for innovation is also to be found in genes, or in their *cis*-regulatory region, encoding proteins with high network centrality, meaning that they have more pleiotropic effects, are more indispensable and in general are at the basis of strong changes as a result of mutations during the initial high-risk high-gain phase of the adaptation process.

ACKNOWLEDGEMENTS

We thankfully acknowledge valuable discussion and corrections from Diego A. Hartasanchez, David A. Hughes and Arcadi Navarro. We thank the National Institute of Bioinformatics (http://www.inab.org) for computational support. This work was funded by the “Ministerio de Ciencia y Tecnología” (Spain) [grant BFU2010-19443; subprogram BMC], and the “Direcció General de Recerca, Generalitat de Catalunya (Grup de Recerca Consolidat 2009 SGR 1101)” awarded to JB. PL is supported by a Ph.D. fellowship from “Acción Estratégica de Salud, en el marco del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2008–2011” from Instituto de Salud Carlos III. D.A.-P. is a “Juan de la Cierva” fellow from the “Ministerio de Economía y Competitividad” (Spain) [JCI-2011-11089]. MAF is supported by a Principal Investigator grant from Science Foundation Ireland (12/IP/1673) and a project from the “Ministerio de Economía y Competitividad” [grant number BFU2012-36346].

REFERENCES

Agrafioti I et al. 2005. Comparative analysis of the Saccharomyces cerevisiae and Caenorhabditis elegans protein interaction networks. BMC Evol. Biol. 5:23.

Alvarez-Ponce D. 2012. The relationship between the hierarchical position of proteins in the human signal transduction network and their rate of evolution. BMC Evol. Biol. 12:192.

Alvarez-Ponce D, Fares M a. 2012. Evolutionary rate and duplicability in the Arabidopsis thaliana protein-protein interaction network. Genome Biol. Evol. 4:1263–74.

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. 18:1585–1592.

Bult CJ, Eppig JT, Kadin J a, Richardson JE, Blake J a. 2008. The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res. 36:D724–8.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics. 134:1289–303.

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. Genome Res. 20:393–402.

Codoñer FM, Fares MA. 2008. Why Should We Care About Molecular Coevolution ? Bioinformatics. 4:29–38.

Cork JM, Purugganan MD. 2004. The evolution of molecular genetic pathways and networks. Bioessays. 26:479–84.

Cui Q, Purisima E, Wang E. 2009. Protein evolution on a human signaling network. BMC Syst. Biol. 3:21.

Dall’Olio GM et al. 2012. Distribution of events of positive selection and population differentiation in a metabolic pathway: the case of asparagine N-glycosylation. BMC Evol. Biol. 12:98.

Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 15:330–40.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. 17:68–74.

Enard D, Messer PW, Petrov D a. 2014. Genome-wide signals of positive selection in human evolution. Genome Res. 24:885–895.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics. 155:1405–13.

Fisher RA. 1930. *The Genetical Theory of Natural Selection*. Clarendon . Oxford.

Fisher RA. 1930. *The Genetical Theory of Natural Selection*. Oxford Univ Press: Oxford.

Flicek P et al. 2010. Ensembl's 10th year. Nucleic Acids Res. 38:D557–62.

Flowers JM et al. 2007. Adaptive evolution of metabolic pathways in Drosophila. Mol. Biol. Evol. 24:1347–54.

Fraser HB. 2013. Gene expression drives local adaptation in humans. Genome Res. 23:1089–96.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary Rate in the Protein Interaction Network. Science. 296:750–752.

Grossman SR et al. 2010. A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. Science (80-.). 327:883–886.

Grossman SR et al. 2013. Identifying recent adaptations in large-scale genomic data. Cell. 152:703–13.

Hahn MW, Conant GC, Wagner A. 2004. Molecular evolution in large genetic networks: does connectivity equal constraint? J. Mol. Evol. 58:203–211.

Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol. Biol. Evol. 22:803–806.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleric Acids Symp. Ser. 41:95–98.

Iyer S, Killingback T, Sundaram B, Wang Z. 2013. Attack robustness and centrality of complex networks. PLoS One. 8:e59613.

Jeong H, Albert R. 2000. The large-scale organization of metabolic networks. Nature. 407:651–654.

Jordan IK, Wolf Y, Koonin E. 2003. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol. Biol. 3:1.

Karolchik D, Hinrichs AS, Kent WJ. 2009. The UCSC Genome Browser. Curr. Protoc. Bioinforma. Chapter1:Unit1:4.

Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. 16:980–9.

Kersey PJ et al. 2012. Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. Nucleic Acids Res. 40:D91–7.

Keshava Prasad TS et al. 2009. Human Protein Reference Database--2009 update. Nucleic Acids Res. 37:D767–72.

Khurana E, Fu Y, Chen J, Gerstein M. 2013. Interpretation of Genomic Variants Using a Unified Biological Network Approach. PLoS Comput. Biol. 9.

Kim PM, Korbel JO, Gerstein MB. 2007. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. Proc. Natl. Acad. Sci. 104:20274–20279.

Kosiol C et al. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet. 4:e1000144.

Liao B, Zhang J. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. Proc. Natl. Acadamy Sci. 105:6987–6992.

Lovell SC, Robertson DL. 2010. An integrated view of molecular coevolution in protein-protein interactions. Mol. Biol. Evol. 27:2567–75.

Luisi P et al. 2012. Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations. Mol. Biol. Evol. 29:1379–92.

MacArthur DG et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 335:823–828.

Martin G. 2014. Fisher's Geometrical Model Emerges as a Property of Complex Integrated Phenotypic Networks. Genetics. 197:237–55.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 148:929–936.

Olson-Manning CF, Lee C-R, Rausher MD, Mitchell-Olds T. 2013. Evolution of flux control in the glucosinolate pathway in Arabidopsis thaliana. Mol. Biol. Evol. 30:14–23.

Olson-Manning CF, Wagner MR, Mitchell-Olds T. 2012. Adaptive evolution: evaluating empirical support for theoretical predictions. Nat. Rev. Genet. 13:867–77.

Orr HA. 2005. The genetic theory of adaptation: a brief history. Nat. Rev. Genet. 6:119–27.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat. Rev. Genet. 7:337–48.

Pérez-Bercoff Å, Hudson CM, Conant GC. 2013. A conserved mammalian protein interaction network. PLoS One. 8:e52581.

Rausher MD. 2012. The evolution of genes in branched metabolic pathways. Evolution. 67:34–48.

Schaffner SF et al. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15:1576–183.

Scheinfeldt LB et al. 2009. Population genomic analysis of ALMS1 in humans reveals a surprisingly complex evolutionary history. Mol. Biol. Evol. 26:1357–67.

Smith JM, Smith NH. 1996. Synonymous nucleotide divergence: what is "saturation"? Genetics. 142:1033–6.

Stark C et al. 2011. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 39:D698–704.

Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics. 168:373–81.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–95.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56:564–77.

Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? Genome Res. 16:702–712.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature. 491:56–65.

Vitkup D, Kharchenko P, Wagner A. 2006. Influence of metabolic network structure and function on enzyme evolution. Genome Biol. 7:R39.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:e72.

Wagner A. 2012. Metabolic networks and their evolution. Adv. Med. Biol. 751:29–51.

Wright KM, Rausher MD. 2010. The evolution of control and distribution of adaptive mutations in a metabolic pathway. Genetics. 184:483–502.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22:1107–1118.

Zeng K, Shi S, Wu C-I. 2007. Compound tests for the detection of hitchhiking under positive selection. Mol. Biol. Evol. 24:1898–908.

FIGURE LEGENDS

Figure 1. Distribution of genes with putative signatures of positive selection within the Protein–Protein Interaction Network. Z_F and $2\Delta\ell$ were used to estimate the likelihood of having evolved under positive selection in human populations and in mammals, respectively. **A.** Average degrees (number of interactions) for genes with and without signatures of positive selection. We represent the mean of centrality measure ± one standard error for the genes with a putative signal of positive selection (in red) and the other genes (in blue). The significance of the differences between the mean of both groups was assessed through 10,000 permutations. Asterisks represent significant differences. *: $P < 0.05$; **: $P < 0.01$;. **B.** Human protein–protein interaction network with genes with signatures of positive selection according to divergence data ($P < 0.05$ estimated from $2\Delta\ell$) represented in red. **C.** Human protein–protein interaction network with genes with signatures of positive selection according to polymorphism data represented in red.

Figure 2. Impact of natural selection among groups of genes divided according to degree quartiles. Genes were divided into four groups according to the degree quartiles. The median positive selection score ± one median absolute deviation for each group is represented in the *y*-axis. Z_F and $2\Delta\ell$ scores were used to estimate the likelihood of positive selection in human populations and in mammals, respectively. DAF and ω were used to estimate the impact of purifying selection in human populations and in mammals, respectively. A non parametric ANOVA analysis was performed to contrast whether the medians of the scores are equal across the groups. A trend test on ranks was also carried out to test for a linear relationship between the four groups (encoded from 1 to 4) and natural selection scores. A Tukey's honestly significant difference test was further applied to test for all pairwise differences. Significantly different pairs are marked with asterisks according to the level of significance. *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.

Figure 3. Comparison of the impact of natural selection between essential and non-essential genes. We performed a Mann-Whitney test to compare the positive selection scores between genes that are lethal (essential, in red) and viable (non essential, in blue) when knocked out in mice (data from the Mouse Genome Database (Bult *et al.* 2008);

"MRK_Ensembl_Pheno.rpt" file downloaded on 7 October 2010). Z_F and $2\Delta\ell$ scores were used to estimate the likelihood of positive selection in human populations and in mammals, respectively. DAF and ω were used to estimate the impact of purifying selection in human populations and in mammals, respectively. In order to put all the scores within the same scale the mean standardized scores are plotted (standardized scores were calculated by subtracting the mean and dividing by the standard deviation). Significant differences between lethal and viable genes pairs are marked with asterisks. *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.

TABLES

Table 1. Relationship between degree and the impact of natural selection

		Positive selection				Purifying selection	
		YRI	CEU	CHB	Mammals	Humans	Mammals
Spearman correlation[a]	ρ	0.0501	0.0410	0.0471	-0.0490	-0.0879	-0.2039
	P-value	1.11×10^{-5}***	0.0004***	3.48×10^{-5}***	0.0002	4.51×10^{-16}***	6.91×10^{-56}***
Partial Spearman correlation[b]	ρ	0.0451	0.0326	0.0374	-0.0340	-0.0668	-0.1697
	P-value	0.0001***	0.0059**	0.0015**	0.0107*	2.38×10^{-09}***	3.08×10^{-37}***
Non parametric ANOVA[c]	*F*	5.324	5.844	5.074	3.780	18.027	77.82
	P-value	0.0012**	0.0006***	0.0016**	0.0101**	1.17×10^{-11}***	2.37×10^{-49}***
Trend test on ranks[c]	*F*	15.88	12.14	14.12	10.23	52.564	229.3
	P-value	6.79×10^{-5}***	0.0005**	0.0002***	0.0014**	4.53×10^{-13}***	7.59×10^{-51}***
Partial non parametric ANOVA[b,c,]	*F*	2.731	3.149	2.080	2.537	6.353	51.87
	P-value	0.0423*	0.0240*	0.1006	0.0548	0.0003***	4.65×10^{-33}***
Partial trend test on ranks[b,c]	*F*	7.794	2.360	5.107	6.281	16.48	153.5
	P-value	0.0053**	0.1246	0.0239*	0.0122*	4.97×10^{-5}***	8.70×10^{-35}***

[a] Spearman correlation between degree and selection scores (Z_F for positive selection in YRI, CEU and CHB populations; $2\Delta\ell$ for positive selection in mammals; DAF for purifying selection in humans; and ω for purifying selection in mammals). High Z_F and $2\Delta\ell$ scores indicate a higher probability of having evolved under positive selection as inferred from polymorphsim and divergence data, respectively. Low DAF and ω scores indicate higher evolutionary constraint estimated from polymorphism and divergence data, respectively.

[b] In order to test for an association between degree and natural selection scores while controlling for putatively confounding factors, we applied a linear regression between the selection scores and protein length, expression level and breadth. The linear regression residuals were then used to perform the Spearman's correlation analysis, the non parametric ANOVA and the linear trend on ranks test.

[c] Non parametric ANOVA and linear trend tests on ranks performed to contrast whether the score used as a proxy of natural selection are equal across the degree groups. *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.

Table 2. Association between gene essentiality and degree and the impact of natural selection

	Lethal *vs*. viable genes[a]			Indispensability score[b]	
	Mean lethal	Mean viable	*P*-value	ρ	*P*-value
Degree	14.55	7.048	6.62×10^{-52}***	0.2311	3.03×10^{-107}***
Positive selection in YRI[c]	6.419	6.154	0.0047**	0.0473	4.34×10^{-05}***
Positive selection in CEU[c]	6.754	6.350	0.0009***	0.0695	2.00×10^{-09}***
Positive selection in CHB[c]	6.712	6.423	0.0248*	0.0380	0.0010**
Positive selection in mammals[d]	1.830	2.270	2.03×10^{-08}***	-0.1157	3.62×10^{-25}***
Purifying selection in humans[e]	0.1041	0.1109	4.66×10^{-08}***	-0.1131	5.14×10^{-25}***
Purifying selection in mammals[f]	0.0768	0.1160	3.70×10^{-29}***	-0.2600	6.67×10^{-89}***

[a] Mann-Whitney test to compare the degree or the natural selection score between genes that are essential and genes that are not essential, i.e. lethal and viable when knocked out in mice, respectively (data from the Mouse Genome Database (Bult et al. 2008) "MRK_Ensembl_Pheno.rpt" file downloaded on 7 October 2010).

[b] Spearman's correlation analysis to test for the relationship between degree or the natural selection score and the functional indispensability score (Khurana et al. 2013).

[c,d] High Z_F and $2\Delta\ell$ scores indicate a higher probability of having evolved under positive selection during human and mammal evolution, respectively.

[e,f] Low DAF and ω scores indicate higher selective constraints during human and mammal evolution, respectively.

*: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.

FIGURES

Figure 1

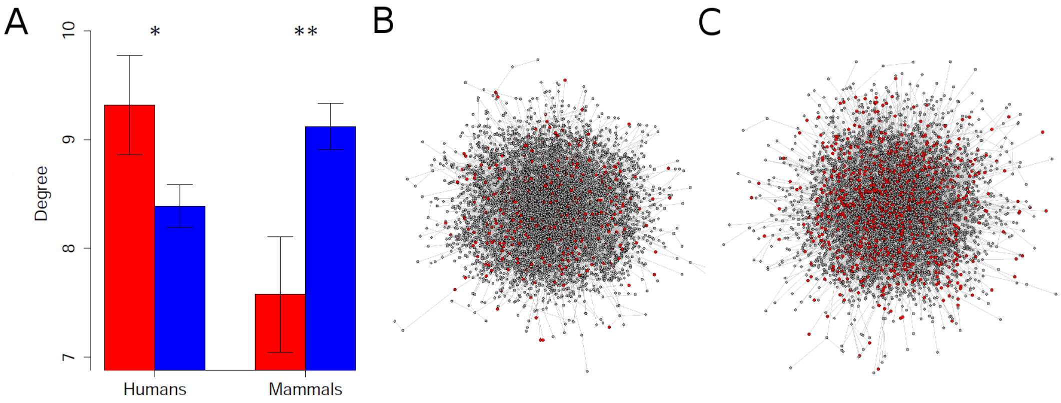


Figure 2

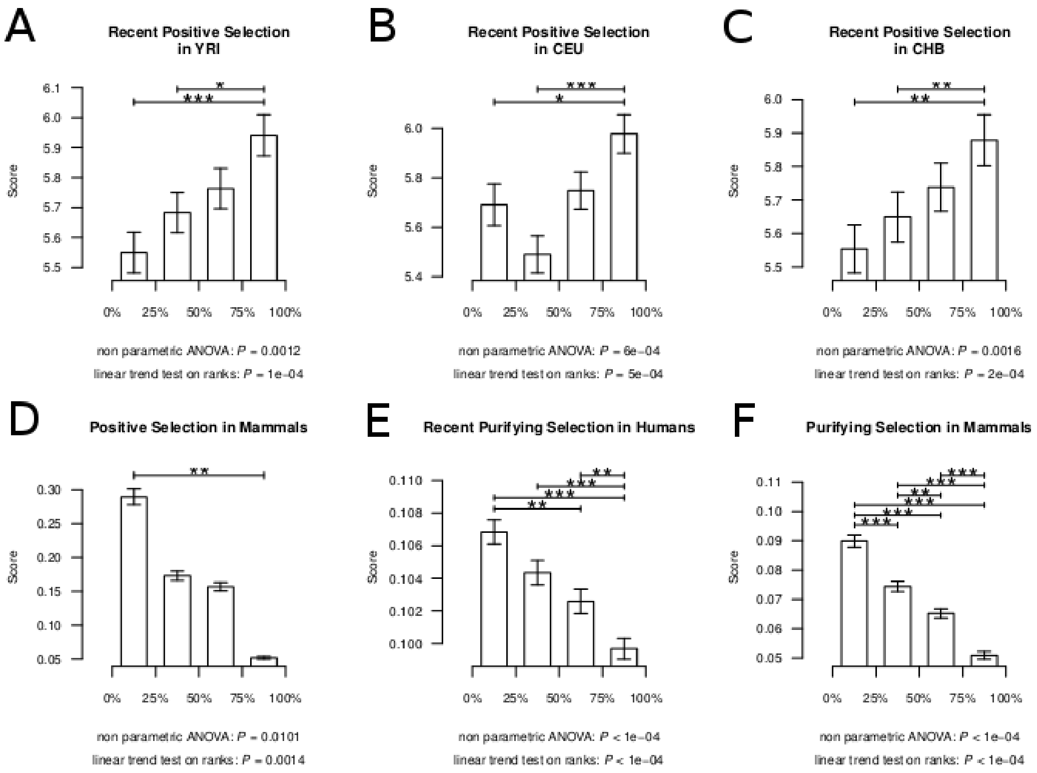
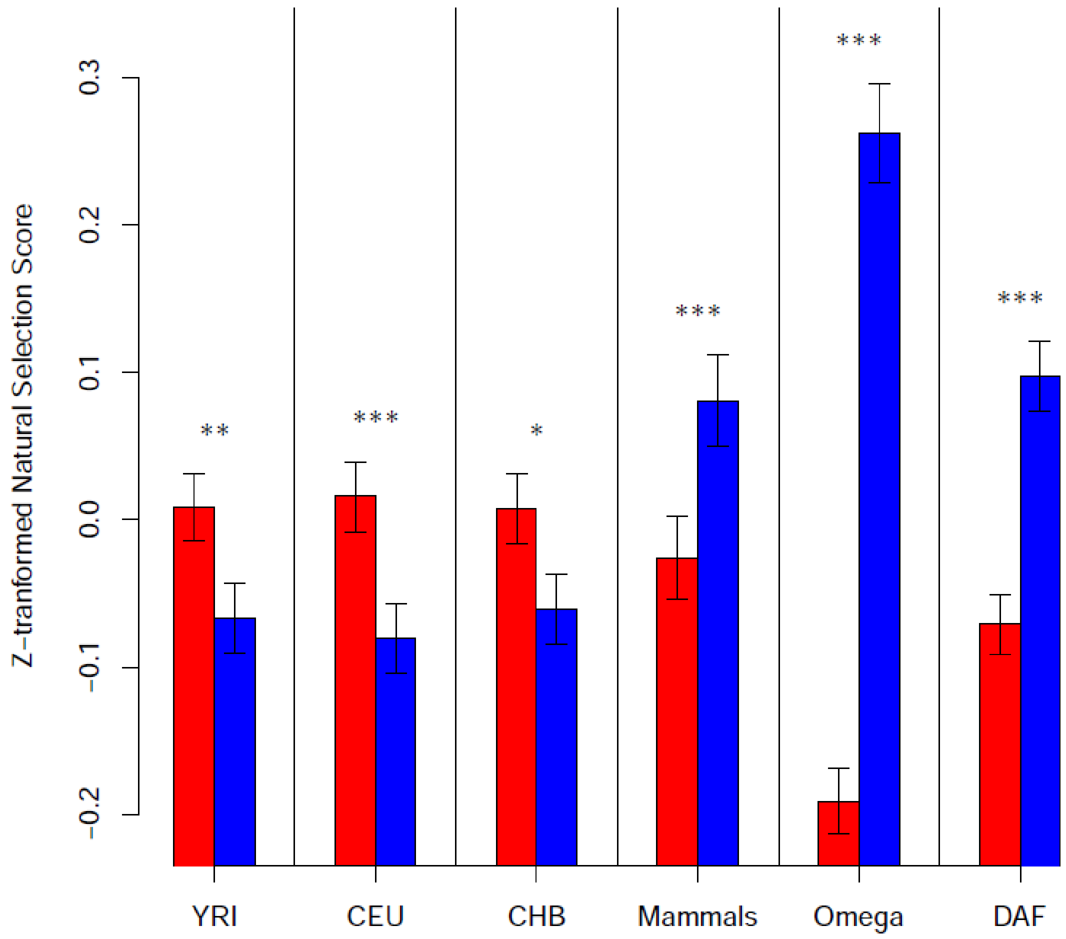

A
Recent Positive Selection in YRI
Score
6.1
6.0
5.9
5.8
5.7
5.6
5.5
*
***
0%
25%
50%
75%
100%
non parametric ANOVA: *P* = 0.0012
linear trend test on ranks: *P* = 1e−04
B
Recent Positive Selection in CEU
Score
6.0
5.8
5.6
5.4
***
*
0%
25%
50%
75%
100%
non parametric ANOVA: *P* = 6e−04
linear trend test on ranks: *P* = 5e−04
C
Recent Positive Selection in CHB
Score
6.0
5.9
5.8
5.7
5.6
5.5
**
**
0%
25%
50%
75%
100%
non parametric ANOVA: *P* = 0.0016
linear trend test on ranks: *P* = 2e−04
D
Positive Selection in Mammals
Score
0.30
0.25
0.20
0.15
0.10
0.05
**
0%
25%
50%
75%
100%
non parametric ANOVA: *P* = 0.0101
linear trend test on ranks: *P* = 0.0014
E
Recent Purifying Selection in Humans
Score
0.110
0.108
0.106
0.104
0.102
0.100
**
***
***
**
0%
25%
50%
75%
100%
non parametric ANOVA: *P* < 1e−04
linear trend test on ranks: *P* < 1e−04
F
Purifying Selection in Mammals
Score
0.11
0.10
0.09
0.08
0.07
0.06
0.05
***
***
**
***
***
***
0%
25%
50%
75%
100%
non parametric ANOVA: *P* < 1e−04
linear trend test on ranks: *P* < 1e−04

Figure 3


Z–tranformed Natural Selection Score
0.3
0.2
0.1
0.0
−0.1
−0.2
**
***
*
***
***
***
YRI
CEU
CHB
Mammals
Omega
DAF

Part III

Discussion

Chapter 7

DISCUSSION

> Doubt is not a pleasant condition, but certainty is absurd.
>
> *Letter to Frederick II of Prussia*
> VOLTAIRE

As outlined in the objectives (Chapter 2), the aim of the work presented in this thesis was to explore how integrating information on gene networks could shed light on adaptive processes at the molecular level. For that purpose, first two gene-level studies have been described: (i) a gene-candidate study to understand the role of positive selection on a region encompassing a specific gene of interest; and (ii) a genome-wide scan for positive selection to identify signals of positive selection across the genome with a functional study to follow-up an outstanding signal thus detected. These two studies are representative examples of traditional analyses performed to understand the impact of adaptive evolution at a particular genomic region encompassing a protein-coding gene. Then, adaptive evolution of genes was studied within a gene-network framework, that is by integrating information on the physical interactions in which are involved the encoded proteins. Two gene-network scales have been considered: (i) a gene-network representing a specific biological pathway, and (ii) one including all the known physical protein-protein in-

teractions occurring in the organism. These two studies come within the scope of an emerging field which could be designed as *evolutionary system biology*. They represent one of the first attempts to describe how the impact of positive selecion on protein-coding genes is related to biological network in which are involved the encoded proteins.
This chapter will first provide a discussion on the strengths and drawbacks for each of the four studies. Second, the importance of the network framework to study adaptive selection will be discussed. Finally, the underlying challenges, and potential perspectives for the *evolutionary system biology* field will be examined.

7.1 General remarks on the four studies described.

7.1.1 Single gene studies.

As already stressed in Chapter 1, most of the studies of adaptive evolution at molecular level have focused on its impact on single genes, as in the two articles presented in Chapters 3 and 4 which illustrate the gene-candidate and genome-wide scan approaches, respectively.

Gene-candidate approach.
The gene-candidate study of the region encompassing *VKORC1* gene shows that detecting the advantageous variant in a region that have undergone a selective event is particularly complicated. This study was performed using the HGDP genotype data (see Section 1.2.2) and allowed to better localize geographically the selective sweep, namely it occurred only in East Asian populations and the signal is shared among all the studied populations in this area. However, it was impossible to precisely pinpoint a putative mutation driving the selective event. A selective sweep leaves an extreme pattern of LD, and, although such molecular pattern is useful to contrast whether a genomic region evolved adaptively, most of

the variants within the region exhibit similar scores for the methods used to detect positive selection. Specifically, an in-depth analysis of the region encompassing *VKORC1* allowed to restrict to 45 Kb the region in which is located the putative advantageous variant. However, four genes are located within this 45 Kb region, thus different functional variants are good candidates for being the target of positive selection. The genotype data from HGDP does not give much precision since the SNPs present on the genotyping array are *tag*-SNPs, reducing the variant density for cost purposes. Therefore, additional SNPs in the studied region were genotyped to increase the density of the variants to be interrogated. Such effort was vain, and if more money could have been invested in the study, sequencing the 45Kb region for the East Asian individuals would have been the best solution. Alternatively, one can study the 1000 Genomes re-sequencing data (see Section 1.2.2) which presents a higher variant density and is free of *ascertainment bias* (see Section 1.5.1). Such analysis is presently performed in the hope of identifying a single variant with an extreme score for several methods to detect positive selection. All together, although the study of the region encompassing *VKORC1* did not reach its initial goals, it provides very useful insights for follow-up studies. There is no doubt that the region has evolved under an adaptive regime, and therefore, a particularly variant within those 45 Kb had enough phenotypic relevance to be selected for in East Asia. Such information might be very useful to anyone with an hypothesis on a phenotype specific to East Asian populations that provided a fitness advantage to the past environment in this geographic area, and in which one of the four genes is involved.

Genome-wide scan approach

The study presented in Chapter 4 is a good example of the alternative to the gene-candidate approach: first the genome is interrogated for positive selection without any *a priori* assumption on the adaptive phenotype, in order to, in turn, pick a region with a putative signal of selection for which an *a posteriori* assumption can be tested. However, the study design here is unique. Indeed, instead of studying an unique population, or different populations each with an assumed independent history, as usually done (e.g. see [70, 71, 119, 154]), here three populations were interrogated simultaneously in order to take profit of their history. Namely, the Rroma/Gypsie population was of particular interest. This population migrated one thousand years ago from North India to the Balkans [141]. In order to have a reference population sharing the selective pressures faced by the Rroma/Gypsies after their migration, a population with European ancestry from Romania was included. Moreover, a Northwest Indian population was used as a reference population sharing most of its history, thus sharing most of the genetic background, with the Rroma/Gypsies. The aim of this study was then to scan the genome for shared signals of positive selection in Rroma/Gypsies and Romanians but absent in Nortwest Indians, assuming that such signals must have emerged from the adaptation to the European environment from different genetic backgrounds. The triangular design of the study allowed then to (1) reduce the FPR in the signals of positive selection detected in two populations sharing the same environment; (2) control for *genetic drift* that occurred before the Rroma/Gypsies emigration using the Northwest Indian population, and (3) be able to infer the selective pressures driving the signal, i.e. any environmental variable that appeared in the laste thousand years in Europe and not in Northwest India. The last point is particularly interesting since in their recent history, European populations faced severe epidemic events (plague, influenza, smallpox, etc.) arguably exercising important selective pressures on the immune system. Several genomic regions have been identified according to the criteria aforementioned. A particular one was blindingly obvious to follow-up: a region containing the gene cluster TLR1/TLR6/TLR10 encoding for the Toll-Like Receptors 1, 6 and 10, respectively. This region had already been described to have undergone a selective event in non-African population in a study which, however, did not include any Central South Asian population [266].

The TLR family is recognized as a key family of innate immunity. After recognition of their ligand(s), TLRs transduce the signalling responses to activate the innate immunity effector mechanisms and the subsequent development of adaptive immunity (for a review see [267]). In humans, ten members compose the TLR family (*TLR1-10*) [267]). They can be classified according to their subcellular distribution: *TLR3, TLR7-9* are located in intracellular compartments (typically in endosomes) whereas *TLR1-2* and *TLR4-6* are mostly expressed on the cell surface [267]. Intracellular or cell surface TLRs have different kinds of agonists: the intracellular TLRs sense nucleic acid-based agonists, and are typically involved in viral recognition, while the cell surface TLRs detect other products such as glycolipids, lipoleptides and flagellins present in bacteria, parasites and fungi [267]. Although functional roles of TLRs are well described, little is known for *TLR10* which is expressed on cell surface. The study described here allowed to broaden our knowledge on this specific gene. Since the *TLR1/TLR6/TLR10* gene cluster has undergone convergent positive selection (but see below) in two populations with different genetic backgrounds and which have lived for the last thousand years in the same environment, one particular selective pressure from this environment must be the driving force. A direct assumption one can make is that *Yersinia pestis*, the agent of plague, has played such role since plague had be the most devastating epidemics in this specific area at that time. Functional analyses allowed to confirm that *TLR1*, *TLR6* and *TLR10* contain genetic variation that modulate *Y. pestis*-induced immune responses. Namely, *Y. pestis* is known to induce proinflammatory cytokines (e.g. TNF, IL-1β and IL-6) that are are modulated by specific combination of variants in the *TLR1/TLR6/TLR10* gene cluster. Interestingly, TLR10 receptors inhibitated the IL-6 induction by IL-1, suggesting that TLR10 may act as an inhibitor of the IL-1 family cytokines. However, the same immunological analysis yet performed using *Y. pseudotuberculosis* showed similar results to that with *Y. pestis*, suggesting that these TLRs may also respond to other deadly bacteria and other diseases might have been the causes of the selective signal observed in this genomic region. However, *Y. pseudotuberculosis* was used because it seems to be the ancestror of *Y. pestis* and the observed reaction to *Y. pestis* was much stronger.

In the article, the convergent selective event affecting *TLR1/TLR6/TLR10* in both European and Rroma/Gypsie populations was suggested. Although such parallel adaptation has been proved to be potentially common [139], one may thing that the shared specific molecular pattern observed at the *TLR1/TLR6/TLR10* and detected in both populations could result from recent admixture between the two populations, i.e. interbreeding between the two populations that were isolated until very recently. "Recent" here refers to admixture that would have occurred during or after the completion of the selective sweep in one population. Thus, the adapted population would provided the adaptive variation to the population with which it admixed. Such scenario has been observed in Tibetans who are a mixture of ancestral populations related to Sherpas and Han Chinese. The ancestral population related to Sherpa was already adapted to the hypoxic environment due to the high altitude of the Tibetan plateau. Jeong *et al.* showed that the Tibetans present the same genetic variants than the Sherpas in both *EGLN1* and *EPAS1* genes conferring better fitness to hypoxia, and could demonstrate that migrants from low altitude acquired the adaptive alleles from the highlanders [160]. Although such hypothesis is evolutionary fascinating, it was not possible to test it on European and Rroma/Gypsie populations. Indeed, the study was performed on the Immunochip which presents a very heterogeneous SNP density across the genome, thereby preventing one to accurately phase the data to obtain information on haplotype variation from genotypes, while such information is essential to perform any admixture inference. Although this scenario has not been tested, this study provides striking results on putative parallel adaptation and illustrates how a genome-wide scan for positive selection can allow to point a specific genomic region for follow-up functional study. The functional study, in turn, sheds light on the immune response to pathogens, and broaden our knowledge on the role played by TLR family members, particularly the poorly characterized *TLR10* gene.

7.1.2 Network-level analyses.

Two analyses of the impact of positive selection in protein-protein interaction networks have been presented in Chapters 5 and 6, the former representative of a particular biological pathway, the later representative of the whole sets of physical interactions among proteins occurring in the human organism. These two biological scales have their own drawbacks and advantages.

Pathway-level.
In Chapter 5, the distribution of signals of positive selection, as inferred from human polymorphism data, within the network representing the Insulin/TOR transduction signalling (IT) pathway has been studied. It was the first published study of this kind. Working at the pathway-level allows good confidence on the proteins and their interactions since such information is retrieved from the literature, provided one is able to define biologically relevant boundaries of the studied pathway (see Section 1.8.1 for details). Such direct manual curation of the information allows to consider different kinds of interactions involved in the biological pathway (see Section 1.7.2). Namely, three types of interactions are participating to the IT pathway: physical protein-protein, metabolic, and transcriptional interactions. Network centrality metrics were computed taking into account the different modes of interaction or only the physical protein-protein interactions, yet the results pointed towards the same direction in both cases. On the other hand, the completeness of the information on protein interactions available in the literature may suffer a bias towards historically more studied proteins. Particularly, paralogous copies resulting from recent gene duplication events have been difficult to identify given the low divergence between the resulting copies, which might have been treated as a single copy during genome assembly. Therefore, one expects to find more information in the literature for the original copy than for its paralogs. To circumvent this issue, the network representing the IT pathway was build from genes known to be involved in the pathway according to the literature or their close paralogs that cluster within these genes in phylogenetic trees. Genes that are known not to play any role, despite having paralogs that are actually involved, were excluded. This illustrates how working at small-biological scale may suffer from interaction ascertainment bias because of incomplete knowledge from specific biochemistry experiments. However, it also illustrates that accurate strategies may exist to circumvent such bias.
Another issue is the accuracy of the interaction annotation in available

pathway databases. For example, Dall'Olio *et al.* manually curated the Asparagine N-linked glycosylation pathway. Although this pathway performs one of the most important forms of protein post-translational modification in eukaryotes and is one of the first metabolic pathways described at a biochemical level, its annotatyion in public databases such as Reactome [168] remained poorly accurate, and several correction had to be performed [268]. The authors advice to use Reactome, rather than other widely used databases (e.g. KEGG [167]) as it operates in a open-source fashion, encouraging feedback from its users and, thereby making it easier to keep the annotation of this pathway updated with future knowledge [268]. The study of the IT pathway did not suffer such caveat since the network was built directly from the literature instead of relying on any database.

Another challenge when studying a specific pathway using polymorphism data arises from the fact that the genes that are involved may be located in either autosomal or sexual chromosome. However, most methods designed to detect positive selection exhibit different sensitivity for sexual and autosomal variants. This complicates the comparison of signals of selection between genes located in sexual and autosomal chromosomes. Usually, while performing a genome-wide scan for positive selection sexual and autosomal chromosomes are studied separately. However, when studying a specific pathway, genes in sexual chromosomes have to be taken into account somehow to perform a proper network-level analysis of the detected signals of selection. To circumvent this issue, the strategy followed in the present study was to simply remove such genes from the network-level, yet obviously accounting for them to calculate network centrality metrics.

When inferring simultaneously the potential impact of positive selection on many genes, one can not perform an in-depth analysis of the signals for each gene such as for single gene studies. A natural way to overpass this issue is to compute for each gene and each positive selection statistics a combination score. In this study, since the genotype data used comes from the HGDP dataset and thus mainly includes *tag*-SNPs (see Section 1.2.2), the Fisher's combination test of the empirical *P*-values for SNPs located in a gene (thus assumed to be independent one form the others) was used as a summary statistics. Moreover, since the number of genes in the study remains relatively limited (~ 70 genes), a visual inspection of the signals was still possible as illustrated in Figure 7.1.

One advantage inherent of a pathway-level analysis is that one can integrate the information on positive selection on each gene together in order to contrast whether the studied endophenotype, i.e. the specific biological function performed by the pathway, has evolved under an adaptive regime. A hypergeometric test performed in each geographic region represented in the HGDP dataset (see Section 1.2.2) suggests that the transduction of Insulin and mTOR protein has been tuned by positive selection in recent human history in West Eurasia populations. Moreover, the results from the network-level analysis point to the fact that such adaptive evolution has occurred through specific selective events in genes located at central positions in the pathway, thus on genes potentially exerting a higher influence on the IT pathway function. Other studies described similar results in other pathways, yet the feature considered to define "genes potentially exerting a higher influence" remains somewhat vague (see Section 1.8.1 and further discussion in Section 7.2.1).

Interactome level.

In order to gain more insights from the potential of considering network topology to understand the impact of positive selection on protein-coding genes, a study at broader scale was performed (Chapter 6). In this study, the whole human interactome was considered, that is, the whole set of identified physical protein-protein interactions occurring in the human organism. As already mentioned, such a large-scale study allows to consider the cross-talks among different biological pathways (provided they involve physical interactions), and thereby take into consideration the gene pleiotropic effects. Gene pleiotropy has been proposed as a major feature for the probability of a mutation to be adaptive ([33], Figure 1.29).

However, in this study, only one type of interactions was taken into account, and therefore all the interactions at stake were not included. The interactome remains a raw simplification of the processes that take place

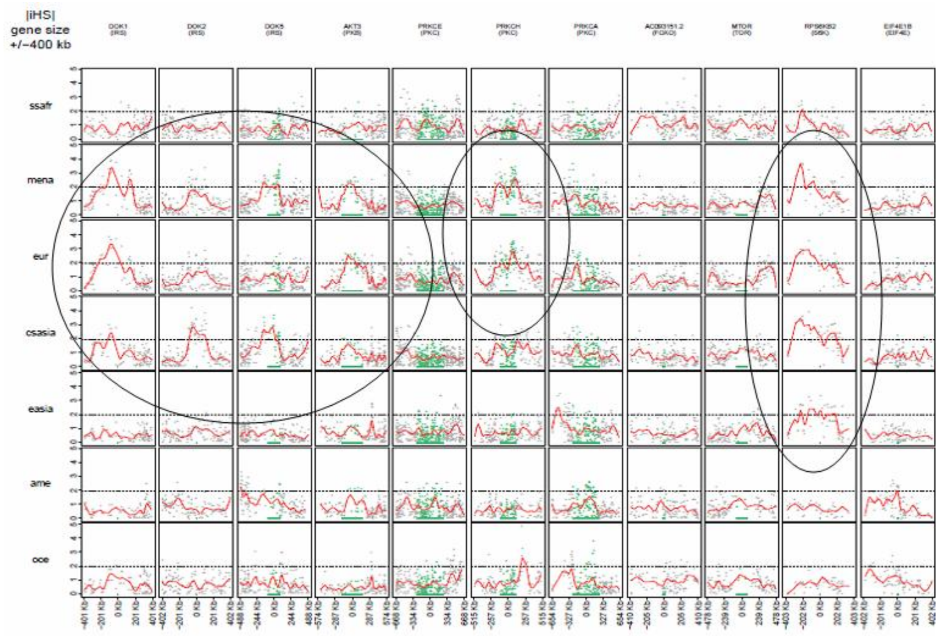


Figure 7.1: Visual inspection of signal of positive selection for the IT pathway genes. $|iHS|$ scores were calculated and plotted for each SNP within the genomic region containing the gene of interest plus 800 flanking Kb. These plots allow comparing observations for SNPs nearby the gene and in its surrounding region. Hence, they provide better visualization of signals of selection. Colour blocks on the bottom represent gene locations. SNPs within the gene are plotted in the same colour as the gene block, whereas SNPs within flanking regions are represented in grey. Red lines represent the spline function computed from SNP scores using the smooth.spline function in R (parameters: df=4, spar=0.7).

within the cell. An arguably more wondering issue is the quality of the interaction data available to date. Although immense efforts have been lately made to more accurately identify the physical interactions, the information remains largely imperfect. Indeed, to build such a large-scale network one must mostly rely on results from high-throughput experiments. Many screening using high-throughput techniques have been conducted in different organisms. Now different databases have curated and compiled the resulting information in order to provide an as exhaustive as possible interaction map in those organisms. In the particular case of this study, the interaction map was recovered from both yeast-two hybrids (Y2H) and mass spectometry experiments to complete individual focused studies available in the litterature [269]. A Y2H experiment follows the following strategy. Two proteins named the bait and prey are coupled to two halves of a transcription factor and expressed in yeast. A reporter gene is activated by the transcription factor when both proteins (prey and bait) are interacting and, thus, are reconstituting the DNA binding and transvaction domains of the transcription factor. On the one hand, such experiments allows the identification of many physical interactions, thus broadening the knowledge on biological processes to great extent. This could not be possible in individual focused studies. On the other hand, the quality of the data remains poor and one must be fully aware of this issue. Indeed, several attempts to identify and discard false positives have shown that the accuracy of the experimental approaches to identify binary protein interactions is underwhelming. For example, a study using a method designed to computationally assign scores to interactions detected through Mass Spectrometry identified an astonishing number of false positives: the original list of 2,553 interactions was narrowed down to 751 [270]. The same procedure reduced the initial list of 2,000 interactions among human mitogen activated protein kinases (MAPKs) down to 641 [271]. The development of more precise procedures is required in order to assessing the false negatives (unidentified true binary protein interactions). The main methodological limitations are: (i) the nature of the interactions (whether the interactions are transient or permanent); (ii) the physiological conditions under which such interactions occur; (iii) the algorithms utilized for assigning scores during the identification of interaction complexes; and (iv) the types of proteins identified (e.g. interaction between plasma membrane).

All together, the currently available datasets have a relatively low quality, being subject to very high false positive and false negative rates, and thus the contained information remains relatively noisy [272–275]. However, the data used in the study presented in this thesis rely on BioGRID database with dedicated efforts to provide curated information [269]. This allowed to perform a second more accurate evaluation of how positive selection, as inferred from divergence data, is distributed across the interactome, in order to validate results presented by Kim *et al.* [234]. Moreover, a strategy to avoid spurious relationship between the impact of positive selection and the interactome topology is to validate it using interactomes retrieved from different databases. For that purpose, data from the Human Protein Reference Database (HPRD [276]) and the High Quality subnetwork from BioGrid (containing only interactions from at least two independent high-throughput experiments or individual focused studies; [269]) were also used.

Besides the more accurate interaction data used as compared to the study by Kim *et al.* [234], the study described in this thesis also relies on the recent wealth in genomic data allowing a more powerful estimation of the putative impact of positive selection on genes using divergence data (for details see Chapter 6). This study also dramatically broaden the knowledge on the distribution of selective events across the interactome by inferring the impact of positive selection using polymorphism data, an unprecedented attempt. Using the recently available re-sequencing data from 1000 Genomes Project ([21]; see Section 1.2.2), the impact of positive selection was inferred by combining methods based on the three patterns expected in a region that underwent a *hard sweep* (see Section 1.4.2). One must be careful to the type of data used: the low-coverage (at ~ 2-6 X) data and the exome data (at ~50-100 X) do not have the same power to detect rare alleles in a population (see Section 1.2.2) and therefore results of statistics of positive selection can not be compared if computed on the two different sets. A reasonable strategy is therefore to only use the low-coverage data which covers the whole genome. Contrary to the HGDP genotype data used for the IT pathway analysis, 1000 Genomes Project made available genotype for a much denser set of SNPs that can not be assumed independent. For this reason the Fisher’s combination test is not suitable to summarize the scores for a given statistics observed in a given gene, and instead more simple statistics, as the average, were used. More interestingly, the 1000 Genomes Project re-sequencing data contains a larger fraction of rare variants than any genotyping data such as HGDP [16] or HapMap (www.hapmap.org). Therefore, the Site Frequency Spectrum (SFS) is unbiased, or at least the bias towards common variants is reduced as compared to traditional publicly available genotype data, and SFS-based methods could be applied. Once computed an unique positive selection score for each gene for DH [117], iHS [71] and $XPCLR$ [79], an empirical *P*-value at gene level was calculated for each gene and statistics using a genome-wide distribution obtained from ~ 13,000 genes using the outlier approach (see Section 1.5.4). To contrast whether a gene underwent a selective sweep in a given population, a visual inspection of the signals in the ~ 9,000 genes was not possible, and the study relies only on the summary statistics and its associated empirical *P*-value.

Another issue arising from using polymorphism data to infer the impact of selection in genes involved in the interactome is the fact that the number of analysed genes is relatively important (~ 9,000 genes) and therefore many are located one very close to others. To study the impact of positive selection on genes we only used the SNVs located within the genomic region corresponding to the longest transcript. However, it is well known that the regions affected by a selective sweep are large, spanning hundreds of kilobases or even megabases and containing many potential variants driving the signal. Thus, several adjacent genes may be affected by an unique event of selection targeting one particular variant. Therefore, some of the genes showing signals of positive selection in our study may be false positives, even though we do not expect that this bias can affect our network-level analysis, since there is no reason why false positives should tend to concentrate in specific parts of the PIN. To confirm that our study does not suffer from this caveat, we first validated our results using the Composite of Multiple Signals (CMS) method [106] calculated in the YRI, CEU and CHB+JPT populations using the Pilot1 genotype data from the 1000 Genomes Project [106]. Although this study used the less accurate Pilot1 data, the implemented method presents the strong advantage of more accurately pinpointing a small number of variants within a large genomic region [106]. Thus, using this test we expect to reduce to a great extent the number of falsely detected genes due to genetic hitchhiking. Our network-level analyses have been confirmed by the use of CMS and, in fact, the association between the impact of selection and network centrality appears to be stronger. To further confirm that hitchhiking does not affect the association between the impact of positive selection and network centrality, we built a subset of unlinked genes, i.e. not in linkage disequilibrium, for the three populations (YRI, CEU and CHB). For that purpose, in each population, we used the population-specific recombination rates estimated genome-wide (recombination map provided by the 1000 Genomes Project Pilot 1 [155]) and defined as a recombination hotspot a region where the observed recombination rates was greater than 10 times the genome average, i.e. greater than 18.36 cM/Mb, 18.55 cM/Mb and 17.61 cM/Mb in YRI, CEU and CHB, respectively. Then, we randomly sampled one PIN gene located between two recombination hotspots and obtained three subsets of most likely unlinked genes.

Relying on the recent wealth in both genomic and interactomic data, this study describes very interesting results. First, it validated the trend already described by Kim *et al.* that genes acting at the periphery of the human interactome are more likely to have evolved under positive selection (as inferred from divergence data from human and chimp) as described in Section 1.8.2 and Figure 1.30 [234]. However, when studying the distribution of the putative events of recent positive selection, as inferred from human polymorphism data, the opposite trend was observed: more central genes are more likely to have been targeted by positive selection. Those results are very interesting and challenge the traditional view that positive selection is active at the network periphery. An association with the Fisher's Geometric Model of Adaptation is discussed in the article (see Discussion in Chapter 6) and will be further developed in Section

7.2.1. However, as mentionned in the article, the signatures of adaptation detected in this study through either a comparative genomics or population genetics approach might correspond to different kinds of changes at the sequence level. Indeed, recent positive selection events detected using polymorphism data are likely to correspond to adaptation through changes in expression patterns (gene expression level or regulation), while selective events detected through divergence analysis may mostly correspond to changes in protein function. To gain more insight on the relationship between network centrality and positive selection at different evolutionary time-scale, one could perform a similar analysis using the asymptomatic McDonald-Kreitman (MK) test recently proposed by Messer and Petrov [277]. The original MK test estimates whether the ratio of functional (i.e. non-synonymous) to neutral (i.e. synonymous) polymorphisms (p_N and p_S, respectively) differs statistically from the ratio of functional to neutral divergence (d_N to d_S). Excess of functional divergence compared to polymorphism is attributable to positive selection. The parameter α ($\alpha = 1 - (p_N/p_S)/(d_N/d_S)$) estimates the proportion of functional substitutions driven by positive selection. Such test is therefore designed to infer the rate of positive selection in a given lineage at protein-coding sequence level. Messer and Petrov performed a simulation-based study of the behaviour of the MK test under different scenarios, making varying the proportion of adaptive variants and their selective coefficient as well as the strength of Background Selection (BGS) by playing with the number of deleterious variants, their negative selective coefficient and the rate of recombination. They found that MK estimates of α severely underestimate the true rate of adaptation and, therefore, proposed the asymptomatic MK test that yielded accurate estimates of $alpha$ in their simulations. Moreover, with the recently published Great Apes Project Data [278], this test could be performed in different ape lineages to contrast whether the same trend for recent positive selection across the interactome holds for different species.

7.2 Evolutionary *system biology is dead*! Long life to *evolutionary system biology*!

In this section, perspectives in *evolutionary system biology* will be discussed. The study of the relationship between the position occupied by a gene within a biological network and the strength of natural selection acting on it receives strong critics. One may think that such studies are vain since network topology appears as a poor predictor of the impact of natural selection. However, one may also consider the half-full glass and see *evolutionary system biology* as an emerging filed with promising potential insights, acknowledging that most methodological tools remains to be developed and more accurate data are getting produced. *Evolutionary system biology* tries to put together new layers of biological complexity to better understand the action of natural selection on protein-coding genes. This makes a lot of sense since it is a way to bridge the gap between genotype and phenotype (see Figure 7.2). However, both the biological data and the statistical framework to consider such complexity can still be dramatically improved as discussed in this section.

7.2.1 Insights from network-level analyses: Is the glass half-full or half-empty?

As already largely described in Section 1.8, several studies have identified a relationship between the position occupied by a protein and the impact of natural selection —mostly purifying, but also positive —on the encoding gene. Those studies have been performed at different biological scales using networks describing either a given biological pathway or the whole set of identified interactions of a given type, i.e. either physical, metabolic or regulatory interactions.

The results of these studies stress the usefulness of using biological networks to capture the epistasis among genes that gives rise to a given phenotype as an emergent property, but also the pleiotropic effects of genes involved in different phenotypes. In particular they have demonstrate that

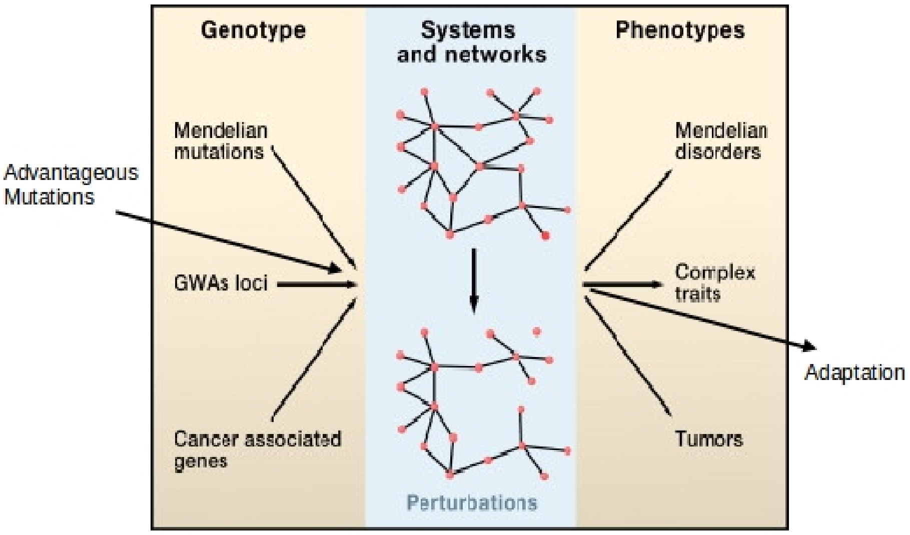


Figure 7.2: Bridge the gap between genotype and phenotype. Adapted from [279].

the position of a gene within its network accounts for a part of the variability in evolutionary rates between genes: network organization imposes constraints on the evolution of its constituent genes. However, universal patterns and general principles can not be derived despite several independent pieces of evidence of the constraint imposed by network structure on genes' evolution. Indeed, the constraints imposed by network structure appears to depend on the specific types of interactions considered, its size and, in case of small-scale networks, on the specific pathway it describes. The effect of network organization on the strength and probability of genes' adaptive evolution have been overlooked. However, the few studies analysing how positive selection distributes across small-scale networks, i.e. describing a given biological pathway, points to the same direction: positive selection often targets genes acting at the most "influential" positions of these pathways, including the most central genes in the human insulin/mTOR pathway [265], genes acting at bifurcation points of the human N-glycosylation pathway [280] and the Drosophila pathways involved in glucose metabolism [202], and the gene encoding the first enzyme of the Arabidopsis glucosinolate pathway [216]. Simulation studies also indicate that adaptation preferentially targets genes acting at the upstream and branch-point parts of pathways, at least when the system is far from the fitness optimum [281, 282]. Proteins occupying these key network positions are expected to exert strong influence over the pathway function, and thus on the associated phenotypes and organism’s fitness [216, 281, 282]. Therefore, positive selection on genes encoding such proteins may lead to rapid adaptation. Although all those studies point to “influential” genes within a given biological pathway being under positive, the gene feature characterizing genes’ “influence” is not universal. As for evolutionary constraint, the relationship between adaptive evolution and small-scale network structure appears to also depend on the specific types of interactions considered, its size and the specific pathway it describes.

Only two studies of the distribution of selective events across a large-scale network, namely the interactome, have been published (see Chapter 6 and [234]). Although different results at two evolutionary time-scales have been observed, a relationship between large-scale network topology and the impact positive selection appears to exist. The consensus since the study by Kim *et al.* [234] was that events of positive selection occur mostly at the periphery of the interactome. The study presented here challenges this view and suggests that there is not an universal pattern which could be applicable at every evolutionary time-scales and/or mode of adaptation. Indeed, it appears that the impact of positive selection within the interactome depends on (1) if only signals at protein-coding level can be identified and/or (2) the range of the selective events age one is able to detect. One possible interpretation with the results presented in Chapter 6 comes from the Fisher’s Geometric Model of Adaptation (FGM) which predicts when and how likely a mutation of given phenotype effects is advantageous, that is, it is selected for [33]. The effects on fitness are measured considering the pleiotropic effects —as a funtion of the number of phenotypes in which a mutation is involved in —and the effect on each specific phenotype. The interactome is most likely a good proxy of genes' fitness effects (see Figure 3 in Chapter 6), with genes acting at the periphery being less "influential", i.e. with lower fitness effects, than genes at the core of the network. Therefore, according to the FGM, one would expect to detect signals of positive selection in different parts of the interactome according to the evolutionary time-scale considered. Namely, if one considers a very long evolutionary time-scale, such as evolution since the divergence of many mammal species, it is likely that the detected selective events correspond to the whole processes of adaptation of the analysed species, and their common ancestors, that made them overall fit. Thus, as predicted by the FGM, it is more likely to detect selective events acting on mutation with lower fitness effects since such events are expected to be more numerous than the ones on mutation with greater fitness effects. On the other hand, when studying recent human evolution, during which human populations had to face drastic environmental changes, one may detect selective events of the new processes of adaptation to the new environments and may expect to observe them on mutations located in genes with important effects on fitness. However, as stated in Section 7.1.2, the study of the interacome in this thesis is not free of some putative confounding factors, such as the methodology used to detect signals of selection at the two evolutionary time-scales. The link between the map of physical interactions among proteins and the FGM appears to be promising and theoretical development is undergoing to bring them together (e.g. [283]; see Figure 7.3)

All together, it is now clear that studying natural selection in the context of biological network is worth. Small-scale networks represent a first approximation to integrate endophenotypes —the function performed by the encoded biological pathway —into evolutionary biology. Large-scale networks allow to account for the genes' pleiotropic effect. However, it is also true that the position occupied by a protein within a network remains a poor predictor of the impact of positive selection or the strength of evolutionary constraint on the encoding gene. Indeed, although the relationship between network topology and both selective processes, that is, positive and purifying selection, is significant in most studies, the described effect is always relatively limited and depends on factors such as

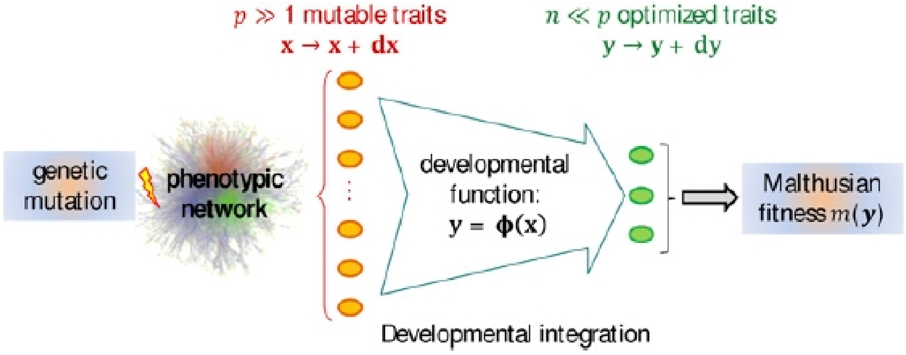


Figure 7.3: Integration of physical interaction maps into a model of genotype-phenotype-fitness map. Schematic representation showing the different levels of integration assumed in the model, from a single mutation (left) to its effect on the fitness of the whole organism. Each mutation affects a large subset of p traits (orange ovals) through the interaction network among proteins because of their pleiotropic effects. The vector x represents the parent phenotype at all these traits. The effect of a mutation (on the offspring's phenotype) is a random small perturbation dx. These basic mutational changes diffuse through the network of interactions to induce changes at a much smaller set of n key integrative traits (optimized traits; green ovals), which are those under selection, represented by the vector y. From [283].

the mode of interaction, the evolutionary time-scale and the biological scale considered.

7.2.2 Other mechanisms at stake: lessons from protein evolutionary rates.

The position occupied by a protein within a biological network accounts for some phenotypic effects of the encoding gene. However, other genes features also account for their phenotypic effects. Those features are not independent one from the others and, more interestingly, are also related to evolutionary constraint (Figure 7.4). Therefore, those features may also be considered when studying adaptive evolution of protein-coding genes. The literature on the putative gene's features affecting evolutive constraint is abundant: for decades, biologists tried to assess why certain proteins accumulate many mutations, whereas others remain unchanged over long evolutionary periods. The following description is only a brief overview on this topic and does not pretend to be exhaustive.

In 1965, Zuckerkandl and Pauling [285] suggested that the differences in the rates of evolution of hemoglobin and cytochrome *c* were attributable to the different levels of selective constraint acting on them. In the following decade, the neutral theory of molecular evolution stated that proteins with low functional importance should evolve faster than more important ones (see Section 1.3). It was also proposed that the proportion of amino acids involved in its function (“functional density”) affects levels of selective constraint [286]: proteins with a low functional density are expected to be less constrained, thereby to evolve faster. However, although it appears logical that proteins' rates of evolution are mainly determined by their importance and/or functional density, it is difficult to test this hypothesis since the measure of these parameters remains experimentally challenging to assess.

The recent emergence of genome-scale datasets allowed to measure an important number of characteristics for most of the genes in model organisms. In turn, a long list of factors correlating with rates of evolution has been drawn up. Factors more or less related to proteins' im-

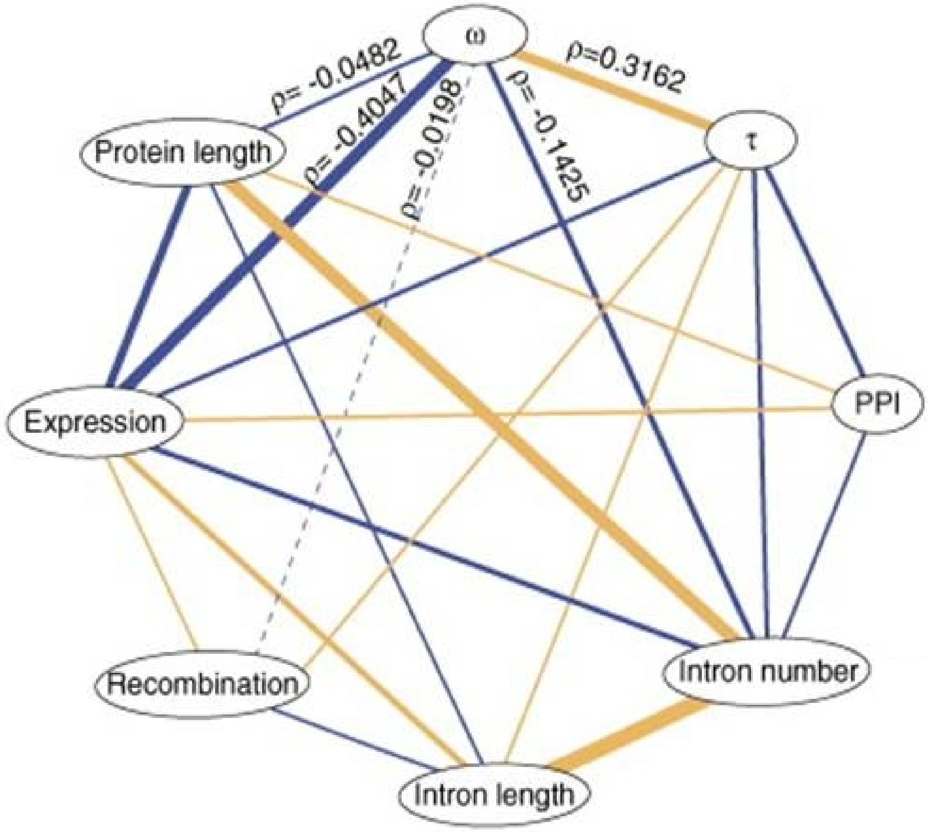


Figure 7.4: Interindependance between gene features that affect evolutionary constraint. PPI, number of protein–protein interactions; τ, range of tissue expression; ω is the d_N/d_S value. Positive correlations are represented in orange, and negative correlations are represented in blue. The width of the lines is proportional to the strength of the correlations. From [284].

portance and/or functional density appear to be relatively poor predictors of rates of evolution. These factors are the following: functional category [193, 243, 287, 288], number of functions [289–291], essentiality for survival [288, 292, 293] and dispensability measured as the fitness effect upon gene knockout [53, 294–296]. On the other hand, patterns and levels of gene expression [297, 298] appear to be the strongest determinants of levels of selective constraint. Some studies in yeasts described that gene expression levels may account for more than 30% of the variability of rates of evolution [299]. Nevertheless, not all analyses reached this conclusion and some of them suggested that certain factors aforementioned may be as determinant as expression for selective constraint (e.g. see [300–303]).

All together, the prevailing view is that patterns and levels of gene expression are the most important factors affecting evolutionary rates while other factors have a relatively minor, yet observable, effect. Several studies already described the relationship between the strength of positive selection acting on a gene and its expression patterns (e.g. see [170, 297, 304, 305]). However, further experimental and theoretical advances are necessary to better understand the contributions of the different factors driving evolutionary rates as well as the probability and the strength of positive selection targeting protein-coding genes. Indeed, currently available datasets have a relatively low quality and are being subject to very high false positive and false negative rates [272–275]. Despite the poor accuracy of those datasets, they allowed to perform a first evaluation of the mechanisms affecting the impact of natural selection on protein-coding genes. Available measures of protein–protein interactions may be particularly noisy, as they are the result on the application of high-throughput techniques, especially for large-scale networks (see Section 7.1.2). This could result in an under-estimation of the relative effect of the network topology on the impact of natural selection on protein-coding genes (see Section 7.1.2). Indeed, a study in which noise levels were equalized across 7 putative predictors of d_N/d_S in yeasts, described a roughly equal contribution to the variability of evolutionary rates [302]. Therefore, the critics stating that using biological network topology to better understand selective processes is uninformative as compared to other biological features (e.g. expression patterns and levels) are not receivable since they do not take into account the discrepancy in the accuracy of the available data.

7.2.3 Perspectives in *evolutionary system biology*.

Consider different modes of adaptive selection.

In order to better understand the mechanisms driving natural selection in a system, it is obviously important to first have a general picture of its impact on the involved genes and their main regulatory regions. As largely discuss in Section 1.3.1, there are three modes of natural selection: purifying, balancing and positive selection. Balancing selection has been overlooked and relating its impact to the position occupied by the encoded proteins in biological network would be informative. Positive selection is mostly studied assuming the *hard sweep* model. Considering other modes of positive selection, such as *soft sweeps* and *polygenic adaptation*, would be interesting to provide a full picture of the distribution of selective events within biological networks. Actually, the study of *polygenic adaptation* in the context of biological network is necessary for obvious reasons. Moreover, the fact that more selective events have been observed in more central genes in the interactome and individual pathways using polymorphism data (see Chapter 6) is arguably not surprising since *hard selective sweeps* have been detected, that is selective events on mutations with a relatively important selective coefficient. The selective coefficient of advantageous mutation is likely to be more important for mutations affecting the function or the regulation of central genes which have greater effects on fitness.

Using methodological tools allowing to infer the action of natural selection on mutations that can affect the regulation of genes may be informative to understand how biological systems evolve through the action of natural selection. While the action of purifying selection in biological system has been well studied, it relies on a methodology biased towards coding regions. Recent efforts have been made to also infer the evolutionary constraint in regulatory regions taking as a reference putatively neutrally evolving regions as ancestral repeats or pseudogenes (e.g. see [306] for micro RNAs and [307] for transcription factor binding sites) from divergence data. The same direction should be followed to estimate positive selection using divergence data.

The study of network topology can shed more light on the mechanisms driving natural selection.

Although the relationship between the impact of natural selection —either positive or negative —and network topology is weak, yet measurable, it is worth keeping studying it across many more biological networks. First, at small-scale, the distribution of selective events across biological pathways has been studied only for very few of them. It is necessary to broaden the range of studied pathways in this framework in order to better understand whether the impact of natural selection on and across a pathway varies according to the pathway size (number of proteins and interactions involved), its linearity (number of branching points), its isolation from other pathways, the type of interactions involved, the cellular compartment in which proteins are active, the tissue specificity of the pathway, etc. However, one must be cautious while performing a systematic analysis of pathways based on database annotation (see Section 7.1.2).

Second, at large-scale, although purifying selection has been relatively well studied the three main interaction maps (transcriptome, metabolome and interactome), more analyses would be informative in other organisms and using networks of increasing accuracy to further validate the view that central genes are more constrained in their evolution. Moreover, the distribution of selective events across such large-scale networks have been overlooked and contrasting whether the patterns observed in the interactome stand in the metabolome and the transcriptome would be interesting. Before proceeding to such studies, one would rather increase the accuracy of the large-scale network of interest. For example, the Bertranpetit Group is currently curating manually the human metabolome in collaboration with biochemists. Such effort is necessary to reduce the noise and, thereby, increase the power of the study. Furthermore, in order to better consider the processes that take place within the cell, it would be interesting to study networks that encompass all the types of interactions, such as the MULTINET built by Khurana *et al.* [200]. Indeed, as illustrated by the IT pathway (see Chapter 5), different types of interactions are at stake within many pathways, mostly the transduction signalling ones. Therefore, considering large-scales networks that include only one type of interactions remains a raw simplification of genes epistasis and pleiotropy. Third and last, when studying network topology, one would have to take into account the fact that interactions do not have the same importance for the processes taking place in the organism: (1) interactions happen in different compartment of the cell; (2) some are tissue specific while others operate within a broad range of cells; (3) proteins may compete one with the others for the same interacting partners; and (4) interactions are not necessarily occurring at any time. Therefore, including into the network studies more information on the expression patterns, the cell compartment and the tempo of interactions, could be important. For now, network-level studies, such as the ones presented in this thesis, settle for only correcting for those parameters through multivariate analyses. More insight would be gain if, instead, this information could be used to build subnetworks to study individually and to compare.

Include dynamics.

Considering the fixed topology of the networks was the first step in *evolutionary system biology*, and it is clear that it is a huge simplification of the protein interaction complexity. Some efforts have been recently made to include dynamical features into the studied system.

First, one can directly rely on measures of dynamical characteristics of the involved proteins. For example Colombo *et al.* studied the relationship between metabolic flux and evolutionary rates of enzyme encoding genes in the human erythrocyte cells [308]. The flux is the movement of matter through metabolic networks that are connected by chemical equilibria, and thus describes the activity of the metabolic network as a whole using a single characteristic. They found that genes encoding enzymes

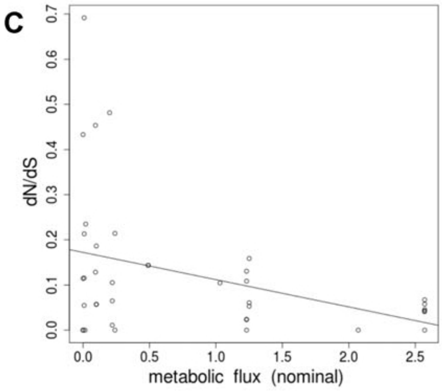


Figure 7.5: Metabolic flux is a determinant of the evolutionary rates of enzyme-encoding genes in human erythrocytes. From [308].

carrying high fluxes evolved under stronger purifying selection, while evolutive constraint was relaxed in genes encoding enzymes carrying low metabolic fluxes (Figure 7.5). This demonstrates the importance of considering the dynamical functioning of gene networks in order to study the action of selection on biological systems. In the Glucosinolate pathway in *Arabidopsis thaliana*, Olson-Manning *et al.* observed that the gene with greater control over metabolic flux and was the only one with signature of adaptive selection [216]. The fact that they demonstrated that the upstream gene in the pathway was the one with greater control on flux [216] is also comforting: network topology is informative. Properties such as metabolic flux has not been well assessed for many systems, and the two aforementioned studies are unique. More importantly the flux measure remains specific to metabolic pathways.

Studies based on simulations of metabolic pathways also pointed to the importance of enzyme control on the dynamic process and the impact of natural selection on the encoding gene, but also described the relationship between the position in the network and control on metabolic flux [281, 282]. A study also based on simulations assessed the relationship between the intensity of regulatory action and the strength of evolutionary constraint in a regulatory network [309], in order to circumvent the fact that a gene connected to many loci but only through weak regulation effects is not expected to be strongly exposed to selection since it does not in practice have a strong effect on the expression of the products of the different genes. The authors observed that an increasing intensity of purifying selection on the phenotype leads to an increased level of regulation between the genes [309]. They also showed that the genes responding more strongly to selection within the network were those evolving towards stronger regulatory action on the other genes and/or those that are the less regulated by the other genes [309]. Although those simulations studies shed light on how gene influence , i.e. the control exerted by the gene on the output of the system, they represent idealized systems that do not characterize real ones operating in an organism.

Less commonly used in the study of evolution are dynamic models of real biochemical systems: a mathematical model can be built in order to simulate known system dynamics. In turn, it can be used as a test of the breadth of knowledge of a system. This model would include known reactions and if the simulations are accurate, it is likely that the mechanisms of the system are well-assessed. Building such models rely on a deep biochemical knowledge of the interactions at stake in a given biomolecular pathway. Invergo and colleagues built a deterministic model of the mammalian phototransduction pathway [210]. In such model, reactions are described by a system of differential equations that track the concentrations of the various molecules in the system. At any time point in the simulation, it is possible to calculate the exact concentrations at the next instant. Using relevant model parameters as a proxy of dynamic influence, perhaps surprisingly, the authors observed that proteins with greater potential to disrupt the system dynamics exhibited a more relaxed evolutionary constraint, in the form of higher evolutionary rates (personal communication). Such biochemical models are useful to better understand the role exerted by the genes in biological systems, to ,in turn, understand the mechanisms affecting the impact of natural selection. However, the phototransduction pathway is unique in the sense that it is isolated, involving specific proteins. Building such dynamical models for other pathways remains challenging.

Chapter 8

CONCLUSIONS

A major challenge in evolutionary biology is to understand how natural selection, which acts on phenotypes, shapes the genome, and particularly the protein-coding genes that contribute to the phenotype. The two first studies of this thesis illustrates how the impact of adaptive selection can be detected at the gene-level. However, here genes are considered as independent entities, while they do not act in isolation but rather interact one with many others to give rise to the phenotype as an emergent property. This concerted contribution to the phenotypes has to be taken into consideration to better understand the mechanisms driving natural selection at the molecular level. With others, the results from the two last studies presented in this thesis demonstrate how encoding biological systems into networks allows a first approximation that captures the concerted action of the genes. Those studies show that the position of a gene within its biological network allows to understand some of the variability in evolutionary rates between genes and the impact of adaptive selection: those works prove that the action of natural selection is somehow circumscribed by network organization.

Several independent pieces of evidence derived from studies of different network types and at different scales did not allow to define universal patterns and general principles. This could be due to the fact that the constraints imposed by network topology likely depend on the type of interaction considered, its size and, for small-scale networks, on the specific system, but also on the tempo of the adaptive process.

On the other hand, the discrepancy between the observed patterns may also come from the poor quality of the interaction maps, particularly from high-throughput experiments allowing to build large-scale networks. Advances in assessing with higher accuracy the interactions taking place in an organism will most likely allow to clarify how network structure has an impact on gene evolution.

Moreover, when natural selection is inferred from divergence or polymorphism data, that is at two different evolutionary time-scales but also relying on different kinds of methods, its relationship with network topology differs: while interspecific studies point to positive selection being particularly active at the periphery of the networks, intraspecific analyses described an enrichment of selective events in highly connected genes. This point needs further inspection in the future.

Despite all the difficulties, much knowledge has been and will be derived from considering the topological organization of the networks, provided an increase accuracy of the interaction maps and the integration of new layers of complexity for both the biological processes and the mode of natural selection. Moreover, a deeper understanding might arise from the dynamics and functioning in space and time of the networks, to move beyond their topology. The *evolutionary system biology* field has still long years ahead. However, much experimental work is needed to gather all the information to allow the integration of new layers of complexity. At the very end, *evolutionary system biology* is likely to contribute significantly to the understanding of the genetic bases of complex adaptation.

Part IV

Appendix

LIST OF PUBLICATIONS

Articles Published in Peer-Reviewed Journals

1. Mayukh Mondal, Garima Juyal, **Pierre Luisi**, Hafid Laayouni, Peter Heutink, Jaume Bertranpetit, Ferran Casals and Thelma BK
Population and genomic lessons from genetic analysis of two Indian populations. Hum Genet. 2014 Jul 1.
`http://www.ncbi.nlm.nih.gov/pubmed/24980708`
2. Vincenza Colonna, Qasim Ayub, Yuan Chen, Luca Pagani, **Pierre Luisi**, Marc Pybus, Yali Xue, Chris Tyler-Smith, The 1000 Genomes Project Consortium
Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol. 2014 Jun 30; 15(6):R88
`http://www.ncbi.nlm.nih.gov/pubmed/24980144`
3. Hafid Laayouni, Marije Oosting, **Pierre Luisi**, Mihai Ioana, Santos Alonso, Isis Ricaño-Ponce, Gosia Trynka, Alexandra Zhernakova, Theo Plantinga, Shih-Chin Cheng, Jos W.M. van der Meer, Thelma BK, Radu Popp, Ajit Sood, Cisca Wijmenga, Leo A.B. Joosten, Jaume Bertranpetit and Mihai G. Netea
Common evolutionary signals in European and Rroma populations: convergent evolution exerted by plague on TLR1/TLR6/TLR10 pattern recognition system. Proc Natl Acad Scien U S A. 2014 Feb 18; 111(7): 2668-73
`http://www.ncbi.nlm.nih.gov/pubmed/24550294`

4. Audrey Sabbagh*, **Pierre Luisi***, Erick C. Castelli, Laure Gineau, David Courtin, Jacqueline Milet, Juliana D. Massaro, Hafid Laayouni, Philippe Moreau, Eduardo A. Donadiand André Garcia
   Worldwide genetic variation at the 3' untranslated region of the HLA-G gene: balancing selection influencing genetic diversity. Genes Immun. 2013 Dec 19
   * Equal contribution
   `http://www.ncbi.nlm.nih.gov/pubmed/24352166`
5. Marc Pybus*, Giovanni M Dall'Olio*, **Pierre Luisi***, Manu Uzkudun*, Angel Carreño-Torres, Pavlos Pavlidis, Hafid Laayouni, Jaume Bertranpetit and Johannes Engelken
   1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res. 2013 Nov 25
   * Equal contribution
   `http://www.ncbi.nlm.nih.gov/pubmed/24275494`
6. Blandine Patillon, **Pierre Luisi**, Audrey Sabbagh and Emmanuelle Génin
   Signatures Of Recent Positive Selection At The VKORC1 Gene Locus. Genetic epidemiology. 2012;36(2), 170-170
7. Blandine Patillon*, **Pierre Luisi***, Hélène Blanché, Etienne Patin, Howard M. Cann, Emmanuelle Génin and Audrey Sabbagh
   Positive selection in the chromosome 16 VKORC1 genomic region has contributed to the variability of anticoagulant response in humans. PLoS One. 2012;7(12):e53049
   * Equal contribution
   `http://www.ncbi.nlm.nih.gov/pubmed/23285254`
8. Giovanni M. Dall'Olio, Hafid Laayouni, **Pierre Luisi**, Martin Sikora, Ludovica Montanucci and Jaume Bertranpetit
   Distribution of events of positive selection and population differentiation in a metabolic pathway: the case of asparagine N-glycosylation. BMC Evol Biol. 2012 Jun 25;12:98
   `http://www.ncbi.nlm.nih.gov/pubmed/22731960`

9. **Pierre Luisi**, David Alvarez-Ponce, Giovanni M. Dall'Olio, Martin Sikora, Jaume Bertranpetit and Hafid Laayouni
Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations. Mol Biol Evol. 2012 May;29(5):1379-92.
`http://www.ncbi.nlm.nih.gov/pubmed/22135191`
10. Lounès Chikhi, Vitor C. Sousa, **Pierre Luisi**, Benoît Goossens and Mark A. Beaumont
The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. Genetics. 2010 Nov;186(3):983-95 `http://www.ncbi.nlm.nih.gov/pubmed/20739713`

Submitted

1. **Pierre Luisi***, David Alvarez-Ponce*, Marc Pybus, Mario A. Fares, Jaume Bertranpetit and Hafid Laayouni
Recent positive selection targets the center of the human protein-protein interaction network.
* Equal contribution
2. Laure Gineau, **Pierre Luisi**, Erick C. Castelli, Jacqueline Milet, David Courtin, Blandine Patillon, Hafid Laayouni, Philippe Moreau, Eduardo A. Donadi and André Garcia
Balancing immunity and tolerance: genetic footprint of natural selection in the HLAG 5' upstream regulatory region.
3. Blandine Patillon, **Pierre Luisi**, Estella Poloni, Sotiria Boukouvala, Pierre Darlu, Emmanuelle Génin
A homogenizing process of selection has maintained an 'ultra-slow' acetylation NAT2 variant in humans.

In Preparation

1. Marc Pybus*, **Pierre Luisi***, Giovanni M Dall'Olio*, Manu Uzkudun, Angel Carreño-Torres, Pavlos Pavlidis, Hafid Laayouni, Jaume Bertranpetit and Johannes Engelken
   A machine-learning framework to detect and classify hard selective sweeps in human populations.
   * Equal contribution
2. Begoña Dobon, Hisham Hassan, Hafid Laayouni, **Pierre Luisi**, Isis Ricaño-Ponce, Alexandra Zhernakova, David Comas, Mihai G. Netea, and Jaume Bertranpetit
   The genetics of human populations of the Sudanese region: a novel Nilotic component in the African landscape and signals of positive selection in the East-African

SUPPLEMENTARY MATERIALS.

Study the impact of positive selection on a candidate gene.

The Supplementary Material for the article presented in Chapter 3 [263] is available at `http://www.plosone.org/article/info\%3Adoi\%2F10.1371\%2Fjournal.pone.0053049#s5`.

Scan the genome for positive selection.

The Supplementary Material for the article presented in Chapter 4 [264] is available at `http://www.pnas.org/content/111/7/2668.long?tab=ds`.

Distribution of selective events within a small-scale protein-protein interaction map.

The Supplementary Material for the article presented in Chapter 5 [265] is available at `http://mbe.oxfordjournals.org/content/29/5/1379/suppl/DC1`.

Distribution of selective events within a large-scale protein-protein interaction map.

The Supplementary Material for the submitted article presented in Chapter 6 is provided below.

RESEARCH ARTICLE

Supplementary Information

Recent Positive Selection Has Acted On Genes Encoding Proteins With More Interactions Within the Whole Human Interactome

Pierre Luisi[1]†, David Alvarez-Ponce[2,3]†, Marc Pybus[1], Mario A. Fares[2,4], Jaume Bertranpetit[1]* and Hafid Laayouni[1,5]*.

(† These authors contributed equally to this work)

[1]Institute of Evolutionary Biology (Universitat Pompeu Fabra-CSIC), CEXS-UPF-PRBB, Barcelona, Catalonia, Spain.

[2]Integrative Systems Biology Group, Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas (CSIC)-Universidad Politécnica de Valencia (UPV), Valencia, Spain.

[3]Current address: Biology Department, University of Nevada, Reno, NV, USA.

[4]Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin, Ireland.

[5]Departament de Genètica i de Microbiologia, Grup de Biologia Evolutiva (GBE), Universitat Autonòma de Barcelona, Bellaterra (Barcelona), Spain

* Authors for correspondence.

SUPPLEMENTARY NOTES

Network-Level Analysis Using Different Methods to Detect Positive Selection from Polymorphism Data

In order to confirm the relationship observed between the position of proteins in the Protein-Protein Interaction Network (PIN) and intraspecific positive selection, we broadened the analysis by using separately the three original tests to detect positive selection iHS (Voight et al. 2006); XP-CLR (Chen et al. 2010) and *DH* (Zeng et al. 2007) used to compute the Fisher's combination test score (Z_F). We also used the Composite of Multiple Signals (CMS) method (Grossman et al. 2010) calculated in YRI, CEU and CHB+JPT populations using Pilot1 genotype data from the 1000 Genomes Project (Grossman et al. 2013). For each gene we used the average score (for iHS, XP-CLR and CMS) or the $-\log_{10}$(*P*-value) (for *DH*) as summary scores. We then applied a Spearman's correlation analysis between gene degree (number of interactions), as an estimator of network centrality, and these scores. Moreover, genes were classified into four groups delimited by their first, second and third degree quartiles. The median summary scores of the four groups were compared using a non-parametric ANOVA test. We also applied a linear trend test to contrast whether the putative differences among groups were due to a trend towards higher summary scores in groups corresponding to higher degrees (Supplementary Table 2; Supplementary Figure 2).

The Spearman's correlation between the iHS score and degree is positive and significantly different from 0 in the three populations ($P \leq 0.0014$ in all three cases; Supplementary Table 2). In addition, iHS values are significantly different in the different groups according to the degree quartiles in all three populations (non-parametric ANOVA; $P \leq 0.0032$; Supplementary Table 2; Supplementary Figure 2), and these differences among groups are due to a clear trend towards higher iHS scores in groups corresponding to higher degrees (linear trend test on ranks; $P \leq 0.0017$*;* Supplementary Table 2). When using the XP-CLR score, we did not observe such a clear relationship between degree and the impact of positive selection. Indeed, although the Spearman's correlation coefficient is positive in the three populations and significantly different from 0 in YRI and CHB (*P* equal to 0.0007 and 0.0127, respectively; Supplementary Table 2), the non-parametric ANOVA reaches significance only in YRI ($P = 0.0211$), and so does the linear trend test on ranks in YRI and CEU ($P = 0.0028$ and $P = 0.0236$, respectively; Supplementary Table 2; Supplementary Figure 2). Moreover, we also observe an association between degree and the impact of selection as measured by *DH*. Indeed, the Spearman's correlation coefficient is positive and significantly different from 0 in the three populations (*P* ≤ 0.0015; Supplementary Table 2), and we observed significantly different *DH* values in different degree groups (non-parametric ANOVA; *P* ≤ 0.0222; Supplementary Table 2; Supplementary Figure 2), as well as a significant linear trend test on ranks in all three populations (*P* ≤ 0.0027; Supplementary Table 2; Supplementary Figure 2). Finally, using the CMS score, the association appears clearer: we observe significantly positive Spearman's correlation coefficients in all three populations (*P* ≤ 0.0009; Supplementary Table 2) as well as significant differences in CMS scores among the degree groups due to a clear tendency towards higher CMS values in groups corresponding to higher degrees (non-parametric ANOVA, *P* ≤ 0.0080; linear trend test on ranks *P* ≤ 0.0016; Supplementary Table 2; Supplementary Figure 2).

In summary, the observed general tendency of central genes to evolve under recent positive selection remains significant when positive selection is inferred separately from different statistics.

Network-Level Analysis for Positive Selection Inferred Using Polymorphism Data in a Subset of unlinked Genes

To study the impact of positive selection on genes we only used the SNVs located within the genomic region corresponding to the longest transcript. However, it is well known that the regions affected by a selective sweep are large, spanning hundreds of kilobases or even megabases and containing many potential variants driving the signal. Thus, several adjacent genes may be affected by a unique event of selection targeting one particular variant. Therefore, some of the genes showing signals of positive selection in our study may be false positives, even though we do not expect that this bias can affect our network-level analysis, since there is no reason why false positives should tend to concentrate in specific parts of the PIN. To confirm that our study does not suffer from this caveat, we first validated our results using the Composite of Multiple Signals (CMS) method (Grossman et al. 2010) calculated in the YRI, CEU and CHB+JPT populations using the Pilot1 genotype data from the 1000 Genomes Project (Grossman et al. 2010). Although this study used the less accurate Pilot1 data, the implemented method presents the strong advantage of more accurately pinpointing a small number of variants within a large genomic region (Grossman et al. 2010). Thus, using this test we expect to reduce to a great extent the number of falsely detected genes due to genetic hitch-hiking. Our network-level analyses have been confirmed by the use of CMS and, in fact, the association between the impact of selection and network centrality appears to be stronger (see Supplementary Table 2; Supplementary Figure 2; see previous part of this supplementary information).

To further confirm that hitch-hiking does not affect the association between the impact of positive selection and degree, we built a subset of unlinked genes, i.e. not in linkage disequilibrium, for the three populations (YRI, CEU and CHB). For that purpose, in each population, we used the population-specific recombination rates estimated genome-wide (recombination map provided by the 1000 Genomes Project Pilot 1 (The 1000 Genomes Project Consortium 2010)) and defined as a recombination hotspot a region where the observed recombination rates was greater than 10 times the genome average, i.e. greater than 18.36 cM/Mb, 18.55 cM/Mb and 17.61 cM/Mb in YRI, CEU and CHB, respectively. Then, we randomly sampled one PIN gene located between two recombination hotspots. We obtained three subsets of most likely unlinked genes involved in the PIN containing 2792, 3106 and 3107 genes in YRI, CEU and CHB, respectively.

For each gene we used the Z_F score as the likelihood of having been targeted by positive selection in the human populations, and observed that it is significantly positively correlated with degree in all three populations (Spearman's correlation analysis; $P \leq 0.0248$). Moreover, when genes were classified into four groups delimited by the first, second and third degree quartiles, we observed significant differences of summary scores among groups in CEU and CHB (non-parametric ANOVA test, *P* equal to 0.0036 and 0.0075, respectively; Supplementary Table 3; Supplementary Figure 3). Through a linear trend test on ranks, we concluded that these differences among groups were due to a trend towards higher summary scores in groups corresponding to higher degrees in these two populations ($P \leq 0.0053$; Supplementary Table 3; Supplementary Figure 3). For YRI, although the non-parametric ANOVA was not significant ($P = 0.2661$), the linear trend test on ranks was marginally significant ($P = 0.0482$).

Network-Level Analysis Correcting for Putative Confounding Factors

Factors such as gene expression level and breadth (tissues in which a gene is expressed), and the length of the encoded proteins, correlate with both network centralities and the likelihood of detecting natural selection, and hence could potentially have an effect on the observed relationship between the impact of natural selection and the gene centrality in the network (Anisimova et al. 2002; Duret & Mouchiroud 2000; Kim et al. 2007; Kosiol et al. 2008; Pál et al. 2006; Subramanian & Kumar 2004). In order to evaluate the effect of these factors, we applied a linear regression between the scores used as the likelihood of having been targeted by positive selection during human and mammalian evolution (Z_F and $2\Delta\ell$, respectively), as well as the scores that estimate the strength of purifying selection (DAF and ω, respectively) and the mentioned putative confounding factors. The linear regression residuals were then used to perform the correlation analysis in each case. The relationship between positive selection inferred using polymorphism data and degree remains significant in all three populations with a Spearman's correlation coefficient, *ρ*, ranging between 0.0326 (P = 0.0059) in CEU and 0.0451 (P = 0.0001) in YRI (Main text Table 1). Moreover, the non-parametric ANOVA and trend tests on ranks provide similar results when using the linear regression residual instead of the Z_F score, although *P*-values tend to be higher (Main text Table 1; Supplementary Figure 4). Indeed, the non-parametric ANOVA test is significant in YRI (P = 0.0423) and CEU (P = 0.0240), while it does not reach significance in CHB (P = 0.1006). Moreover, we observe a trend towards higher residuals in groups corresponding to higher degree (Supplementary Table 3; Supplementary Figure 4). Indeed, in YRI and CHB the linear trend test on ranks reaches significance (P = 0.0053 and P = 0.0239, respectively). Taken together, these observations indicate that the association observed between the Z_F score and degree within the PIN cannot be attributed to the three putative confounding factors. Similarly, the association observed between the impact of positive selection inferred using divergence data (as estimated by $2\Delta\ell$) and degree remains significant when we correct for the three putative confounding factors. Indeed, we observed a significant negative Spearman's partial correlation coefficient (ρ = -0.0340; P = 0.0107). Moreover, although the non-parametric ANOVA test is not significant (P = 0.0548), the trend test remains significant (P = 0.0122) with lower residuals in groups corresponding to higher degree. Finally, the relationship between purifying selection and degree also remains significant when using the residuals of the linear regression of either DAF or ω with protein length, expression level and expression breadth. Indeed, the correlation tests remain significant (ρ = -0.0668 and -0.1697, respectively; P < 0.0001 in both cases), as well as the non-parametric ANOVA ($P \leq 0.0003$) and the linear trend tests (P < 0.0001).

Network-Level Analysis Using Different Protein-Protein Interaction Networks

To validate the association between network position and the impact of positive selection, analyses were repeated using two additional high-quality networks: a high-quality (HQ) subnetwork from BioGRID (Stark et al. 2011), in which we retained only interactions discovered by low-throughput techniques, plus those reported in at least two independent high-throughput analyses, and the network from the Human Protein Reference Database (HPRD) (Keshava Prasad et al. 2009), derived from the literature. As in the main analysis, we calculated the number of interactions in which each protein is involved (degree centrality), considering the whole set of non-redundant interactions.

For both the HQ and HPRD networks, the Spearman's correlation coefficient between degree and the recent positive selection Z_F scores is positive and significantly different from 0 in all three populations ($\rho \geq 0.0257$; $P \leq 0.0294$; Supplementary Table 4), except for CHB when using the HPRD data set ($\rho = 0.0229$; P=0.0533). Moreover, for the HQ sub-network the non-parametric ANOVA test is significant in the three populations ($P \leq 0.00950$) (Supplementary Table 4; Supplementary Figure 5). These differences among groups are due to a trend towards higher Z_F scores in groups corresponding to higher degree. Indeed, the linear trend test on ranks reaches significance in all three populations ($P \leq 0.0027$ in the three cases). When using the HPRD network, although most of the non-parametric ANOVA tests do not reach significance, the overall pattern also points to higher Z_F scores in groups corresponding to higher degrees: the linear trend test on ranks is significant in YRI and CEU (P equal to 0.0364 and 0.0053, respectively; Supplementary Figure 6; Supplementary Table 4) and marginally significant in CHB ($P = 0.0628$).

When studying the association between positive selection as inferred from divergence data (estimated by $2\Delta\ell$) and degree in the HQ and HPRD networks, we observed a negative Spearman's correlation coefficient ($\rho = -0.0620$ and $\rho = -0.0577$, respectively; $P < 0.0001$; Supplementary Table 4). We also observed differences in the $2\Delta\ell$ scores among degree groups (non-parametric ANOVA; $P \leq 0.0011$; Supplementary Table 4; Supplementary Figures 5 and 6), with higher $2\Delta\ell$ scores in groups corresponding to lower degrees. Indeed, the linear trend test on ranks reaches significance in both networks ($P \leq 0.0001$) . Finally, when studying the association between purifying selection and degree in both the HQ and HPRD networks, we observed a significantly negative Spearman's correlation coefficient ($\rho \leq -0.0488$; $P < 0.0001$; Supplementary Table 4). Moreover, we observed clear differences in both DAF and ω among degree groups (non-parametric ANOVA, $P \leq 0.0007$), due to a clear tendency towards lower scores in groups corresponding to higher degrees (linear trend test on ranks; $P \leq 0.0001$; Supplementary Table 4; Supplementary Figures 5 and 6).

Network-Level Analysis Using Different Centrality Measures

We explored whether the association found between the impact of natural selection and network centrality, as estimated by degree (the number of interactions in which a protein is involved), was also significant when using other centrality measures. For that purpose, we calculated two other centrality measures: betweenness (the number of shortest paths between other proteins passing through a given protein), and closeness (the inverse of the average distance to all other proteins in the network). The association between the impact of natural selection and these network centrality measures remains similar, regardless of the centrality measure considered. Indeed, the Spearman's correlation coefficient between either betweenness or closeness and Z_F, the score used as the likelihood of having been targeted by positive selection in recent human evolution, is significantly positive in all three populations ($\rho \geq 0.0295$; $P \leq 0.0096$; Supplementary Table 5). Moreover, we observed Z_F differences among betweenness groups in the three populations performing a non-parametric ANOVA (Supplementary Table 5; Supplementary Figure 7), which reaches significance in the three populations ($P \leq 0.0332$; Supplementary Table 5; Supplementary Figure 7). These differences are due to a clear tendency for higher Z_F scores in groups corresponding to higher betweenness (linear trend test on ranks, $P \leq 0.0157$; Supplementary Table 5; Supplementary Figure 7). When comparing the Z_F scores among closeness groups, the non-parametric ANOVA test reaches significance in YRI and CHB ($P = 0.0093$ and $P = 0.0113$, respectively; Supplementary Table 5; Supplementary Figure 8). These differences among groups are also due to a trend towards higher $\mathrm{Z_F}$ scores in groups corresponding to higher closeness. Indeed, the linear trend test on ranks is significant in all three populations ($P \leq 0.0091$; Supplementary Table 5; Supplementary Figure 8).

When studying the association between positive selection during mammalian evolution (as estimated by $2\Delta\ell$) and network centrality, using both betweenness and closeness, we observed a significantly negative Spearman's correlation coefficient (ρ equal to -0.0645 and -0.0726, respectively; $P < 0.0001$; Supplementary Table 5). We observed differences in the $2\Delta\ell$ scores among betweenness groups (non-parametric ANOVA, $P < 0.0001$; Supplementary Table 5; Supplementary Figure 7), and among closeness groups (non-parametric ANOVA, $P < 0.0001$; Supplementary Table 5; Supplementary Figure 8), with a clear tendency for higher $2\Delta\ell$ scores among groups corresponding to lower betweenness or closeness (linear trend test on ranks , $P < 0.0001$; Supplementary Table 5; Supplementary Figures 7-8).

Finally, when studying the association between purifying selection and either betweenness or closeness, we observed significantly negative Spearman's correlation coefficients ($\rho \leq -0.0641$; $P < 0.0001$; Supplementary Table 5). Moreover, we observed clear differences in both DAF and ω values among either betweenness or closeness groups (non-parametric ANOVA, $P < 0.0001$; Supplementary Table 5; Supplementary Figures 7 and 8), due to a clear tendency towards lower scores in groups corresponding to higher centrality measures (linear trend test on ranks; $P < 0.0001$; Supplementary Table 5; Supplementary Figures 7 and 8).

Network-Level Analysis for Positive Selection Inferred Using Polymorphism Data Correcting for the Action of Purifying Selection

The action of purifying selection on a genomic region can leave some patterns that are similar to the ones expected under recent positive selection (e.g. an excess of rare variants). Therefore, we wanted to confirm that the association found between positive selection estimated from polymorphism data and degree is not a by-product of the already described association between purifying selection and network centrality. For that purpose, we applied a linear regression between Z_F, the score used as the likelihood of having been targeted by the impact of recent positive selection in human populations, and ω, which estimates the strength of purifying selection during mammalian evolution. The linear regression residuals were then used to perform the analysis. The relationship between positive selection and degree remains positive in all three populations, with a Spearman's correlation coefficient, ρ, greater than or equal to 0.0195 (Supplementary Table 6). It is significantly different from 0 in YRI and CHB (P equal to 0.0020 and 0.0032, respectively). Moreover, the residuals are marginally different among degree groups (non-parametric ANOVA; P ranging from 0.0371 to 0.0578; Supplementary Table 6; Supplementary Figure 9). These differences are due to a trend towards higher residuals in groups corresponding to higher degrees in YRI and CHB (linear trend test on ranks; P equal to 0.0081 and 0.0100, respectively; Supplementary Table 6; Supplementary Figure 9). In summary, the observed association between Z_F scores and protein degree cannot be attributed to the association between network centrality and the action of purifying selection.

We further confirmed that background selection (BGS), a process that removes neutral variation linked to deleterious mutations, thus reducing levels of polymorphism in regions of high functional density and low recombination (Charlesworth et al. 1993), does not confound the association observed between network centrality and positive selection estimated from polymorphism data. We estimated the level of BGS acting on each gene using two correlates of BGS: GC content and recombination rate. Note that we did not take into account the level of functional constraint because the present study focuses on protein-coding genes. For each gene, we calculated the average of GC content from the 5-mer table downloaded from the UCSC browser (Karolchik et al. 2009) (table "gc5Base" downloaded on the 10th of July, 2013), as well as the average recombination rate from the population-specific recombination rates estimated genome-wide (recombination map provided by the 1000 Genomes Project Pilot 1 (The 1000 Genomes Project Consortium 2010)). We then applied a linear regression between Z_F, the score used as the likelihood of having been targeted by positive selection, and both recombination rate average and GC content average. The linear regression residuals were then used to perform the analysis. The relationship between positive selection and degree remains positive in all three populations, with a Spearman's correlation coefficient, ρ, greater than 0.0369 (Supplementary Table 7) and significantly different from 0 ($P \leq 0.0013$). Moreover, the residuals are different among degree groups in all three populations (non-parametric ANOVA; P ranging from 0.0017 to 0.0089; Supplementary Table 7; Supplementary Figure 10), and these differences are due to a trend towards higher residuals in groups corresponding to higher degree (linear trend test on ranks; P ranging from 0.0007 to 0.0016; Supplementary Table 7; Supplementary Figure 10). Therefore, the association observed between Z_F scores and degree cannot be attributed to the association between network centrality and the action of BGS.

References

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. 18:1585–1592.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics. 134:1289–303.

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. Genome Res. 20:393–402.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. 17:68–74.

Grossman SR et al. 2010. A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. Science. 327:883–886.

Grossman SR et al. 2013. Identifying recent adaptations in large-scale genomic data. Cell. 152:703–13.

Karolchik D, Hinrichs AS, Kent WJ. 2009. The UCSC Genome Browser. Curr. Protoc. Bioinforma. Chapter1:Unit1:4.

Keshava Prasad TS et al. 2009. Human Protein Reference Database--2009 update. Nucleic Acids Res. 37:D767–72.

Kim PM, Korbel JO, Gerstein MB. 2007. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. Proc. Natl. Acad. Sci. 104:20274–20279.

Kosiol C et al. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet. 4:e1000144.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat. Rev. Genet. 7:337–48.

Stark C et al. 2011. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 39:D698–704.

Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics. 168:373–81.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature. 467:1061–1073.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:e72.

Zeng K, Shi S, Wu C-I. 2007. Compound tests for the detection of hitchhiking under positive selection. Mol. Biol. Evol. 24:1898–908.

Supplementary Table 1. Number of interactions (Degree) for genes with putative signal of positive selection test and for the others.

	Humans[a]				
	Global[b]	YRI	CEU	CHB	Mammals[c]
Mean degree for genes with signals of positive selection	9.637	10.34	8.844	9.263	7.578
Mean degree for genes without signals of positive selection	8.107	8.438	8.526	8.456	9.122
Permutation test (*P*-value)[d]	0.0254*	0.0108*	0.2862	0.0929	0.0067**

[a] Positive selection is invoked when the *P*-value associated to Z_F score is below 5%.
[b] Positive is invoked at global level when the *P*-value associated with the Z_F score is below 5% in at least one of the three studied populations (YRI, CEU or CHB).
[a] Positive selection is invoked when the *P*-value associated to $2\Delta\ell$ score is below 5%.
[d] *P*-values were calculated using permutations. In each permutation a set of genes is randomly drawn, with the sampling size corresponding to the number of genes with signals of positive selection. Then, the average of their degree is compared to the one obtained for the genes with signals of positive selection. *P*- values are computed as the proportion of permutations with an average degree higher or equal to the observed one.
*: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.

Supplementary Table 2. Relationship between degree and the likelihood of havingevolved under recent positive selection in human populations as estimated from four different statistics.

			YRI	CEU	CHB
iHS	Spearman correlation[a]	*ρ*	0.0347	0.0433	0.0357
		P-value	0.0014**	6.79×10^{-05}***	0.0010**
	Non-parametric ANOVA[b]	*F*	4.609	5.761	4.942
		P-value	0.0032**	0.0006***	0.0020**
	Trend test on ranks[b]	*F*	9.844	15.65	10.41
		P-value	0.0017**	7.70×10^{-05}***	0.0013**
XPCLR	Spearman correlation[a]	*ρ*	0.0385	0.0149	0.0282
		P-value	0.0007***	0.1906	0.0127*
	Non-parametric ANOVA[b]	*F*	3.241	0.7999	1.592
		P-value	0.0211*	0.4941	0.0849
	Trend test on ranks[b]	*F*	8.918	1.999	2-209
		P-value	0.0028**	0.1574	0.0236
DH	Spearman correlation[a]	*ρ*	0.0343	0.0360	0.0436
		P-value	0.0015**	0.0009***	5.82×10^{-05}***
	Non-parametric ANOVA[b]	*F*	3.206	3.648	4.612
		P-value	0.0222*	0.0121*	0.0032**
	Trend teston ranks[b]	*F*	8.980	9.467	12.91
		P-value	0.0027**	0.0021**	0.0003***
CMS	Spearman correlation[a]	*ρ*	0.0700	0.0566	0.0368
		P-value	1.61×10^{-10}***	2.59×10^{-07}***	0.0009***
	Non-parametric ANOVA[b]	*F*	12.46	8.597	3.945
		P-value	1.09×10^{-09}***	1.07×10^{-05}***	0.0080**
	Trend test on ranks[b]	*F*	37.24	25.61	9.998
		P-value	1.09×10^{-09}***	4.27×10^{-07}***	0.0016**

[a] Spearman correlation between degree and the positive selection score given in the fist column.

High scores indicate a higher probability to have evolved under positive selection.

[b] Non-parametric ANOVA and trend tests on ranks performed to contrast whether the medians of the positive selection scores are equal across the degree groups.

*: *P* < 0.05; **: *P* < 0.01; ***: *P* < 0.001.

Supplementary Table 3. Relationship between degree and the likelihood of having evolved under positive selection in human populations using a subset of independent genes.

		YRI	CEU	CHB
Spearman correlation[a]	ρ	0.0450	0.0600	0.0605
	P-value	0.0248*	0.0017**	0.0014**
Non-parametric ANOVA[b]	*F*	1.320	4.521	4.000
	P-value	0.2661	0.0036**	0.0075**
Trend test on ranks[b]	*F*	3.908	8.901	7.791
	P-value	0.0482*	0.0029**	0.0053**

We obtained a subset of most likely unlinked genes represented in the network containing 2,792, 3,106 and 3,107 genes in YRI, CEU and CHB, respectively, by randomly sampling one network gene located between two recombination hotspots (defined as a region where the observed recombination rates is greater than 10 times the genome recombination rate average).

[a] Spearman correlation between degree and Z_F in YRI, CEU and CHB.
High Z_F scores indicate a higher probability of having evolved under positive selection.
[b] Non-parametric ANOVA and trend tests on ranks performed to contrast whether the medians of the Z_F score are equal across the degree groups (calculated on the whole set of genes).
*: $P < 0.05$; **: $P < 0.01$.

Supplementary Table 4. Relationship between degree and the impact of natural selection using two alternative high-quality protein-protein interaction networks.

			Positive selection				Purifying selection	
			YRI	CEU	CHB	Mammals	Humans	Mammals
High-Quality network from BioGRID	Spearman correlation[a]	*ρ*	0.0379	0.0420	0.0441	-0.0620	-0.0715	-0.1730
		P-value	0.0034**	0.0012**	0.0006***	2.53×10^{-05}***	5.82×10^{-09}***	5.50×10^{-32}***
	Non-parametric ANOVA[b]	*F*	3.819	4.194	4.237	5.363	10.29	45.42
		P-value	0.0095**	0.0057**	0.0053**	0.0011**	9.35×10^{-07}***	6.44×10^{-29}***
	Trend test on ranks[b]	*F*	9.492	9.504	8.995	14.91	28.10	131.2
		P-value	0.0021**	0.0021**	0.0027**	0.0001***	1.19×10^{-07}***	5.39×10^{-30}***
Network from Human Protein Reference Database	Spearman correlation[a]	*ρ*	0.0257	0.0365	0.0229	-0.0577	-0.0488	-0.1353
		P-value	0.0294*	0.0021**	0.0533	1.04×10^{-05}***	6.88×10^{-06}***	7.97×10^{-21}***
	Non-parametric ANOVA[b]	*F*	2.223	2.787	1.443	5.733	5.729	27.39
		P-value	0.0833	0.0392*	0.2281	0.0006***	0.0007***	1.49×10^{-17}***
	Trend test on ranks[b]	*F*	4.380	7.765	3.463	16.90	15.03	80.29
		P-value	0.0364*	0.0053**	0.0628	3.99×10^{-05}***	0.0001	4.56×10^{-19}***

[a] Spearman correlation between degree and selection scores (Z_F for positive selection in YRI, CEU and CHB populations; $2\Delta\ell$ for positive selection in mammals; DAF for purifying selection in humans; and ω for purifying selection in mammals). High Z_F and $2\Delta\ell$ scores indicate a higher probability of having evolved under positive selection. Low DAF and ω scores indicate higher selective constraint during human and mammalian evolution, respectively.

[b] Non-parametric ANOVA and trend tests on ranks performed to contrast whether the medians of the natural selection scores are equal across the degree groups.

*: *P*-value < 0.05; **: *P*-value < 0.01; ***: *P*-value < 0.001.

Supplementary Table 5. Relationship between network centrality and the impact of natural selection using betweenness and closeness.

			Positive selection				Purifying selection	
			YRI	CEU	CHB	Mammals	Humans	Mammals
Betweenness	Spearman correlation[a]	ρ	0.0295	0.0353	0.0385	-0.0645	-0.0641	-0.1848
		P-value	0.0096**	0.0020**	0.0007***	6.84×10^{-07}***	3.15×10^{-09}***	4.92×10^{-46}***
	Non-parametric ANOVA[b]	*F*	3.406	4.445	5.728	11.33	15.29	91.33
		P-value	0.0332*	0.0118*	0.0033**	1.23×10^{-05}***	2.36×10^{-07}***	8.77×10^{-40}***
	Trend test on ranks[b]	*F*	5.835	8.853	10.69	22.41	30.55	182.0
		P-value	0.0157*	0.0029**	0.0011**	2.26×10^{-06}***	3.34×10^{-08}***	7.23×10^{-41}***
Closeness	Spearman correlation[a]	ρ	0.0413	0.0332	0.0427	-0.0726	-0.0802	-0.1679
		P-value	0.0003***	0.0037**	0.0002***	2.28×10^{-08}***	1.25×10^{-13}***	3.30×10^{-38}***
	Non-parametric ANOVA[b]	*F*	3.839	2.362	3.698	9.923	16.73	54.87
		P-value	0.0093**	0.0693	0.0113*	1.60×10^{-06}***	7.82×10^{-11}***	5.71×10^{-35}***
	Trend test on ranks[b]	*F*	9.702	6.809	10.59	26.40	49.72	153.9
		P-value	0.0018**	0.0091**	0.0011**	2.85×10^{-07}***	1.91×10^{-12}***	6.81×10^{-35}***

[a] Spearman correlation between degree and natural selection scores (Z_F for positive selection in the YRI, CEU and CHB populations; $2\Delta\ell$ for positive selection in mammals; DAF for purifying selection in humans; and ω for purifying selection in mammals). High Z_F and $2\Delta\ell$ scores indicate a higher probability of having evolved under positive selection, respectively. Low DAF and ω scores indicate higher selective constraint during human and mammalian evolution, respectively.

[b] Non-parametric ANOVA and trend tests on ranks performed to contrast whether the medians of the natural selection scores are equal across the connectivity measure groups. For Betweenness, the 1st and 2nd quartiles were merged due to the uneven distribution of values.

*: *P*-value < 0.05; **: *P*-value < 0.01; ***: *P*-value < 0.001.

Supplementary Table 6. Relationship between degree and the impact of recent positive selection in human populations controlling for ω in mammals.

		YRI	CEU	CHB
Spearman correlation[a]	ρ	0.0432	0.0195	0.0412
	P-value	0.0020**	0.1655	0.0032**
Non-parametric ANOVA[b]	*F*	2.51	2.827	2.499
	P-value	0.0569	0.0371*	0.0578
Trend test on ranks[b]	*F*	7.012	2.032	6.646
	P-value	0.0081**	0.1541	0.0100**

In order to test for an association between degree and the impact of positive selection in humans controlling for ω, we used the Z_F as the likelihood of having been targeted by positive selection. We then applied a linear regression between this score and ω. High Z_F values indicate a higher probability of having evolved under positive selection. Low ω scores indicate higher selective constraint. The linear regression residuals were then used to perform the Spearman's correlation analysis, the non-parametric ANOVA and the linear trend test on rank.

[a] Spearman correlation between degree and the residuals.

[b] Non-parametruc ANOVA and trend tests performed to contrast whether the medians of the residuals across the degree groups.

*: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.

Supplementary Table 7. Relationship between degree and the impact of recent positive selection in human populations controlling for covariates of background selection.

		YRI	CEU	CHB
Spearman correlation[a]	ρ	0.0427	0.0369	0.0428
	P-value	0.0002***	0.0013**	0.0002***
Non-parametric ANOVA[b]	*F*	3.872	5.043	4.110
	P-value	0.0089**	0.0017**	0.0064**
Trend test on ranks[b]	*F*	11.53	9.947	11.48
	P-value	0.0007***	0.0016**	0.0007***

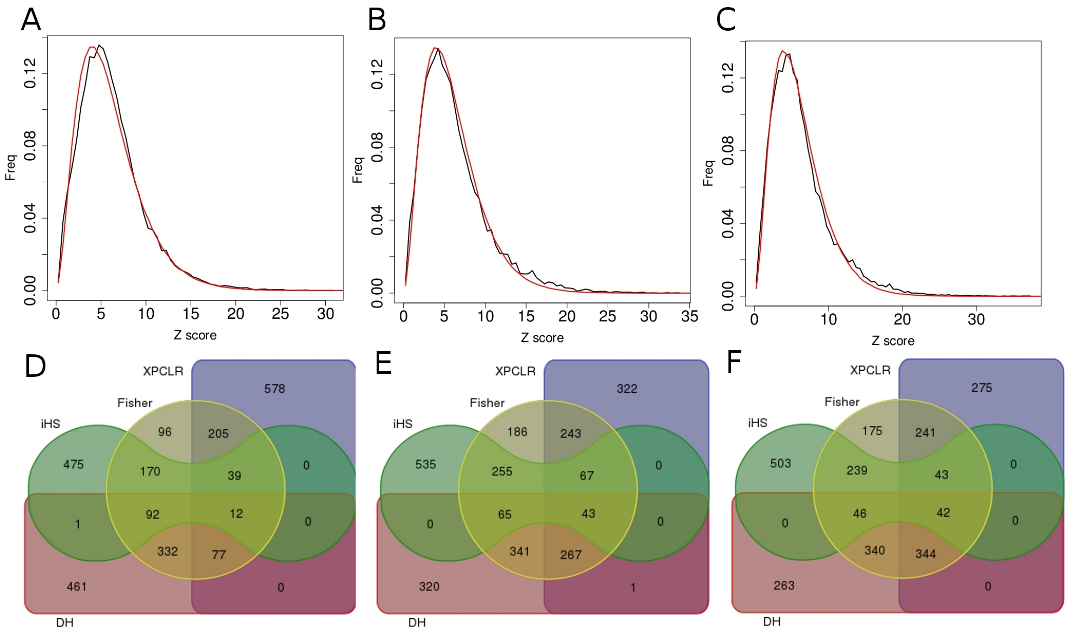
In order to test for an association between degree and the impact of positive selection in humans controlling for background selection, we used Z_F as the likelihood of having been targeted by positive selection. We then applied a linear regression between this score and both population-specific recombination rate average across the gene and GC content average across the gene. High Z_F values indicate a higher probability of having evolved under positive selection.

The linear regression residuals were then used to perform the Spearman's correlation analysis, the ANOVA and the linear trend test.

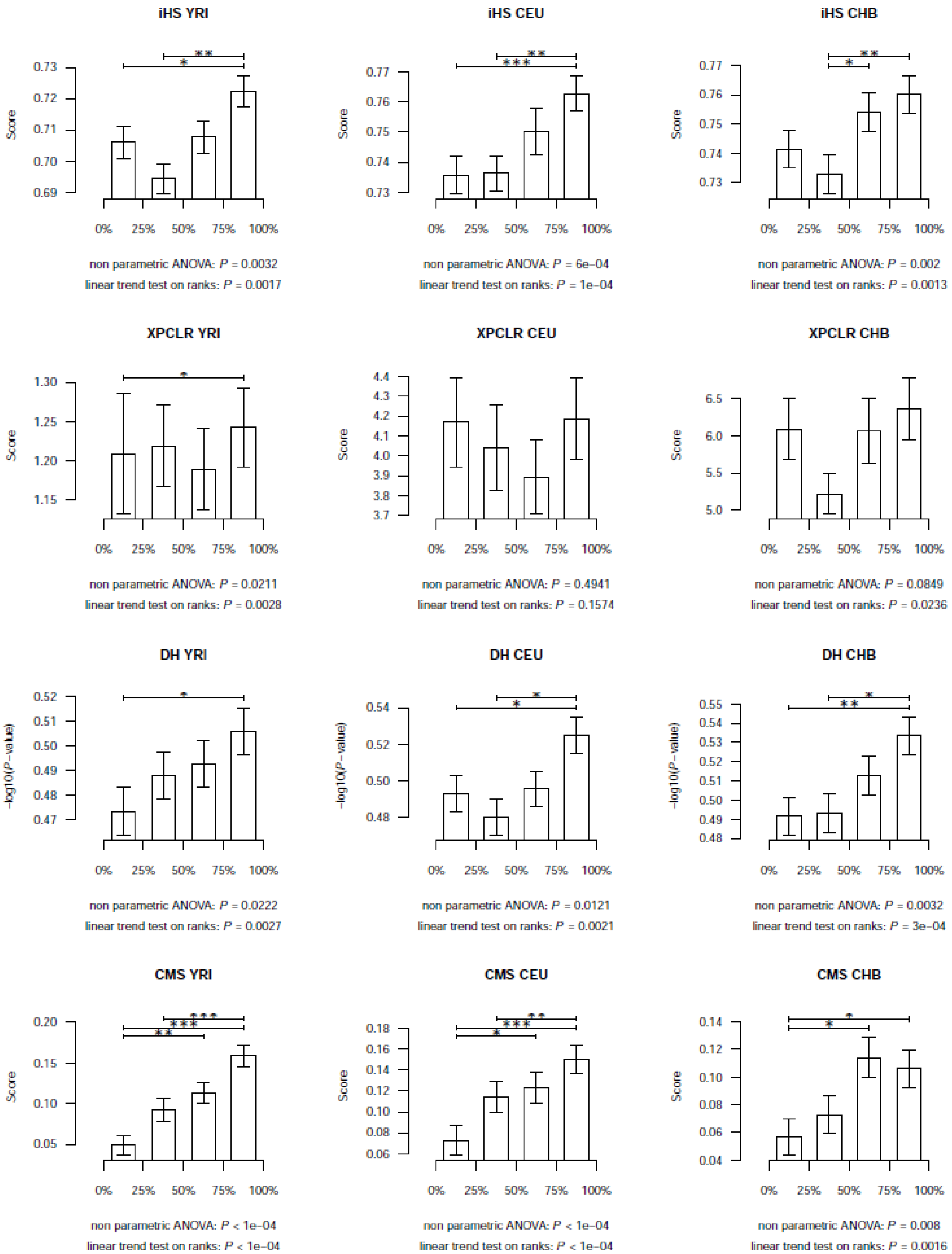
[a] Spearman correlation between degree and the residuals.

[b] Non-parametric ANOVA and trend tests performed to contrast whether the medians of the residuals across the degree groups.

*: *P* < 0.05; **: *P* < 0.01; ***: *P* < 0.001.



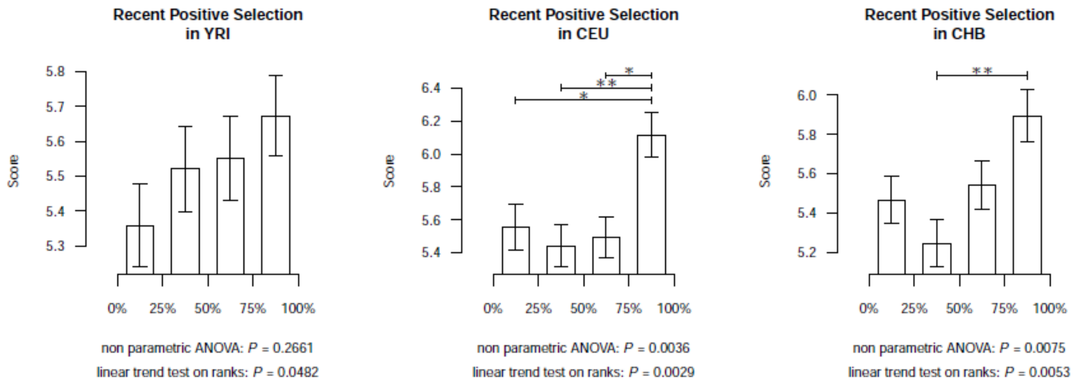
Supplementary Figure 1. Confirmation of the accuracy of the Fisher's combination test score. A-C: Comparison of the Fisher's combination Z_F score distribution observed for the genes within the interactome and the genome background set (in black) to the $\chi^2_{(6)}$ expected distribution (in red) in YRI, CEU and CHB populations, respectively. **D-F**: Venn diagram of the genes with a signal of positive selection (*P*-value < 0.05) for the four tests in YRI, CEU and CHB, respectively.



Supplementary Figure 2. Impact of positive selection in human populations as measured by four different tests based on polymorphism data among groups of genes divided according to the degree quartiles.

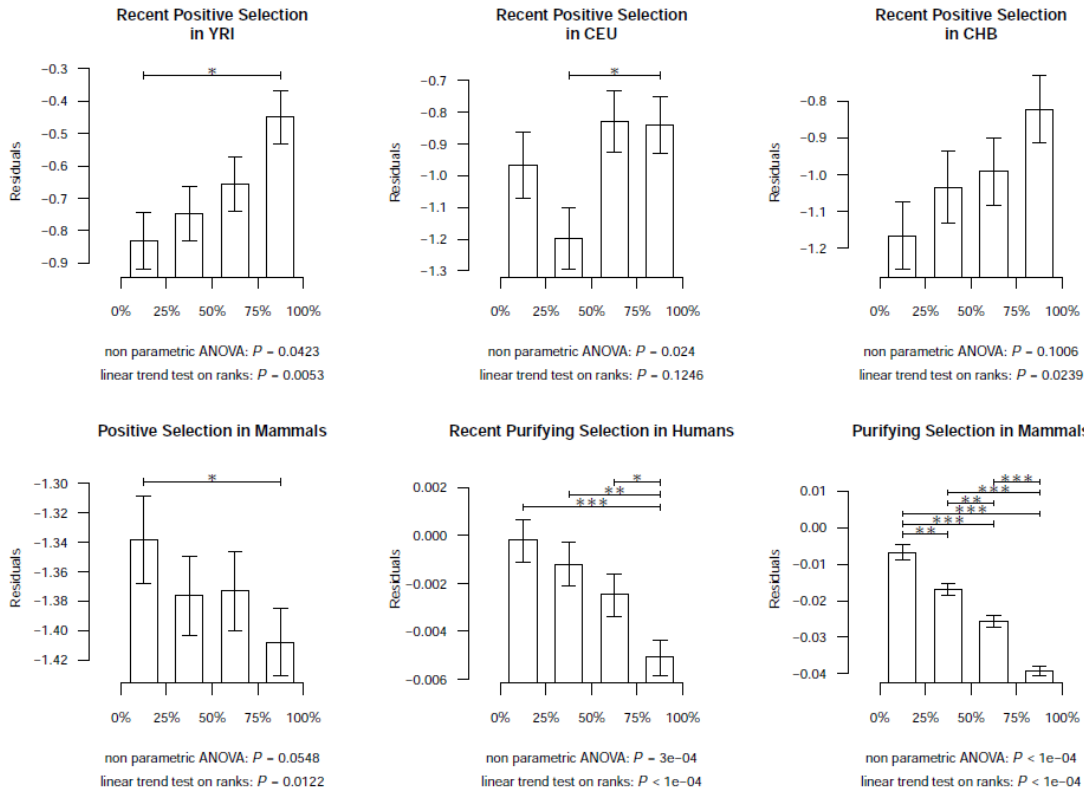
Genes were classified into four groups according to the degree quartiles calculated in the network.

The median of the positive selection score ± one median absolute deviation within each group is represented in the *y*-axis. A non-parametric ANOVA analysis was performed to contrast whether the medians of the positive selection scores were equal across the groups. A trend test on ranks was also been carried out to test for a linear relationship between the four quartiles (encoded from 1 to 4) and natural selection scores. A Tukey's honestly significant difference test was further applied to test for all pairwise differences. Significantly different pairs are marked with asterisks. *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.



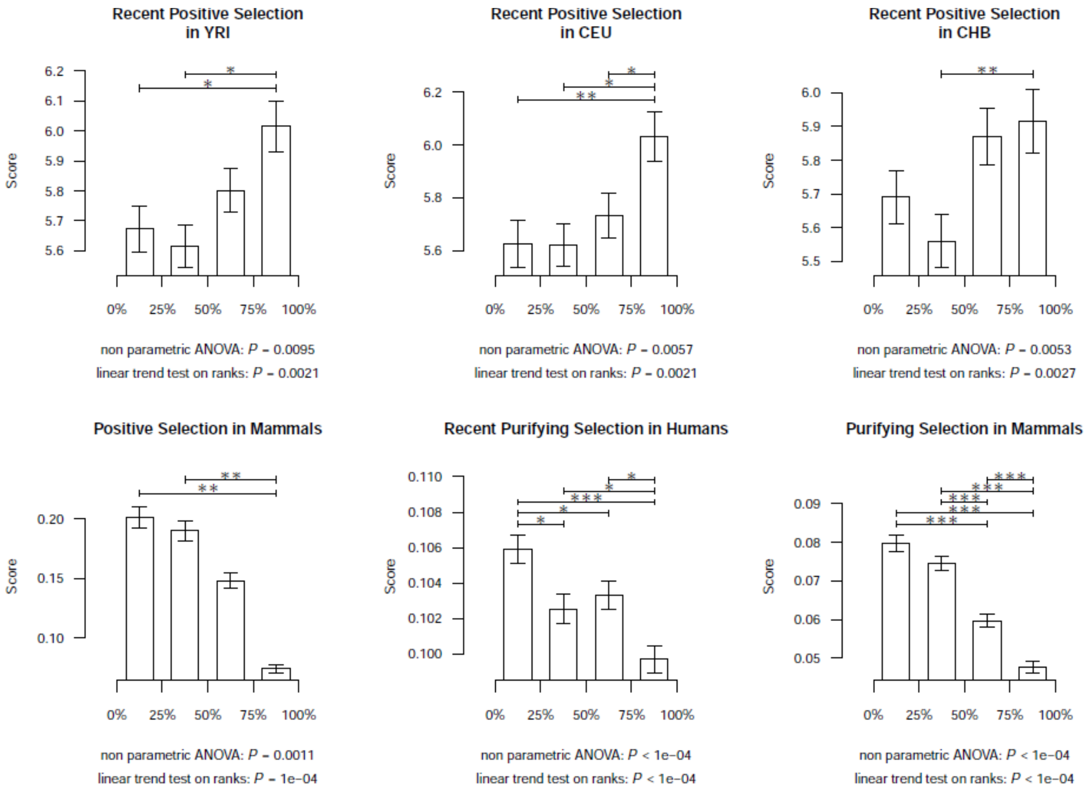
Supplementary Figure 3. Impact of positive selection during recent human evolution among groups of genes classified according to their degree using a subset of independently evolving genes.

We obtained a subset of most likely unlinked genes involved in the network containing 2,793, 3,107 and 3,108 genes in YRI, CEU and CHB, respectively, by randomly sampling one network gene located between two recombination hotspots (defined as a region where the observed recombination rates is greater than 10 times the genome recombination rate average). Genes were classified into four groups according to the degree quartiles. The median of the Z_F scores ± one median absolute deviation within each group is represented in the *y*-axis. A non-parametric ANOVA analysis was performed to contrast whether the median scores are equal across the groups. A trend test on ranks was also carried out to test for a linear relationship between the four groups (encoded from 1 to 4) and natural selection scores. A Tukey's honestly significant difference test was further applied to test for all pairwise differences. Significantly different pairs are marked with asterisks. *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.



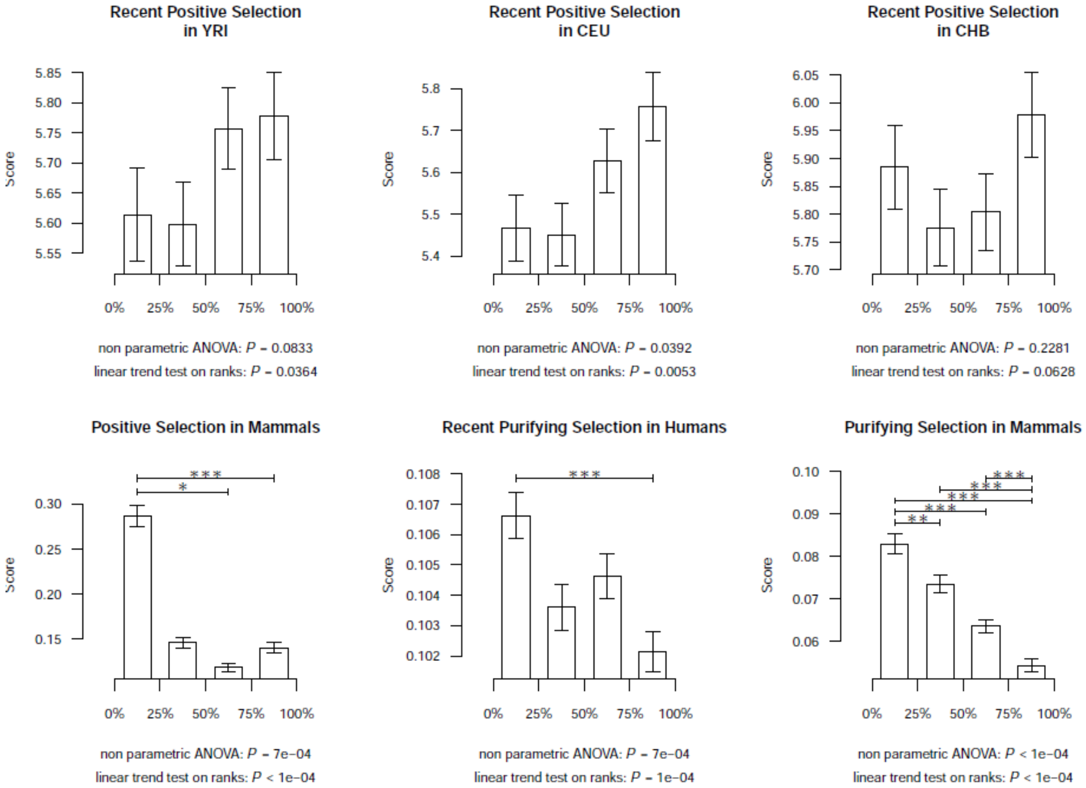
Supplementary Figure 4. Impact of natural selection among groups of genes classified according to their degree controlling for confounding factors.

Z_F and $2\Delta\ell$ were used to estimate the impact of positive selection in human populations and in mammals, respectively. DAF and ω were used to estimate the impact of purifying selection in human populations and in mammals, respectively. In order to test for an association between degree and positive selection scores controlling simultaneously for protein length, expression level and expression breadth, we applied a linear regression between positive selection scores and these factors. The linear regression residuals were then used to perform the Spearman's correlation analysis, the non-parametric ANOVA and the linear trend test on ranks. Genes were classified into four groups according to the degree quartiles. The median of the residuals ± one median absolute deviation within each group are represented in the *y*-axis. A non-parametric ANOVA analysis was performed to contrast whether the medians of the scores are equal across the groups. A trend test was carried out to test for a linear relationship between the four groups (encoded from 1 to 4) and natural selection scores. A Tukey's honestly significant difference test was further applied to test for all pairwise differences. Significantly different pairs are marked with asterisks. *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.



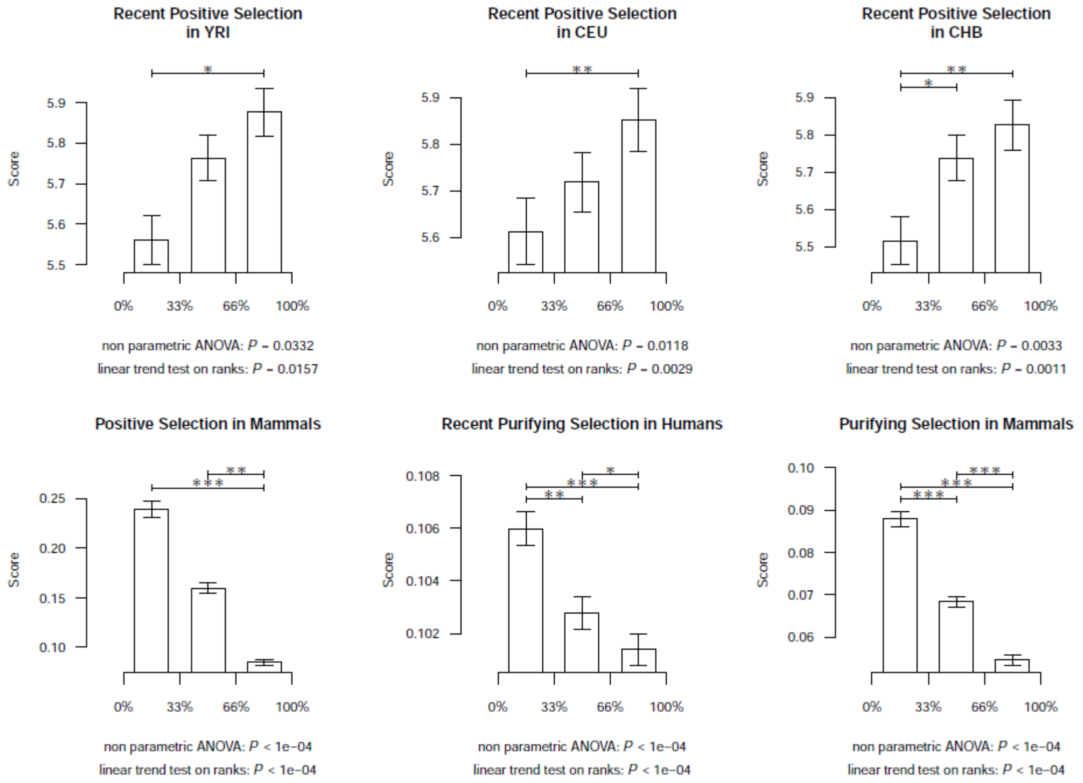
Supplementary Figure 5. Impact of natural selection among groups of genes classified according to their degree in the BioGRID high quality network.

Genes were classified into four groups according to the degree quartiles calculated in the network HQ. The median of the positive selection score used as likelihood of having been targeted by natural selection ± one median absolute deviation within each group is represented across the *y*-axis. Z_{F} and $2\Delta\ell$ were used to infer the impact of positive selection in human populations and in mammals, respectively. DAF and ω were used to estimate the impact of purifying selection in human populations and in mammals, respectively. A non-parametric ANOVA analysis was been performed to contrast whether the medians of the scores were equal across the groups. A trend test on ranks was also carried out to test for a linear relationship between the four groups (encoded from 1 to 4) and natural selection scores. A Tukey's honestly significant difference test was further applied to test for all pairwise differences. Significantly different pairs are marked with asterisks. *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.



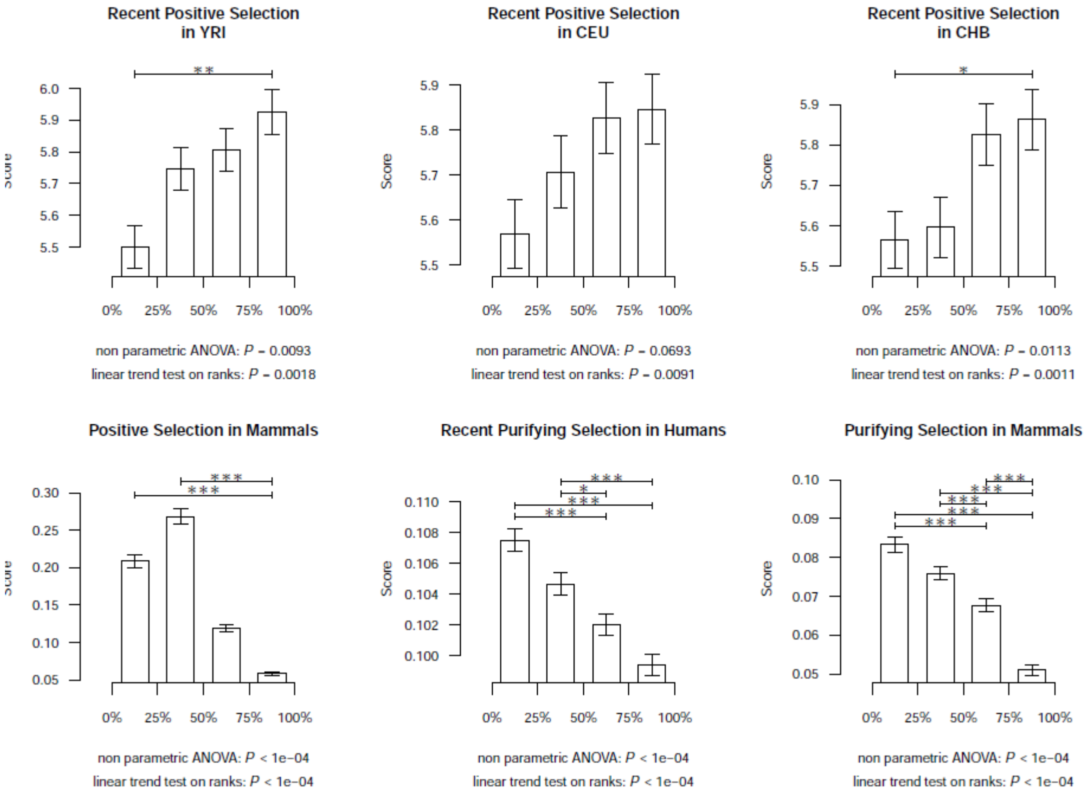
Supplementary Figure 6. Impact of natural selection among groups of genes classified according to their degree in the Human Protein Reference Database network.

Genes were classified into four groups according to the degree quartiles calculated in the HPRD network. The median of the positive selection scores ± one median absolute deviation within each group is represented across the *y*-axis. Z_F and $2\Delta\ell$ were used to estimate the likelihood of having been targeted by positive selection in human populations and in mammals, respectively. DAF and ω were used to estimate the impact of purifying selection in human populations and in mammals, respectively. A non-parametric ANOVA analysis was performed to contrast whether the medians of the positive selection scores are equal across the groups. A trend test on ranks was also carried out to test for a linear relationship between the four groups (encoded from 1 to 4) and natural selection scores. A Tukey's honestly significant difference test was further applied to test for all pairwise differences. Significantly different pairs are marked with asterisks. *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.



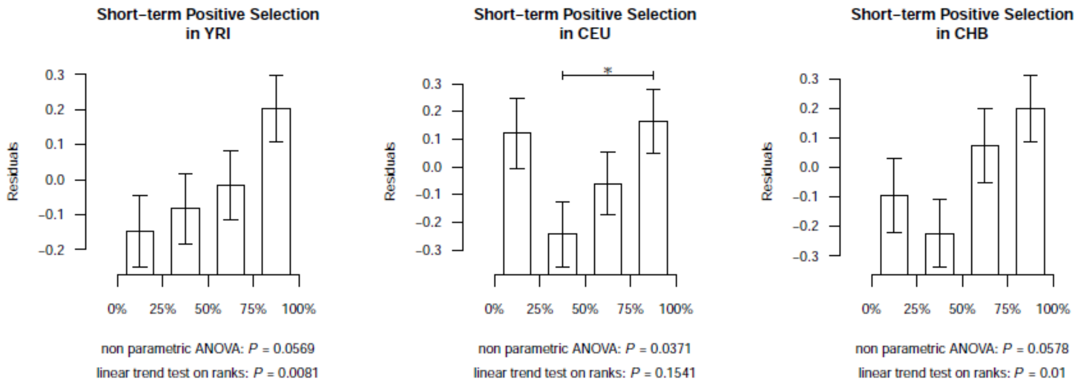
Supplementary Figure 7. Impact of natural selection among groups of genes classified according to their betweenness in the BioGRID network.

Genes were classified into four groups according to the betweenness quartiles calculated in the interactome. The 1st and 2nd groups were merged due to the uneven distribution of values. The median of the positive selection scores ± one median absolute deviation within each group is represented across the *y*-axis. Z_F and $2\Delta\ell$ were used to estimate the likelihood of having been targeted by positive selection in human populations and in mammals, respectively. DAF and ω were used to estimate the impact of purifying selection in human populations and in mammals, respectively. A non-parametric ANOVA analysis was performed to contrast whether the medians of the scores are equal across the groups. A trend test on ranks was also carried out to test for a linear relationship between the four quartiles (encoded from 1 to 3) and natural selection scores. A Tukey's honestly significant difference test has been further applied to test for all pairwise differences. Significantly different pairs are marked with asterisks. *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.



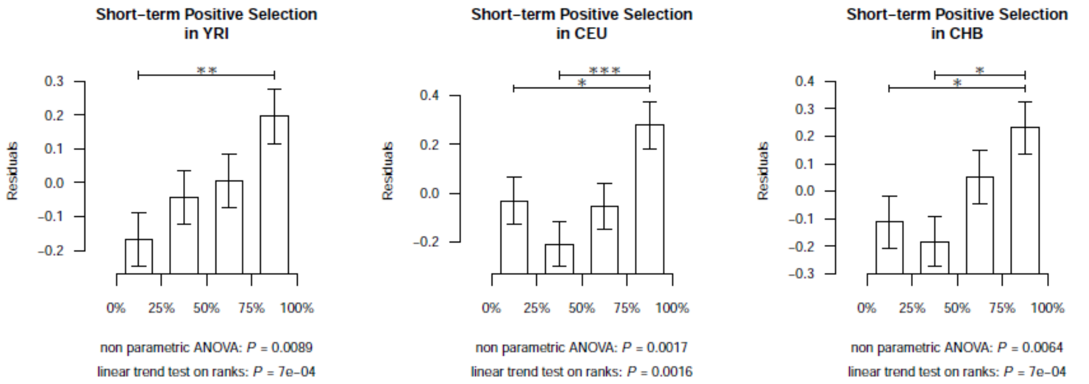
Supplementary Figure 8. Impact of natural selection among groups of genes classified according to their closeness in the BioGRID network.

Genes were classified into four groups according to the closeness quartiles. The median of the positive selection scores ± one median absolute deviation within each group is represented across the *y*-axis. Z_F and $2\Delta\ell$ were used to estimate the likelihood of having been targeted by positive selection in human populations and in mammals, respectively. DAF and ω were used to estimate the impact of purifying selection in human populations and in mammals, respectively. A non-parametric ANOVA analysis was performed to contrast whether the medians of the scores are equal across the groups. A trend test on ranks was also carried out to test for a linear relationship between the four quartiles (encoded from 1 to 4) and natural selection scores. A Tukey's honestly significant difference test was further applied to test for all pairwise differences. Significantly different pairs are marked with asterisks. *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.



Supplementary Figure 9. Impact of positive selection during recent human evolution among groups of genes classified according to their degree controlling for the effect of purifying selection during mammalian evolution.

In order to test for an association between degree and the Z_F score controlling for purifying selection, we applied a linear regression between Z_F and ω. The linear regression residuals were then used in a Spearman's correlation analysis, a non-parametric ANOVA and a linear trend test on ranks. Genes were classified into four groups according to the degree quartiles. The median of the residuals ± one median absolute deviation within each group is represented across the *y*-axis. A Tukey's honestly significant difference test was further applied to test for all pairwise differences. Significantly different pairs are marked with an asterisk. *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.



Supplementary Figure 10. Impact of positive selection during human evolution among groups of genes divided according to their degree controlling for covariates of background selection. In order to test for an association between degree and the Z_F scores controlling for background selection, we applied a linear regression between Z_F and both the GC content and the average population-specific recombination rate across the gene. The linear regression residuals were then used in a Spearman's correlation analysis, a non-parametric ANOVA and a linear trend test on ranks. Genes were classified into four groups according to their degree. The median of the residuals ± one median absolute deviation within each group is represented across the *y*-axis. A Tukey's honestly significant difference test was further applied to test for all pairwise differences. Significantly different pairs are marked with asterisks. *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.

A MACHINE-LEARNING FRAMEWORK TO DETECT AND CLASSIFY HARD SELECTIVE SWEEPS IN HUMAN POPULATIONS.

Marc Pybus, Pierre Luisi, Giovanni Dall'Olio1, Manu Uzkudun, Hafid Laayouni, Jaume Bertranpetit, Johannes Engelken

In Preparation

A Machine-Learning Framework to Detect and Classify Hard Selective Sweeps in Human Populations

Marc Pybus[1,+] , Pierre Luisi[1,+], Giovanni Dall’Olio[1,2,+] , Manu Uzkudun[1] , Hafid Laayouni[1] , Jaume Bertranpetit[1,*], Johannes Engelken[1,*]

[1] Institut de Biologia Evolutiva (UPF-CSIC), Parc de Recerca Biomèdica de Barcelona (PRBB), Dr Aiguader, 88, 08003, Barcelona, Catalonia, Spain.

[2] Division of Cancer Studies, King's College of London, London (UK)

+ Those authors contributed equally to this work

* Corresponding authors

Jaume Bertranpetit:

IBE, Institut de Biologia Evolutiva (UPF-CSIC)
CEXS-Universitat Pompeu Fabra-PRBB
Doctor Aiguader 88
08003 Barcelona, Catalonia, Spain

email: jaume.bertranpetit@upf.edu

Johannes Engelken:

IBE, Institut de Biologia Evolutiva (UPF-CSIC)
CEXS-Universitat Pompeu Fabra-PRBB
Doctor Aiguader 88
08003 Barcelona, Catalonia, Spain

email: johannes.engelken@upf.edu

Abstract (or Author Summary)

Detecting Darwinian selection in human genomic regions has been a recurrent topic in human population genetic studies. Over the years, many positive selection tests have been implemented to highlight specific genomic patterns left by a selective event when compared to neutral expectations. However, there is little consistency among the regions detected in several genome-wide scans using different tests: population-specific demographic dynamics, local genomic features or different types of selection acting along the genome might explain such discrepancies.

We have implemented a machine-learning classification framework that exploits the combined ability of some positive selection tests to uncover different features of a given selective sweep (such as completeness and oldness). Our simulation-calibrated framework estimates composite scores of several positive selection tests while controlling for population-specific demographies within a hard sweep model context. As a result, we increase the sensitivity toward hard selective sweeps while adding insights about the completeness and oldness of the sweep. Our method also allows to interpret the relevance of a given positive selection test under specific selection scenarios. We calibrated and applied the method to three reference populations from The 1000 Genome Project to generate a genome-wide classification map of hard selective sweeps that can be used to find putative regions under positive selection in the human lineage. Different genomic patterns arise under specific demographies and different time-spanning hard selective sweeps and, probably, different selection models (soft, balancing). This study is aimed at improving the way a selective sweep is inferred by taking into account that such differences may exist and can be used as proxies to understand better how natural selection has shaped our genome. We found very few signals of hard sweep in the African population analyzed, putatively appointing to alternative modes of adaptation at stake.

Author Summary (or Abstract)

Almost all the current methods to detect positive selection are designed to detect a very specific type of selective sweep: the recent strong hard selective sweep. A hypothetical beneficial mutation appears in a population (hard sweep model) and is increased in frequency until it reaches fixation in a relatively short period of time (strong selection coefficient). Allele fixation occurs at the present time (recent selective event), leaving no time for mutation and recombination to recover previous diversity levels and linkage disequilibrium patterns. While this type of sweeps appears to be common in some animal species (e.g. in some *Drosophila* species), in human populations seem to be not so common: few have been detected in Out-of-Africa populations and almost none have been found in African populations. Selection on standing variation (soft sweep model), balancing selection or different types of hard sweeps (such as partial or old sweeps) may have played a bigger role

in human evolution. Thus, we have developed a calibrated framework based on realistic simulations that builds composite scores from different positive selection tests to detect and classify different types of hard selective sweeps. With this new approach we are able to increase sensitivity toward almost-fixed selective sweeps and get insights about the relative oldness of a given selection signal. Once validated, we applied the method to empirical data of three reference populations from The 1000 Genomes Project to generate a genome-wide classification map of human hard sweeps. Very few signals were observed in the African population studied, while our method presents higher sensitivity in this population. In the future, our framework implementation could be used to include more types of selection (soft sweeps, balancing selection) so they could be properly classified and used to further understand how Darwinian evolution has shaped our genome.

Introduction

Over the last few decades, many different methods to detect positive selection in genomic regions have been developed. Such methods rely on the different genomic patterns left by an hypothetical selection event occurring in an idealized human population: a beneficial *de-novo* mutation arises and increases its frequency in a relatively low number of generations until it reaches population fixation at present times, in what was called the hard sweep model. This process leaves some characteristic patterns in the region surrounding the beneficial allele (selective sweep), such as skewed site frequency spectrum towards low frequency variants, long linkage disequilibrium haplotypes and population differentiation. Over the years, methods aimed at distinguishing such patterns have been developed and the genetic basis of some examples of human adaptation were confirmed, such as lactase persistence allele [1,2] or malaria resistance gene variants [3–6]. However, most of such methods usually lack consistency in reporting the same selective events along the genome [7]. This disagreement might appear due to the specific method capacity to uncover selection patterns under some local features of a local genomic region (such as specific recombination maps) or due to specific demographic dynamics of a given natural population. Thus, during the last decade, special effort was made to incorporate population-specific demographic models and region-specific recombination maps to approximate the neutral model to more complex scenarios [8]. However, while this approach improved the sensitivity to detect positive selection, it did not explain the lack concordance between methods. Other hypothesis pointed that other types of selection acting in the genome might explain such discrepancies, like selection on standing variation and polygenic adaptation [9] or balancing selection. Recently, a Bayesian method called CMS was developed and trained with neutral and selection simulations with the aim to integrate the signals of different positive selection tests in a composite score [10,11]. This Bayesian method combines the common signals shown by different positive selection tests under many selection scenarios so that the resulting sensitivity to general hard selective sweeps is increased. Nonetheless, the method did not try to uncover the specific features of a given selective sweep, such as the extent of completeness (final allele frequency of the selected allele) or the oldness of a selective event. We hypothesize that these internal features of the hard sweep model might explain, in part, the observed inconsistency between statistics signals so we can use them to uncover specific hard sweep features. Here, we applied a similar strategy (combination of neutrality scores through a machine-learning algorithm trained with simulations) with the difference that our framework works by automatically giving more weight to those tests that perform better at distinguishing between two simulated scenarios. This approach increases the sensitivity toward hard selective sweeps and uncovers some specific features of such sweeps. In the framework, we train different boosting models with very specific coalescent simulations to use the resulting regression functions as classification methods embedded in a hierarchical classification tree. A boosting algorithm is a widely used machine-learning algorithm that estimates a linear regression function of different input variables (here, positive selection test scores) so it maximizes the differences between two competing scenarios (for example, neutrality versus selection). Given an empirical genomic region for which we know their positive selection test scores, our framework sequentially applies different boosting functions in order to classify it based on the patterns observed when simulating different selection scenarios. Accordingly, we trained our framework with selection simulations with different final allele frequencies of the selected allele (completeness of a selective sweep) and with different time-spanning selective events (oldness of a selective sweep). After evaluating our method performance through independent simulations we applied it to empirical genome-wide data from The 1000 Genomes Project [12] with the objective of obtaining an improved genome-wide map of positive selection in three reference human populations. To our knowledge, this is the first genome-wide attempt to build a machine-learning classification method for hard sweeps in human populations.

A conservative average estimation of **0.47%** of the genome exhibiting signals of hard sweep in the three studied populations suggests that such selective events were rare during human evolution. Moreover, we detected **13-fold** and **15-fold** decrease of signals in the African population as compared to the European and East Asian populations, respectively, underlying the fact that adaptive processes may have been different within and out of Africa.

Materials and Methods

Reference Empirical Dataset

We downloaded genome-wide single nucleotide variant (SNV) data representing three continental populations -- Yoruba in Ibadan, Nigeria (YRI), Han Chinese in Beijing, China (CHB) and Utah residents with Northern and Western European ancestry, USA (CEU) -- from the low-coverage Phase I release (April 2012) of The 1000 Genomes Project [12]. After extracting all polymorphic SNVs we counted more than 24M segregating sites genome-wide. The SNV data was already phased by The 1000 Genomes Consortium and its phasing state was kept in other to apply haplotype-based statistics. We also downloaded both the ancestral allele state genome and the global genetic map provided by the consortium since this information is required by some of the positive selection statistics used.

Coalescent Simulations

Coalescent simulations covered a total of 46 scenarios (45 with selection + 1 neutral) simulating each one of the selected populations (CEU, CHB and YRI) under population-specific demographic models (**Figure 1**). Additionally, the same 46 scenarios were replicated to obtain an independent dataset that was used for the evaluation process. We used the software *cosi* version 1.2.1 [13] to generate such simulations since it includes a tuned human demography for three continental populations of Northern Europe, Asian and African ancestry (CEU, JPT/CHB, and YRI). In addition to the neutral scenario, this version of *cosi* is able to simulate classic selective sweeps (hard sweep model) under specific constraints [10]. Three parameters are required to simulate a hard selective sweep in *cosi*: selection coefficient, time when the sweep ends and final allele frequency of the selected allele. We set up these parameters to build 45 selection scenarios in a similar way to Grossman *et al.* 2010 [10]. *Cosi* does not allow any population effective size change or migration between populations while selection is occurring. Accordingly, we removed migration between populations from the provided "best fit" demographic model and simulated selective sweeps in a period when population effective sizes do not change in any population (between 10 Kya and 45 Kya). Generation time was set to 25 years, as described in the original paper. In short, nine classes of selective sweeps were simulated (grouped as Recent, Recent Long and Ancient selective sweeps) spanning different time periods between 10 Kya and 45 Kya, and for each class we also simulated five different final allele frequencies (FaF) for the selected allele (grouped as Complete, Incomplete and Partial selective sweeps). More details on the demographic and selection parameters used in our simulations are provided in **Supporting Text S1**. We also used the hotspot recombination model implemented in the simulation package (recosim) in order to obtain more realistic linkage disequilibrium patterns: this model incorporates features of recombination hotspots observed in human populations data instead of a mean genome-wide recombination rate. We computed 3000 replicates for the neutral scenario and 100 replicates for each one of the 45 selection scenarios. For each replicate, we simulated regions of 600 Kbp, to allow extended homozygosity statistics to calculate properly the EHH decay, and 97, 85 and 88 diploid individuals for CEU, CHB and YRI populations respectively, thus matching the sample size found in the reference empirical dataset. Note that our demographic model, contrary to Grossman *et al.* 2010 [10], allowed for the last agricultural population size increase as in the original *cosi* publication. *Cosi*'s 'best fit' demography was designed to produce sequence data that match different genome-wide evolutionary statistics distributions but it does not simulates the site frequency spectrum (SFS) bias towards low-frequency variants found in the low-coverage release of 1000 genomes Phase I [12]. Consequently, we applied a SFS thinning strategy to all simulations and replicates that randomly removed 48% of the singletons present in the all-together populations site frequency spectrum to adjust simulation data to the SFS found in the reference empirical dataset. **Figure 2** shows the difference in relative site frequency spectrum of empirical genome-wide data, original neutral simulations and thinned neutral simulations for each population.

Implemented Positive Selection Tests

In order to detect positive selection in empirical genome-wide and simulated SNV data we implemented an informatic pipeline including 21 different positive selection statistics (**Table 1**) that allowed us to parallelize the analysis in a computer cluster. Pipeline design and detailed descriptions of the tests can be found in Pybus *et al.*, 2014 [14]. All the included tests were applied as described in their respective papers except for XP-EHH and iHS algorithms, which were modified to speed up the calculations by increasing the EHH threshold from 0.05 to 0.15. Both modified versions were then verified to report signal of positive selection through simulations. Positive selection tests based on regions, such those based on allele frequency spectrum, were ran applying a sliding window approach, also described in Pybus *et al.*, 2014 [14]**.** For the empirical genome-wide data we developed a further parallelization strategy which consisted in splitting the genome-wide data in overlapping regions of 5Mb. Once the regions were analyzed with our 21 positive selection statistics, their outputs were re-merged seamlessly to retrieve concatenated genome-wide results. For the simulation datasets we ran each positive selection test to the 600 Kbp simulated sequence, although we only used the results from the central 25 Kbp region to train the algorithm and evaluate the results. For that purpose, we used the positive selection scores obtained from the central 25 Kbp region containing the selected allele. To get a unique score per region we used a specific summary statistic to each test so that it maximizes the selective sweep signal in the central region, as explained below. After the validation process, only 11 positive selection statistics were used to train the machine-learning algorithm as explained below.

Simulation and Positive Selection Test Validation

One important step to use this method in empirical data was to validate all the positive selection tests we implemented. By comparing statistics’ scores on neutral and selection simulations we were able to confirm that most of the tests were showing incremented signals on selective sweep simulations when compared to neutral ones. This approach confirmed their suitability to report regions under positive selection in a human demographic context (**Figure 3 and Supporting Figures S2**). 'Best fit' demography was fine-tuned to match SFS, Fst and LD decay in Schaffner *et al.* 2005 [13]. Having implemented those and more evolutionary statistics we wanted to check that the score distributions at neutral simulations of our positive selection tests were similar to those found in empirical genome-wide data. The genome-wide distributions were obtained from analyzing 13,969 autosomal 25 Kbp regions separated by a distance of 200 Kbp each other. We also selected 102 autosomal 25kb regions that have putatively evolved neutrally and dispersed throughout the genome. Those regions were selected to meet the following criteria: (1) to be a least 100 Kbp away from any known or predicted gene or expressed sequence tag or region transcribed into mRNA; (2) to be outside any segmental duplication or region transcribed into long noncoding RNA or conserved noncoding element (as defined in Woolfe et al. 2007 [15]); (3) to be distant from each other by at least 100 Kbp and not in LD each other. By manually checking the score distribution of the positive selection tests implemented in the framework, we confirmed a good correspondence between empirical and simulated datasets, thus confirming the suitability of the chosen demographic model. **Figure 4** shows distribution plots, box plots and violin plots for neutral simulations, 1000 genomes genome-wide data and the neutral subset of 1000 genomes data for **dDAF, Tajima's *D* and XP-EHH** statistics in CEU / EUR population as reference examples. The rest of the plots for the other statistics and populations can be found in **Supporting Figures S3**.

Score combination into window-based summary statistics

Positive selection statistics usually report scores for genomic regions of different lengths and, in some cases, for individual SNVs. However, a common region size was needed in order to compare and combine different positive selection tests within the boosting analysis. Accordingly, different region summarizing approaches were evaluated for each positive selection test. We chose to work with a region of 25 Kbp and considered as summary statistics the maximum, minimum or mean score across the region. To identify the best of these summary statistics for each positive selection test, we considered the 25 Kbp window located at the center of the simulated sequences, thus containing the selected allele. We then performed a sensitivity vs sensibility analysis comparing neutral simulations to estimate the false discovery rate to selection simulations with final allele frequencies of 0.8 and 1.0 to estimate the true positive rate. Thus, for each positive selection statistic we computed three Receiver Operating Characteristic Curves (ROC Curves) and, for each method, we chose summary statistics showing the highest Area Under the Curve (AUC) score (**Table 3**).

The Hierarchical Boosting Framework

We define a boosting function as a linear regression function of positive selection test scores that can be used as a classification method and that has been estimated through a boosting algorithm. In our framework, the four different boosting functions are sequentially considered within a hierarchical decision tree implementation (**Figure 5**). These boosting functions optimally combine each positive selection test ability to uncover specific properties of different selective sweeps. In order to estimate the different boosting functions, we grouped the 45+1 simulated datasets according to some common selective sweep features. After analyzing several simulations under different selection parameters we found that the main force driving statistics' scores was the final allele frequency of the selected allele (**Figure 3**). Hence, we decided to create groups of different selection scenarios according to the final allele frequency of the selected variant as main property (Complete: FAF=1.0, Incomplete: FAF=0.8 and 0.6, Partial: FAF=0.4 and 0.2 and Neutral scenarios) and then, according to the oldness of the simulated sweep (Recent: sweep ends 10 Kya, Ancient: sweep ends 30 Kya). Having our simulation scenarios groups defined (**Figure 5**) and their positive selection test scores calculated, we ran the machine-learning algorithm to train boosting functions which allowed to classify competing groups of scenarios. To do so we used the positive selection scores obtained from the central 25 Kbp region containing the selected allele using an specific summary statistic to each test so that it maximizes the selective sweep signal in the central region, as explained above (**Table 3**). Boosting algorithm estimates regression coefficients for input positive selection tests scores from two competing sets of scenarios (training datasets) so that the resulting regression score maximizes the differences between the two. This allows to set up a regression score threshold to, in turn, classify an unknown empirical or simulated genomic region for which we already know its individual positive selection scores. We systematically verified coefficient convergence for every estimated boosting function (**Supplementary Figures S8**). To circumvent a putative convergence to local instead of global optimal, and thus, to obtain a more robust regressions, we developed a bootstrapping strategy which is explained below. Thus, the estimated coefficients (**Figure 6**) have an amplitude according to their performance to distinguish among distinct scenarios. We used the mean coefficient value for each positive selection test to build our boosting functions. Then, using the reference set of simulation replicates at a given step, we calculated the thresholds for estimated regression scores that were needed to classify the evaluation datasets allowing 1% of false discovery rate. According to the chosen decision tree scheme, the iterative classification of a genomic region of interest is done as following:

(1a) if the Complete Boosting score is above the 99th percentile of the distribution of the Complete Boosting scores for the training simulations under Neutral scenario and Partial and Incomplete sweep scenarios, the region is classified as Complete Sweep and go to step 2a, otherwise go to step 1b.

(1b) if the Incomplete Boosting score is above the 99th percentile of the distribution of the Incomplete Boosting scores for the training simulations under Neutral scenario and Partial sweep scenarios, the region is classified as Incomplete Sweep and go to step 2b, otherwise go to step 1c.

(1c) If not classified at iteration 1a or 1b, the genomic region is left unclassified and the algorithm stops.

(2a) If the Ancient/Recent Complete Boosting score is above the 99th percentile of the distribution of the Ancient/ancient Complete Boosting scores for the training simulations under Complete Recent scenario the region is classified as Ancient Complete Sweep, while if it is below the 1th percentile of the distribution of the Ancient/Recent Complete Boosting scores for the training simulations under Complete Ancient scenario the region is classified as Recent Complete Sweep , otherwise the region remains only classified as Complete Sweep.

(2b) If the Ancient/Recent Incomplete Boosting score is above the 99th percentile of the distribution of the Ancient/ancient Incomplete Boosting scores for the training simulations under Incomplete Recent scenario the region is classified as Ancient Incomplete Sweep, while if it is below the 1th percentile of the distribution of the Ancient/Recent Incomplete Boosting scores for the training simulations under Incomplete Ancient scenario the region is classified as Recent Incomplete Sweep , otherwise the region remains only classified as Incomplete Sweep.

Two more alternative classification tree configurations were tested but showed lower performance (**Supplementary Text S4**).

Boosting Algorithm, Bootstrapping and Quality Control

We have used the boosting algorithm as exactly implemented in Lin *et al*, 2011 [16]: among different versions of boosting, they chose a logistic regression model with only one predictor a time as base procedure, thus permitting an easy interpretation of the relevance of each input variable. The loss of function used was the squared error loss function, as described in the manuscript. Because the boosting algorithm is predicted to be robust to overfitting we did not use an information-based iteration stopping criteria: we allowed boosting to iterate enough times until we observed that regression coefficients reached stable convergence. While testing different combinations of positive selection tests to build our boosting predictors we noted that, when highly correlated input variables were used, boosting algorithm never reached coefficient convergence. Thus, once we removed correlated positive selection statistics, we achieved coefficient convergence in only few hundreds of iterations (**Supplementary Figures S8**). The boosting algorithm implementation we used and other relevant functions are available in the R package *mboost* **[17]**. Each boosting was trained with different number of simulations depending on each scenarios different number of replicates (**Table 2**). We decided to apply a bootstrapping strategy to evaluate and correct differences in input sizes as well as to show the robustness of the coefficients estimated. We trained 1000 times each boosting algorithm performing a 90% resampling of input data from each competing scenarios. To build the final boosting regression function we used the mean coefficient values from the resampling procedure (**Figure 6**). All together, the following quality control analysis was performed to choose the positive selection statistics to be included in the boosting framework:

1. Comparison of distributions from Neutral simulations with Genome-wide and Putatively Neutral Real data: a poor match between the scores distributions of a given test in genome-wide empirical data compared with neutral or selection simulations was considered as strong reason for excluding such test from an empirical analysis. We removed Fu's F statistic according to this criteria (**Supporting Figures S3**).

2. Correlation among positive selection statistics and with recombination rate, GC content and read coverage: highly correlated input statistics had to be removed to facilitate coefficient convergence and avoid overfitting during the training process. We analyzed the correlation between statistics and other variables under neutral, selection and empirical genome-wide data. Purifying selection, through background selection (BGS), can produce signatures that can be confounded with positive selection by tests based on DNA polymorphism [18]. We controlled for such the putative confounding effect of BGS using genomic covariates (recombination rate and GC content). Moreover, the power of detecting rare variants depends on the read coverage [12], making the SFS-based tests putatively correlated of the sequencing depth. Those putative correlations could bias the final boosting score regarding local properties of the analyzed region. **Figure 9** shows the correlation analysis between positive selection tests and with recombination in empirical genome-wide data: **Wall's B, Wall's Q, Za, ZnS, R2, Fu & Li's F** were removed following this criteria. Population-specific correlation plots in empirical and simulated data are shown in **Supporting Figures S5.**

3. Intrinsic nature of the statistic: cross-population non-directional statistics show the same signal either selection is happening on the target population or on the reference one. **Fst** is a clear example of this, so it can not be used for training population-specific boosting algorithms. **XP-EHH** and **dDAF** are cross-population directional tests but since boosting algorithm can handle negative and positive values, all negative values for these two tests were set to zero to avoid confounding signals before the boosting training was applied. **XP-CLR** is a cross-population directional test reporting only positive values, so its output did not need to be modified. **Dh** statistic was discarded because it is not a positive selection test.

At the end, 11 positive selection statistics were selected as being suitable for a combined boosting analysis. **Table 4** summarizes the quality control process applied to all the implemented statistics and its outcome.

Results

Method Performance

Using an independent set of evaluation simulations (same parameters as the training dataset) we evaluated our framework performance through two methods: its ability to classify selection scenarios and its sensitivity compared to the positive selection tests used as input variables. To evaluate its classification power we calculated population-average false and true positive rates for each scenario in the evaluation dataset (**Table 5** and **Table 6**). Equivalent population-specific tables are found in **Supporting Text S6**. Note that unclassified cases are also taken into account yet we do not consider them as negative results. Population-averaged, our hierarchical boosting implementation was able to classify the evaluation scenarios with low false positive rates (**5.37%**). Many cases were left unclassified (false negatives, **28.14%**), making the hierarchical boosting a conservative method. It also showed different true positive rates depending on the scenario to classify: complete sweeps were easier to classify (**89.58%**) than incomplete sweeps (**43.04%**), explained because most of the positive selection tests were implemented to detect hard sweep upon important final allele frequency of the selected variants. Recent complete sweeps (**25.41%**) were better classified than ancient complete sweeps (**23.76%**). The same pattern was observed with incomplete recent sweeps (**18.89%**) and incomplete ancient sweeps (**11.00%**). A selection signal is expected to be stronger in recent sweeps because the recovery phase have less time to affect genomic region diversity. When looking at population-specific performance we noted that hierarchical boosting was performing better in the simulated African-ancestry population (**93.44%** for complete and **52.27%** for incomplete sweeps) than in the simulated Out-of-Africa populations (**87.64%** for complete and **38.97%** for incomplete sweeps). On the other hand, we compared independently the power of each statistic used in our boosting functions and the method itself. For that purpose we calculated the true positive rate observed among a group of selective sweep simulations at a given false discovery rate assessed on neutral simulations alone. We observed an improvement of our method to detect sweeps regardless final allele frequencies (**Figure 7, Supporting Text S6**). While individual positive selection statistics may show a comparable sensitivity to Complete and Incomplete boosting functions, like **XP-EHH** and **iHS** respectively, our method showed higher sensitivity at different final allele frequencies overall, probably because the unifying power of the hierarchical classification scheme. We controlled for confounding factors and local genomic features that might bias the application of the method to empirical genome-wide data. To evaluate the robustness of our method to such features we performed a correlation analysis between the obtained boosting scores and recombination rate, GC content and read coverage. Neither feature showed correlation with selected individual positive selection tests or the hierarchical boosting itself (**Supporting Figures S5**).

Application to 1000 Genomes Data

We applied our population-specific hierarchical boosting method to the reference empirical genome-wide data that was used to calibrate the simulations on which relies the framework. We obtained a list of 25 Kbp widows per population that were classified according to the different boosting functions described above (**Table 7).** Overall, higher number of windows were classified as Incomplete rather than Complete. This is expected as complete sweeps should be surrounded by incomplete signals due to the genetic hitchhiking effect. Yet more windows were classified as being Recent than Ancient, probably explained because the higher sensitivity of the positive selection tests towards recent selective events than to old ones. From a population-specific point of view, much less windows were reported to be under selection in YRI population than in the CEU and CHB populations. A population average of **0.47%** of over **103,000** genome-wide 25 kbp windows were classified as being either complete or incomplete (**CEU:0,6%, CHB:0.8%, YRI:0,03%**). We also observed more windows showing recent rather than ancient selection signal, as expected (**Table 7**). However, the number of 25 Kbp windows showing evidence of selection does not inform about the number of selective events that happened or are happening along the genome and that are detectable using this framework. The strength of selection and the recombination hotspot map of a local region determine how much a selective sweep signal would span for a given selective event as observed in many cases of positive selection. Thus, we implemented an algorithm that concatenates consecutive 25 Kbp windows according to their proximity, allowing for valley of non-significant scores as long as they do not contain any recombination hotspot (for details see **Supplementary Text S7**). After applying the algorithm we counted **27, 355** and **424** selective events in YRI, CEU and CHB populations, respectively **(Table 8; Supporting Table S9).** Roughly, a **13-fold** and **15-fold** decrease in the number of selective events was detected in YRI population with respect to CEU and CHB populations, respectively. Additionally, we classified the selective events according to boosting function scores showing significance in the genomic region encompassing the selective sweep signal (**Table 8**). One satisfying result is the excellent classification of the selective events detected. Indeed, we observed few signals with any ambiguity for the Class (Complete or Incomplete): only **10.2%** of the identified selective sweeps in any of the three populations (**0%, 7.6% and 12.1%** in only YRI, CEU and CHB, respectively) encompass significant scores for both Complete and Incomplete Boosting functions **(Table 8**). Moreover, those ambiguous signals exhibit a much longer size and lower proportion of significant scores as compared to unambiguous signals (**Supplementary Text 7**); hence, most of those signals may actually arise from different adjacent independent selective events. On the other hand, most of the selective events could not be assigned to a selective sweep Type (Ancient or Recent): **59.4%** of the identified selective sweeps in any of the three populations (**44.4%, 71.5%** and **50.2%** in only YRI, CEU and CHB, respectively) could be assigned a given Type. This demonstrates the difficulty to asses the oldness of a sweep. However, we have designed here a very conservative framework for that purpose, as demonstrated by the very low number of regions with a signal that has been assigned to both Recent and Ancient selective sweep (**0%, 0.6%, 2.4%** and **1.5%** in YRI, CEU, CHB and any of the three populations, respectively). Finally, as signaled when describing the results for individual 25 Kbp windows, we detected overall more than twice Recent sweeps than Ancient ones (**28.0%** and **11.0%**, respectively). However, this trends is clearly driven by selective events in CEU and CHB, while in YRI we observe a slightly higher number of Ancient selective sweeps than Recent ones (**29.6%** and **25.9%**, respectively)

We generated UCSC supertracks to easily visualize our hierarchical boosting results in any UCSC Genome Browser server. Visualizing selective sweeps in a genome browser helps to properly evaluate their genomic context and allows to interpret their strength and properties to propose putative candidate genes under positive selection. **Figure 8** shows four known examples of selection in the populations analyzed: *LCT* and *SLC24A5* genes in CEU population, *EDAR* gene in CHB and *DARC* gene in YRI. These supertracks and the raw hierarchical boosting results can be found in a local implementation of the UCSC Browser of our institute. We also provide raw scores of the 21 original positive selection tests used in this project and their estimated p-value under population-specific neutral expectations (**http://hsb.upf.edu/**). We did not attempted to speculate about the phenotypic consequences of the detected selective events because extensive functional analysis are usually needed to do so.

Discussion

Interpretation of the Estimated Boosting Functions

In order to properly compare the regression coefficients assigned to each positive selection statistics within a given boosting function, we standardized them as in Lin *et al.* 2011 [16]**.** We multiplied the estimated coefficients of each test by the square root of the variance observed in the related statistic distribution across replicates from both competing scenarios. The standardized coefficient assigned to a positive selection test in a boosting function (**Figure 6**) gives some insight on how well a given test is performing to distinguish between the two competing scenarios. We observe that to highlight complete sweeps against the other scenarios the two most important statistics are, in order of importance, **XP-EHH, dDAF, Fu & Li's D, Omega** and **CLR**. On the other hand, to detect incomplete sweeps, our boosting functions relied mostly on **iHS, XP-CLR** and **Fay & Hu's H**. For recent complete sweeps, **XP-EHH** contributed the most along with **EHH Average** and **Fu & Li's D**. Instead, for complete ancient sweeps, **dDAF** and **Tajima's D** are the ones more relevant. Within incomplete sweeps cases, **iHS** highlights recent selection patterns while **dDAF** defines more ancient ones. We observe that all three populations show very similar boosting function coefficients, indicating that the method is robust to continental human demography (**Figure 6**). These results are also in agreement with the estimated power of each statistic to detect a selective sweep with a given final allele frequency (**Supporting Text S6**).

Missing Hard Sweep Signals in Yoruba Population

We report less selective events or regions under selection (**13-fold** and **15-fold** reduction) in African-ancestry populations (YRI) than Out-of-Africa populations (CEU and CHB, respectively). A similar pattern was observed in previous works [19]. In these studies, the authors hypothesize that this lack of hard sweep signals might be attributed to (i) the SNV ascertainment bias present in the chip-array data they used in their studies or (ii) the use of positive selection tests with putatively less power in African ancestry populations because particular genomic features (shorter LD, higher diversity) . They also proposed as an alternative explanation, a scenario in which selection acted on standing variation. The present study is able to overcome both methodological caveats since (i) it relies on sequencing data without explicit ascertainment bias to a specific population, and (ii) the hierarchical boosting framework shows greater power to uncover selective events for African-ancestry populations than Out-of-Africa populations (EUR and ASN) as shown in **Supporting Text S6**. Assuming that there is no biological reason for African populations to have suffered from less selection pressure than Out-of-Africa populations, we suggest that this 10-fold difference is likely due to selection acting on standing variations (soft sweep), rather than *de-novo* mutations (hard sweep), segregating in African-ancestry populations and, unfortunately, out of the scope of our framework. Furthermore, the Out-of-Africa human diaspora likely occurred through serial founder effects, a specific case of population bottlenecks. Such demographic scenario seems to increase the rates of fixation of favored alleles [20]. This would imply that complete hard sweeps were more frequent Out-of-Africa, as observed in the present study. Alternatively, Wilson et al. 2014 [21] recently showed that population bottlenecks can lead a soft sweep to leave molecular footprints expected under the hard sweep model. Indeed, under this specific demographic scenario, it is likely that only one unique haplotype carrying the standing favored mutation is sampled. Unfortunately, the coalescent framework used here to generate training simulations does not allow the implementation of selective scenarios on a standing variant. Therefore, we could not assess whether the hard sweeps signals observed in Out-of-Africa populations are driven by selective sweep on *de-novo* or standing mutations, that is by hard or soft sweeps, respectively. Further work is necessary to include even more alternative selective scenarios in our framework implementation. This could provide more insights on the modes of adaptation at stake and their relative importance during human population evolution. However, this study already appoints that hard sweeps have definitely not been common in African populations (or at least in Yoruba studied here), and underlines the dramatic role of demography in understanding human adaptation.

References

1. Tishkoff S a, Reed F a, Ranciaro A, Voight BF, Babbitt CC, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39: 31–40. Available: http://www.ncbi.nlm.nih.gov/pubmed/17159977.

2. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, et al. (2004) Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. Am J Hum Genet 74: 1111–1120.

3. Tishkoff S a, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, et al. (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science (80-) 293: 455–462. Available: http://www.ncbi.nlm.nih.gov/pubmed/11423617. Accessed 30 April 2014.

4. Hamblin MT, Di Rienzo a (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66: 1669–1679. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1378024&tool=pmcentrez&rendertype=abstract.

5. Sabeti P, Usen S, Farhadian S, Jallow M, Doherty T, et al. (2002) CD40L association with protection from severe malaria. Genes Immun 3: 286–291. Available: http://www.ncbi.nlm.nih.gov/pubmed/12140747. Accessed 10 May 2014.

6. Ayodo G, Price AL, Keinan A, Ajwang A, Otieno MF, et al. (2007) Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. Am J Hum Genet 81: 234–242. doi:10.1086/519221.

7. Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19: 711–722. Available: http://www.ncbi.nlm.nih.gov/pubmed/19411596. Accessed 29 January 2013.

8. Sabeti P, C, Varilly P, Fry B, Lohmueller J, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918. doi:10.1038/nature06250.

9. Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20: R208–15. doi:10.1016/j.cub.2009.11.055.

10. Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, et al. (2010) A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. Science (80-) 327: 883–886.

11. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, et al. (2013) Identifying recent adaptations in large-scale genomic data. Cell 152: 703–713. Available: http://www.ncbi.nlm.nih.gov/pubmed/23415221. Accessed 27 February 2013.

12. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. Available: http://www.nature.com/nature/journal/v491/n7422/full/nature11632.html. Accessed 2 November 2012.

13. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15: 1576–183. Available: http://www.ncbi.nlm.nih.gov/pubmed/16251467.

14. Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, et al. (2014) 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res 42: D903–9. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965045&tool=pmcentrez&rendertype=abstract. Accessed 8 May 2014.

15. Woolfe A, Goode DK, Cooke J, Callaway H, Smith S, et al. (2007) CONDOR: a database resource of developmentally associated conserved non-coding elements. BMC Dev Biol 7: 100. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2020477&tool=pmcentrez&rendertype=abstract. Accessed 15 July 2014.

16. Lin K, Li H, Schlötterer C, Futschik A (2011) Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics 187: 229–244. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3018323&tool=pmcentrez&rendertype=abstract. Accessed 15 March 2013.

17. Bühlmann P, Hothorn T (2007) Boosting Algorithms: Regularization, Prediction and Model Fitting. Stat Sci 22: 477–505. Available: http://projecteuclid.org/euclid.ss/1207580163. Accessed 15 July 2014.

18. Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289–1303. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1205596&tool=pmcentrez&rendertype=abstract.

19. Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, et al. (2012) Limited evidence for classic selective sweeps in African populations. Genetics 192: 1049–1064. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3522151&tool=pmcentrez&rendertype=abstract.

20. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19: 826–837. Available: http://www.ncbi.nlm.nih.gov/pubmed/19307593.

21. Wilson BA, Petrov D, Messer PW (2014) Soft selective sweeps in complex demographic scenarios Corresponding author : bioRxiV: 0–34.

22. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.

23. Pavlidis P, Jensen JD, Stephan W (2010) Searching for Footprints of Positive Selection in Whole-genome SNP Data from Non-equilibrium Populations. Genetics 110: 908–922. Available: http://www.ncbi.nlm.nih.gov/pubmed/20407129. Accessed 27 July 2010.

24. Sabeti P, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.

25. Wall JD (1999) Recombination and the power of statistical tests of neutrality. Genet Res 74: 65–79.

26. Wall JD (2000) A comparison of estimators of the population recombination rate. Mol Biol Evol 17: 156–163. Available: http://www.ncbi.nlm.nih.gov/pubmed/10666715.

27. Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147: 915–925. Available: http://www.genetics.org/content/147/2/915.short. Accessed 29 November 2012.

28. Nei M (1987) Molecular Evolutionary Genetics. New York, NY: Columbia University Press.

29. Rozas J, Gullaud M, Blandin G, Aguadé M (2001) DNA variation at the rp49 gene region of Drosophila simulans: evolutionary inferences from an unusual haplotype structure. Genetics 158: 1147–1155. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461709&tool=pmcentrez&rendertype=abstract.

30. Kelly JK (1997) A Test of Neutrality Based on Interlocu Associations. Genetics 1206: 1197–1206.

31. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1203831&tool=pmcentrez&rendertype=abstract. Accessed 21 November 2012.

32. Weir BS, Cockerham C (1984) Estimating F-statistics for the analysis of population structure. Evolution (N Y) 38: 1358–1370.

33. Hofer T, Ray N, Wegmann D, Excoffier L (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. Ann Hum Genet 73: 95–108. Available: http://www.ncbi.nlm.nih.gov/pubmed/19040659. Accessed 28 February 2013.

34. Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Res 20: 393–402. Available: http://www.ncbi.nlm.nih.gov/pubmed/20086244.

35. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, et al. (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15: 1566–1575.

36. Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461156&tool=pmcentrez&rendertype=abstract. Accessed 29 November 2012.

37. Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133: 693–709. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1208208&tool=pmcentrez&rendertype=abstract.

38. Ramos-Onsins SE, Rozas J (2002) Statistical properties of new neutrality tests against population growth. Mol Biol Evol 19: 2092–2100. Available: http://www.ncbi.nlm.nih.gov/pubmed/12446801.

Figure Legends

Figure 1. Coalescent simulations.

Simulations were run following a calibrated human demography that resembles population genetic data from three reference continental populations (YRI, CEU and CHB, from left to right) **[Schaffner et al. 2005]**. Nine different time-spanning selective sweep were simulated (grouped as Neutral, Recent, Recent Long and Ancient) allowing for five different final allele frequencies (FaF = 0.2 ,0.4, 0.6, 0.8 and 1.0).

Figure 2. Demography validation.

The difference in relative site frequency spectrum (SFS) of neutral simulations and genome-wide 1000 genomes data is shown. In dashed lines, the original *cosi* demography (no migration) without thinning applied. In solid lines, same demography after singleton thinning. Incomplete matching at the low-frequency region after thinning process is explained by inaccuracies of the ’bestfit’ model. Whereas the slight deviation at the high-frequency region is explained due to the lack of migration in the demographic model used, which would bring back ancestral alleles already fixed in the target population. However, the SFS match is overall adequate to use the neutral simulations as a reference selectively neutral model.

Figure 3. Comparison of positive selection score along the simulated sequence under selective and neutral scenarios.

Summary statistics for three neutrality tests (dDAF, Tajima’s D and XP-EHH, respectively) in simulated EUR population for the Complete scenario (brown), the Incomplete scenario (red), the Partial scenario (orange) and the Neutral scenario (blue) along the simulated sequence length (600 Kbp). Simulated sequence was divided in 25 Kbp regions and a specific summary statistic was applied for each test. The thick line indicates the mean score across the replicates whereas the shape represents two mean standard errors. **(A)** dDAF uses the maximum of individual scores within the 25 Kbp regions; **(B)** Tajima’s D uses the minimum of individual scores within the 25 Kbp regions and **(C)** XP-EHH uses the average of individual scores within the 25 Kbp regions. Population-specific plot for each one of the statistics implemented can be found in **Supporting Figures S2**.

Figure 4. Positive selection test and simulation validation.
Distribution plots, box plots and violin plots for dDAF **(A-B)**, Tajima's D **(C-D)** and XP-EHH **(E-F)** summary scores (**Table 3**) in neutral simulation data, genome-wide 1000 Genomes data and the neutral subset of 1000 genomes data for EUR / CEU population. Plots for the rest of the statistics and populations can be found in **Supporting Figures S3**.

Figure 5. Hierarchical boosting classification tree.
The implemented classification tree was organized in two levels: an unknown genomic region is firstly classified according to the completeness of the sweep, being either Complete, Incomplete or Unclassified. In a second step, it is then classified according to the oldness of the sweep, being Ancient, Recent or Unclassified. The algorithm can be described as following: (1a) if the Complete Boosting score is above the 99th percentile of the distribution of the Complete Boosting scores for the training simulations under Neutral scenario and Partial and Incomplete sweep scenarios, the region is classified as Complete Sweep and go to step 2a, otherwise go to step 1b. (1b) if the Incomplete Boosting score is above the 99th percentile of the distribution of the Incomplete Boosting scores for the training simulations under Neutral scenario and Partial sweep scenarios, the region is classified as Incomplete Sweep and go to step 2b, otherwise go to step 1c. (1c) If not classified at iteration 1a or 1b, the genomic region is unclassified and stop. (2a) If the Ancient/Recent Complete Boosting score is above the 99th percentile of the distribution of the Ancient/ancient Complete Boosting scores for the training simulations under Complete Recent scenario the region is classified as Ancient Complete Sweep, while if it is below the 1th percentile of the distribution of the Ancient/Recent Complete Boosting scores for the training simulations under Complete Ancient scenario the region is classified as Recent Complete Sweep , otherwise the region remains only classified as Complete Sweep. (2b) If the Ancient/Recent Incomplete Boosting score is above the 99th percentile of the distribution of the Ancient/ancient Incomplete Boosting scores for the training simulations under Incomplete Recent scenario the region is classified as Ancient Incomplete Sweep, while if it is below the

1th percentile of the distribution of the Ancient/Recent Incomplete Boosting scores for the training simulations under Incomplete Ancient scenario the region is classified as Recent Incomplete Sweep , otherwise the region remains only classified as Incomplete Sweep.

Figure 6. Standardized coefficients for the three populations and implemented boosting functions.
Estimated coefficients for each population in the four boosting functions used in the classification tree: Complete **(A)**, Incomplete **(B)**, Complete Recent / Ancient **(C)** and Incomplete Recent / Ancient **(D)**. The relevance of the positive selection statistics to classify the different scenarios is given by the strength of its standardized coefficient.

Figure 7. Sensitivity analysis for Complete and Incomplete Boostings and other positive selection statistics.
Upper panel: for each method, ROC curves were separately performed for the different selective scenarios, as defined by the Final Allele Frequency (FaF). For that purpose, we assessed the thresholds corresponding to a given specificity (false positive rate across the *x*-axis) using as a reference the summary statistic distribution observed for the 25 Kpb central regions of the 3000 replicates for the neutral scenario; in turn, the sensitivity (true positive rate across the *y*-axis) was calculated as the proportion of the summary statistics for the 25 Kpb central regions across the 600 replicates of the analyzed selective scenario above this threshold. The line colors appear as described in the lower panel. **Lower panel:** The Area Under the Curve (AUC) score for each for each method and selective scenario (as defined by the FaF). For direct visualization of the performance, we plotted a circle proportional to the AUC besides the AUC value is provided as well.

Figure 8. Classic examples of positive selection as seen by hierarchical boosting.
A UCSC supertrack containing our hierarchical boosting results for each target population was generated: Complete signal is shown in red, Incomplete signal is shown in orange, Recent signal is shown in blue and Ancient signal is shown in vio-

let. The 1% significance thresholds for each boosting were also included. Changing the supertrack overlaying visualization to ”none” splits the hierarchical boosting results into individual tracks. **(A)** Complete Recent sweep signal in the region surrounding *LCT* in CEU; **(B)** Complete sweep signal in the region surrounding *SLC24A5* in CEU; **(C)** Incomplete sweep signal in the region surrounding *EDAR* in CHB and **(D)** Complete Ancient sweep signal in the region surrounding *DARC* in YRI. More information about how to interpret the supertracks can be found in our blog (**http://hsb.upf.edu/)**

Supporting Information

Supporting Text S1. Selection and demographic parameters for *cosi* simulator.

Supporting Figures S2. Population-specific selection signal plots along the simulated sequence length.

Supporting Figures S3. Distribution plots of positive selection tests in simulated and empirical genome-wide population-specific data.

Supporting Text S4. Testing alternative classification tree configurations.

Supporting Text S5. Population-specific correlation matrices for neutral and empirical genome-wide data.

Supporting Text S6. Performance analysis for each population-specific hierarchical boosting.

Supporting Text S7. Estimating the number of selective events detected in 1000 genomes data

Supporting Figures S8. Coefficient convergence for each population-specific hierarchical boosting.

Supporting Table S9. Regions encompassing putative selective events.

Tables

Table 1. Implemented positive selection tests.

Method Family	**Method**	**Reference**	**Analyzed window size**	**Reporting window size**
Allele Frequency Spectrum	Tajima's D	[31]	30 Kbp	3 Kbp
	CLR	[35]	variable size	2 Kbp
	Fay and Wu's H	[36]	30 Kbp	3 Kbp
	Fu and Li's F*	[37]	30 Kbp	3 Kbp
	Fu and Li's D*	[37]	30 Kbp	3 Kbp
	R^2	[38]	30 Kbp	3 Kbp
Linkage Disequilibrium	XP-EHH	modified from [24]	variable size	SNV-based
	diHH	modified from [22]	variable size	SNV-based
	iHS	modified from [22]	variable size	SNV-based
	Omega	[23]	variable size	0.1 Kbp
	EHH Average	modified from [24]	30 Kbp	3 Kbp
	Wall's B	[25]	30 Kbp	3 Kbp
	Wall's Q	[26]	30 Kbp	3 Kbp
	Fu's F	[27]	30 Kbp	3 Kbp
	Dh	[28]	30 Kbp	3 Kbp
	Za	[29]	30 Kbp	3 Kbp
	ZnS	[30]	30 Kbp	3 Kbp
	ZZ	[29]	30 Kbp	3 Kbp
Population Differentiation	Fst (global and pairwise)	[32]	SNV-specific	SNV-based
	dDAF	[33]	SNV-specific	SNV-based
	XP-CLR	[34]	0.1 cM (max. window)	2 Kbp

Table 2. Number of replicates used in each boosting analysis.

Boosting Scheme	**Number of replicates used**
Complete* *vs* Incomplete*+Partial*+Neutral	900 *vs* 6600
Incomplete* *vs* Partial*+Neutral	1800 *vs* 4800
Complete Ancient *vs* Complete Recent	300 *vs* 300
Incomplete Ancient *vs* Incomplete Recent	600 *vs* 600

* Recent, Recent Long and Ancient

Table 3. Selection of the best region summarizing approach for each positive selection test.

Statistic	**Mean**	**Min / Max**	**Best Approach**
dDAF	0.5945	0.9893 (Max)	MAX
Fst	0.9052	0.9901 (Max)	MAX
XP-CLR	0.7009	0.6990 (Max)	MEAN
diHH	0.8220	0.6562 (Max)	MEAN
iHS	0.9673	0.8130 (Max)	MEAN
XP-EHH	0.9947	0.9839 (Max)	MEAN
EHH Average	0.9253	0.8495 (Max)	MEAN
Omega	0.7882	0.7983 (Max)	MAX
CLR	0.7457	0.7327 (Max)	MEAN
Tajima's D	0.9684	0.9735 (Min)	MIN
Fu & Li's D*	0.8793	0.8828 (Min)	MIN
Fu & Li's F*	0.9541	0.9504 (Min)	MEAN
Fay & Wu's H	0.8042	0.7811 (Min)	MEAN
R^2	0.9689	0.9743 (Min)	MIN
Fu's F	0.7734	0.7549 (Min)	MEAN
Dh	0.9949	0.9948 (Min)	MEAN
Wall's B	0.4879	0.5651 (Max)	MAX
Wall's Q	0.4594	0.5354 (Max)	MAX
Za	0.4504	0.5301 (Max)	MAX
ZnS	0.4401	0.5215 (Max)	MAX
ZZ	0.4516	0.5147 (Max)	MAX

Positive selection test scores were unified for a fixed region size in order to be used in the boosting analysis. Region size was set to 25 Kbp. Mean, maximum and minimum summary statistics were evaluated in EUR simulations.

Table 4. Quality Control analysis.

Statistic	Distribution Analysis	Correlation among Statistics	Statistic Properties
dDAF	----	Fst	cross-population directional
Fst	----	**dDAF**	**cross-population non-directional**
XP-CLR	----	----	cross-population directional
diHH	----	----	fails on complete sweeps
iHS	----	----	fails on complete sweeps
XP-EHH	----	----	cross-population directional
EHH Average	----	----	----
Omega	----	----	only sensitivity at FaF=1.0
CLR	----	----	artifact at low SNV density [A]
Tajima's D	----	R^2 / Fu & Li's F*	----
Fu & Li's D*	----	Fu & Li's F*	----
Fu & Li's F*	----	**Fu & Li's D* / Tajima's D / R^2**	----
Fay & Wu's H	----	----	----
R^2	----	**Fu & Li's F* / Tajima's D**	----
Fu's F	**inconsistent**	----	----
Dh	----	----	**not a positive selection test**
Wall's B	----	Wall's Q / Za / ZnS	**low sensitivity**
Wall's Q	----	Wall's B / Za / ZnS	**low sensitivity**
Za	----	Wall's Q / Wall's B / ZnS	**low sensitivity**
ZnS	----	Wall's Q / Za / Wall's B	**low sensitivity**
ZZ	----	----	**low sensitivity**

Criteria used to ascertain suitability of a positive selection test to be used in our boosting analysis in combination with the other tests. Removed statistics and main reasons are marked in **bold**

[A] masked in low-density regions in empirical data

Table 5. Average hierarchical boosting classification power for the three populations in True Positives evaluation scenarios.

Scenarios to Classify		**Hierarchical Boosting Classification**				
Class	**Type**	**Correct Class**	**Correct Class & Type**	**Wrong Class**	**Wrong Type**	**Partial/Neutral**
Complete	Recent	87.59%	25.41%	1.83%	1.81%	10.58%
Complete	Recent Long	91.10%	---------	2.67%	----------	6.23%
Complete	Ancient	90.05%	23.76%	1.78%	0.44%	8.17%
Incomplete	Recent	55.67%	18.89%	10.45%	1.50%	33.89%
Incomplete	Recent Long	50.67%	---------	4.44%	----------	44.89%
Incomplete	Ancient	23.89%	11.00%	11.05%	0.00%	65.06%
	average	66.49%	19.77%	5.37%	1.18%	28.14%

Population-specific True Positive tables can be found in **Supporting Text S6**.

Table 6. Average hierarchical boosting classification power for the three populations in True Negative evaluation scenarios.

Scenarios to Classify		**Hierarchical Boosting Classification**	
Class	**Type**	**Wrong Class**	**Partial/Neutral**
Neutral	Neutral	0.20%	99.80%
Partial	Recent	0.95%	99.05%
Partial	Ancient	1.17%	98.83%

Population-specific True Negative tables can be found in **Supporting Text S6**.

Table 7. Regions of 25 Kbp in genome-wide 1000 genomes data classified as being under selection.

Classified as	**CEU**	**CHB**	**YRI**
Complete	263	427	18
- Complete Ancient	1	18	6
- Complete Recent	77	119	0
Incomplete	394	451	16
- Incomplete Ancient	27	69	5
- Incomplete Recent	29	81	9
number of 25 Kbp regions analyzed	103215	103617	103496

Due to the hierarchical nature of the method, a region classified as Complete cannot be classified as Incomplete at the same time. Within each of these two classes, regions can be classified as Recent, Ancient or not classified.

Table 8. Classification of the putative selective events detected.

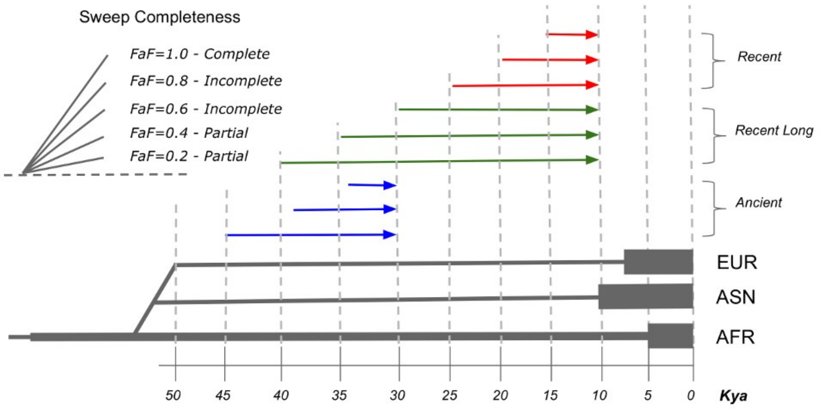
Population	Class	Ancient	Recent	Ancient & Recent	Undefined	Total	Proportion
YRI	Complete & Incomplete	0	0	0	0	0	0.000
	Complete	4	0	0	10	14	0.519
	Incomplete	4	7	0	2	13	0.481
	Total	8	7	0	12	27	1.000
	Proportion	0.296	0.259	0.000	0.444	1.000	1.000
CEU	Complete & Incomplete	3	9	2	13	27	0.076
	Complete	1	42	0	76	119	0.335
	Incomplete	21	23	0	165	209	0.589
	Total	25	74	2	254	355	1.000
	Proportion	0.070	0.208	0.006	0.715	1.000	1.000
CHB	Complete & Incomplete	10	25	6	14	55	0.130
	Complete	8	76	1	98	183	0.432
	Incomplete	38	44	3	101	186	0.439
	Total	56	145	10	213	424	1.000
	Proportion	0.132	0.342	0.024	0.502	1.000	1.000
Global[a]	Complete & Incomplete	13	34	8	27	82	0.0102
	Complete	13	118	1	184	316	0.392
	Incomplete	63	74	3	268	408	0.506
	Total	89	226	12	479	806	1.000
	Proportion	0.110	0.280	0.015	0.594	1.000	1.000

A putative selective event (as defined in **Supporting Text S7**) is assigned to a given class , i.e. Complete or Incomplete, when the genomic region encompassing the signal contains 25 Kbp windows with a significant score only for Complete or Incomplete boosting, respectively. If significant scores are observed for both Complete and Incomplete Boostings, the selective event is defined as both Complete & Incomplete. Similarly, the type of the selective event (Ancient or Recent) is assigned when the genomic region encompassing it only contains 25 Kbp windows with a significant score for the related boosting, i.e. Ancient/Recent Complete, Ancient/Recent Incomplete boosting or one of the two, if the selective event was classified as Complete, Incomplete or both, respectively.

[a] Global refers to the analysis overall the three populations.

FIGURES

Figure 1.



		TYPES		
CLASSES	**Final Allele Frequency**	**Recent** (25 to 10 Kya)	**Recent Long** (40 to 10 Kya)	**Ancient** (45 to 30 Kya)
Complete Sweep	1.0	300 replicates	300 replicates	300 replicates
Incomplete Sweep	0.8 0.6	300 replicates 300 replicates	300 replicates 300 replicates	300 replicates 300 replicates
Partial Sweep	0.4 0.2	300 replicates 300 replicates	300 replicates 300 replicates	300 replicates 300 replicates
Neutral		3000 replicates		

* This simulation scheme was duplicated for evaluation purposes.
* Selection simulation were run for each population demography independently

Figure 2.

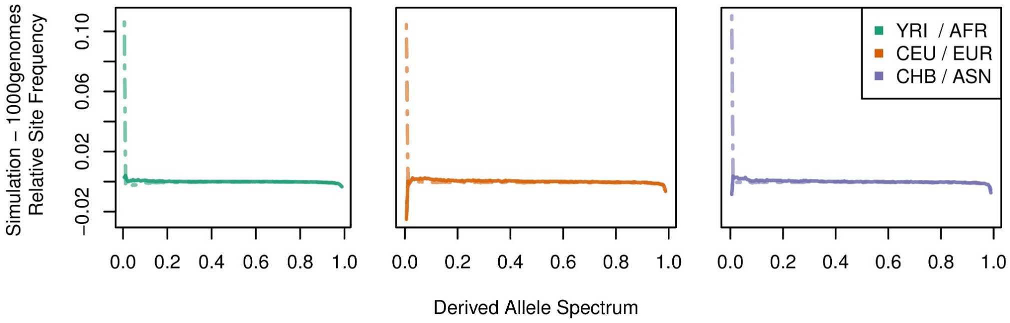

Simulation – 1000genomes
Relative Site Frequency
−0.02
0.02
0.06
0.10
0.0
0.2
0.4
0.6
0.8
1.0
Derived Allele Spectrum
■ YRI / AFR
■ CEU / EUR
■ CHB / ASN

Figure 3.

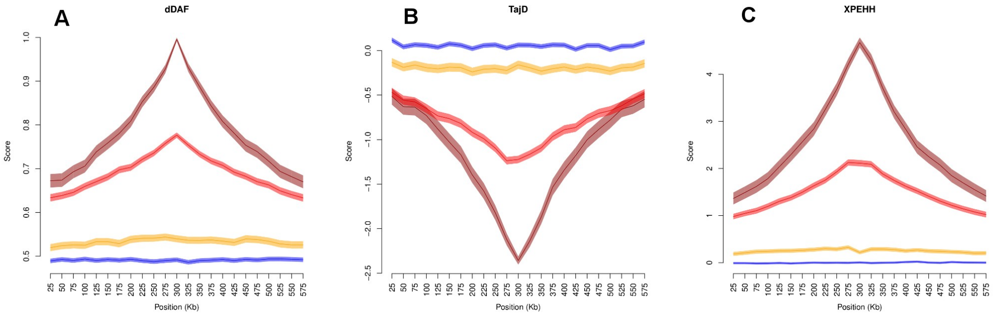

A
dDAF
Score
1.0
0.9
0.8
0.7
0.6
0.5
25 50 75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550 575
Position (Kb)
B
TajD
Score
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
25 50 75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550 575
Position (Kb)
C
XPEHH
Score
4
3
2
1
0
25 50 75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550 575
Position (Kb)

Figure 4.

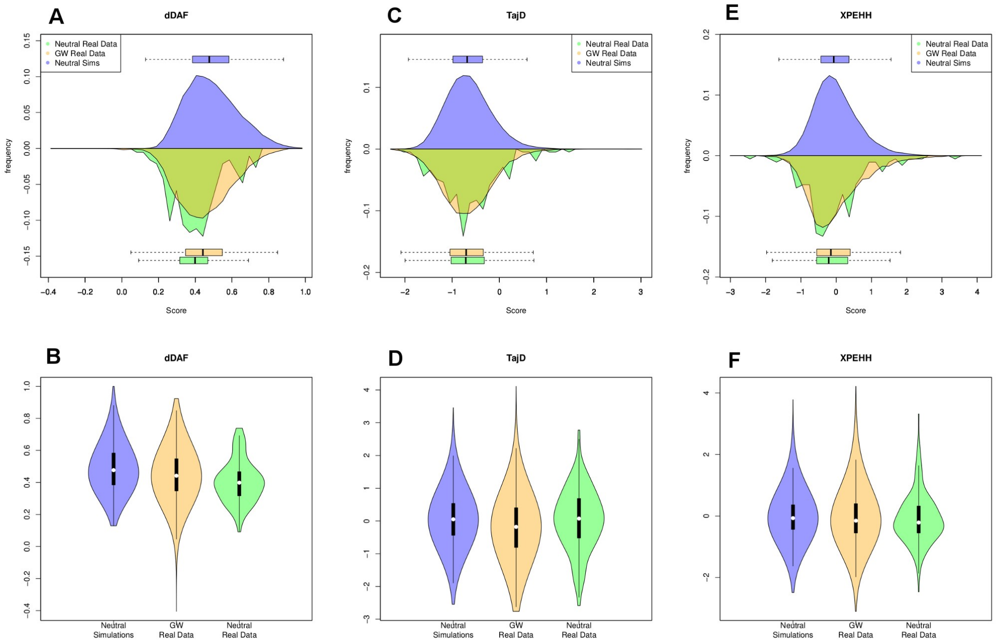
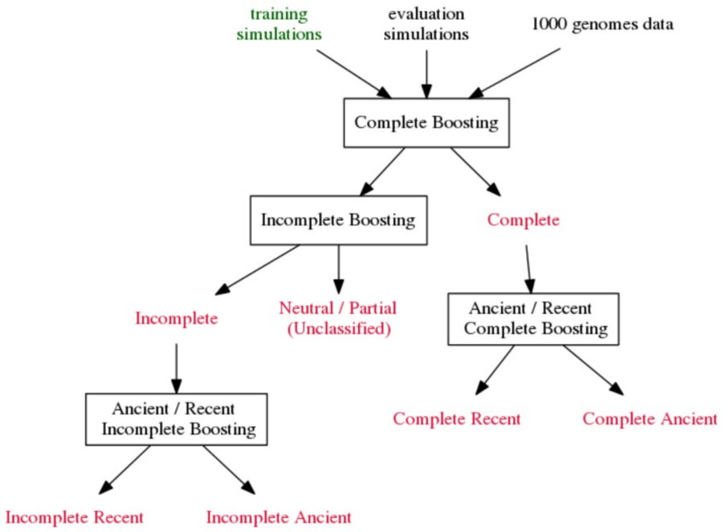

A
dDAF
Neutral Real Data
GW Real Data
Neutral Sims
frequency
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Score
C
TajD
Neutral Real Data
GW Real Data
Neutral Sims
frequency
0.1
0.0
−0.1
−0.2
−2
−1
0
1
2
3
Score
E
XPEHH
Neutral Real Data
GW Real Data
Neutral Sims
frequency
0.2
0.1
0.0
−0.1
−0.2
−3
−2
−1
0
1
2
3
4
Score
B
dDAF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Neutral Simulations
GW Real Data
Neutral Real Data
D
TajD
4
3
2
1
0
−1
−2
−3
Neutral Simulations
GW Real Data
Neutral Real Data
F
XPEHH
4
2
0
−2
Neutral Simulations
GW Real Data
Neutral Real Data

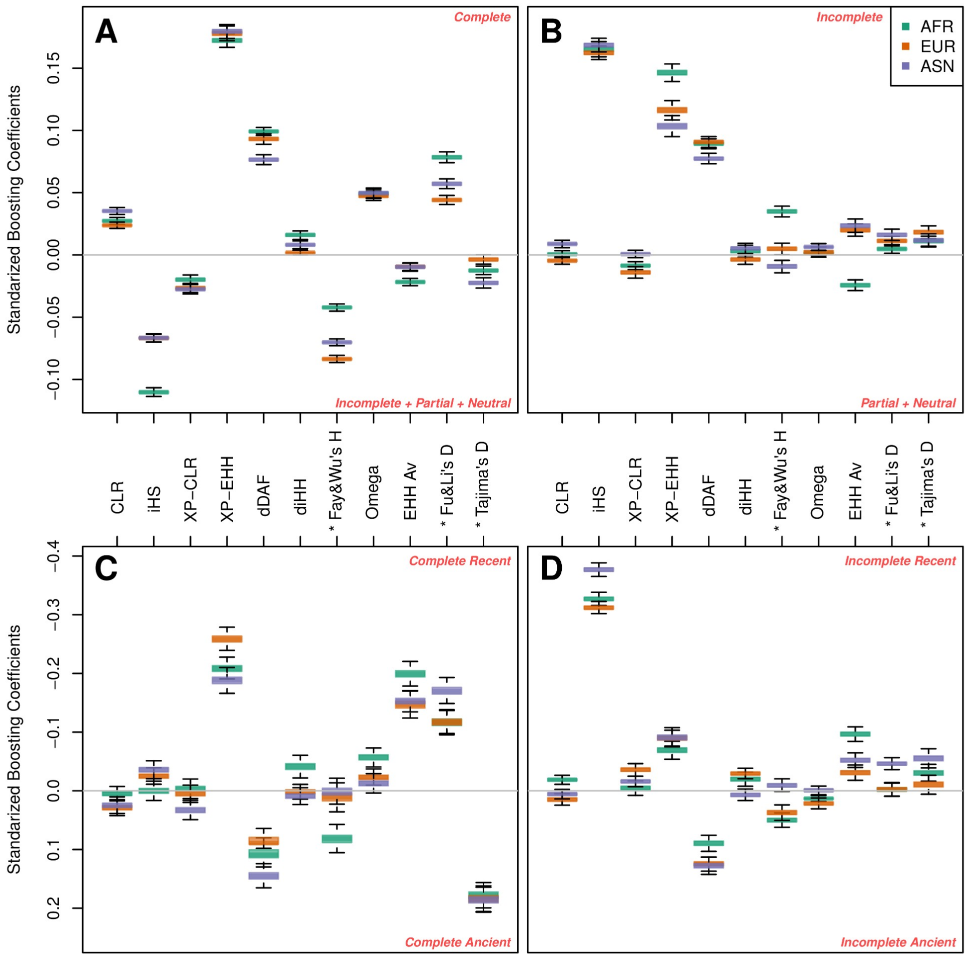
Figure 5.



	Competing A		Competing B
Complete Boosting	Complete*	*vs*	Incomplete* Partial* Neutral
Incomplete Boosting	Incomplete*	*vs*	Partial* Neutral
Ancient/Recent Complete Boosting	Complete Ancient	*vs*	Complete Recent
Ancient/Recent Incomplete Boosting	Incomplete Ancient	*vs*	Incomplete Recent

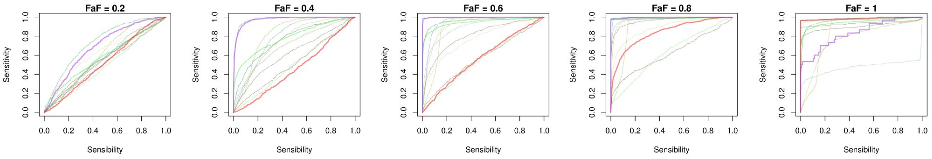
**Recent, Recent Long and Ancient*

Figure 6.



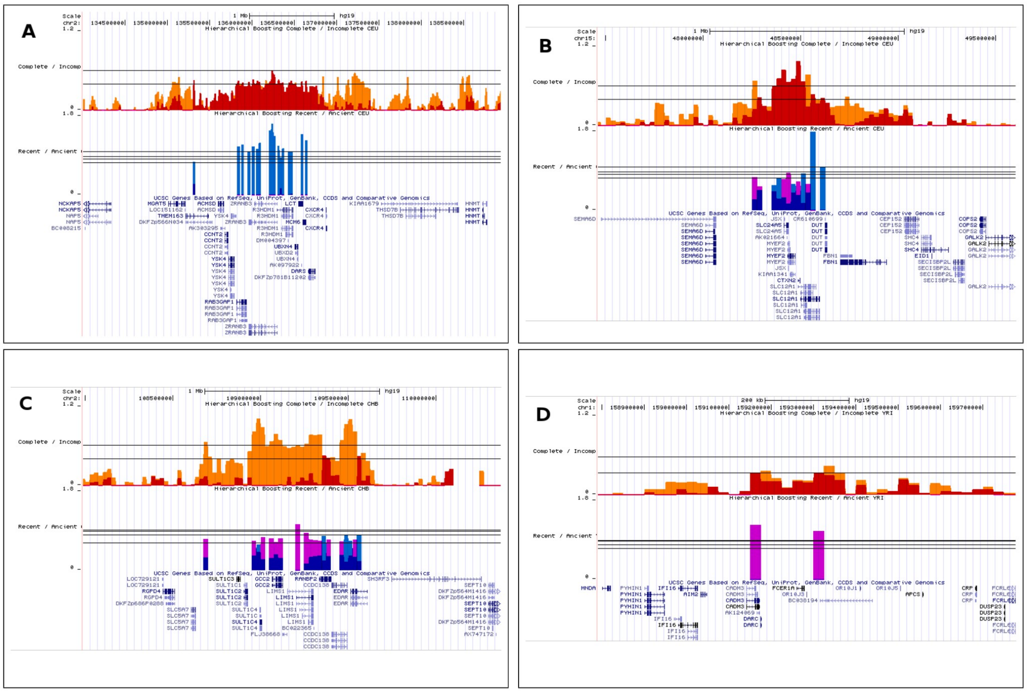
* Lower Tail based Statistic: boosting coefficient is multiplied by −1 to facilitate interpretation.

Figure 7.



FAF	dDAF	XP−CLR	diHH	iHS	XP−EHH	EhhAv	Omega	CLR	Tajima's D	Fu & Li's D	Fay & Wu's H	Complete Boosting	Incomplete Boosting
0.2	0.561 ●	0.552 ●	0.609 ●	0.734 ●	0.551 ●	0.583 ●	0.486 ●	0.511 ●	0.557 ●	0.635 ●	0.542 ●	0.52 ●	0.689 ●
0.4	0.788 ●	0.7 ●	0.82 ●	0.979 ●	0.764 ●	0.785 ●	0.501 ●	0.513 ●	0.617 ●	0.704 ●	0.748 ●	0.456 ●	0.975 ●
0.6	0.947 ●	0.893 ●	0.89 ●	0.997 ●	0.963 ●	0.945 ●	0.556 ●	0.568 ●	0.807 ●	0.768 ●	0.92 ●	0.569 ●	0.998 ●
0.8	0.986 ●	0.977 ●	0.909 ●	0.998 ●	0.988 ●	0.985 ●	0.624 ●	0.621 ●	0.963 ●	0.884 ●	0.971 ●	0.839 ●	0.997 ●
1	0.988 ●	0.472 ●	0.756 ●	0.965 ●	0.978 ●	0.946 ●	0.986 ●	0.901 ●	0.934 ●	0.983 ●	0.853 ●	0.985 ●	0.835 ●

Figure 8.


A
Hierarchical Boosting Complete / Incomplete CEU
Complete / Incomp
Hierarchical Boosting Recent / Ancient CEU
Recent / Ancient
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
B
Hierarchical Boosting Complete / Incomplete CEU
Complete / Incomp
Hierarchical Boosting Recent / Ancient CEU
Recent / Ancient
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
C
Hierarchical Boosting Complete / Incomplete CHB
Complete / Incomp
Hierarchical Boosting Recent / Ancient CHB
Recent / Ancient
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
D
Hierarchical Boosting Complete / Incomplete YRI
Complete / Incomp
Hierarchical Boosting Recent / Ancient YRI
Recent / Ancient
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics

Bibliography

[1] Lalueza-Fox C, Gilbert MTP. Paleogenomics of archaic hominins. Current biology : CB. 2011 Dec;21(24):R1002–9.

[2] McDougall I, Brown FH, Fleagle JG. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature. 2005;433(February):733–736.

[3] Veeramah KR, Hammer MF. The impact of whole-genome sequencing on the reconstruction of human population history. Nature Reviews Genetics. 2014 Feb;15(3):149–162.

[4] Stringer C. What makes a modern human. Nature. 2012;485(7396):33–35.

[5] Nei M, Roychoudhury AK. Evolutionary relationships of human populations. Molecular biology and evolution. 1993;10(5):927–943.

[6] von Cramon-Taubadel N, Lycett SJ. Brief communication: human cranial variation fits iterative founder effect model with African origin. American journal of physical anthropology. 2008 May;136(1):108–13.

[7] Liu H, Prugnolle F, Manica A. A Geographically Explicit Genetic Model of Worldwide Human-Settlement History. American journal of human genetics. 2006;79(August):230–237.

[8] Manica A, Amos W, Balloux F, Hanihara T. The effect of ancient population bottlenecks on human phenotypic variation. Nature. 2007 Jul;448(7151):346–8.

[9] Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science (New York, NY). 2008 Mar;319(5866):1100–4.

[10] Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome research. 2005 Nov;15(11):1576–183.

[11] Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS genetics. 2013 Oct;9(10):e1003905.

[12] Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics (Oxford, England). 2010 Aug;26(16):2064–5.

[13] Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics. 2000 Sep;156(1):297–304.

[14] Conrad DF, Keebler JEM, DePristo Ma, Lindsay SJ, Zhang Y, Casals F, et al. Variation in genome-wide mutation rates within and between human families. Nature genetics. 2011 Jul;43(7):712–4.

[15] Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. American journal of human genetics. 2003 Nov;73(5):1162–9.

[16] Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg Na, et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nature genetics. 2006 Nov;38(11):1251–60.

[17] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb;409(6822):860–921.

[18] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001 Feb;291(5507):1304–51.

[19] Cann H, de Toma C, Cazes L, Legrand M, Morel V, Piouffre L, et al. A human genome diversity cell line panel. Science. 2002;12(296(5566)):261–262.

[20] The International Hapmap Consortium. A haplotype map of the human genome. Nature. 2005 Oct;437(7063):1299–320.

[21] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.

[22] Kimura M. The Neutral Theory of molecular evolution. Cambridge: Cambridge University Press; 1983.

[23] Kimura M. Evolutionary Rate at the Molecular Level. Nature. 1968;217:624–626.

[24] Kreitman M. The neutral theory is dead. Long live the neutral theory. BioEssays : news and reviews in molecular, cellular and developmental biology. 1996 Aug;18(8):678–83.

[25] Fay JC, Wu CI. Sequence divergence, functional constraint, and selection in protein evolution. Annual review of genomics and human genetics. 2003 Jan;4:213–35.

[26] Hardy GH. Mendelian Proportions in a Mixed Population. Science. 1908;28(706):49–50.

[27] Weinberg W. Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg. 1908;64:368–382.

[28] Smith NGC, Webster MT, Ellegren H. Deterministic mutation rate variation in the human genome. Genome research. 2002 Sep;12(9):1350–6.

[29] Darwin CR, Wallace AR. On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. Journal of the Proceedings of the Linnean Society of London. 1858;3(9):46–50.

[30] Darwin CR. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London: John Murray; 1859.

[31] Quintana-murci L, Clark AG. Population genetic tools for dissecting innate immunity in humans. Nature Reviews Immunology. 2013;13(4):280–293.

[32] Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Philosophical Transactions of the Royal Society of Edinburgh. 1918;52:399–433.

[33] Fisher RA. The Genetical Theory of Natural Selection. Oxford: Oxford Univ Press; 1930.

[34] Haldane JBS. A Mathematical Theory of Natural and Artificial Selection. Part I. Transactions of the Cambridge Philosophical Society. 1924;23:19–41.

[35] Wright S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: Proceedings of the sixth international congress of genetic. 1932;p. 356–366.

[36] Wright S. Evolution in Mendelian populations. Genetics. 1931;16(2):97–159.

[37] Duret L. Neutral Theory : The Null Hypothesis of Molecular Evolution. Nature Education. 2008;1(1):1–6.

[38] Kimura M. The neutral theory of molecular evolution: a review of recent evidence. Japanese Journal of Genetics. 1991;66:367–386.

[39] Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. Science. 2006;312(5780):1614–1620.

[40] Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution. 1986 Sep;3(5):418–426.

[41] Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN, editor. Mammalian Protein Metabolism. vol. III. New York: Academic Press; 1969. p. 21–132.

[42] Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution. 1994 Sep;11(5):725–736.

[43] Seo TK, Kishino H, Thorne JL. Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. Molecular Biology and Evolution. 2004 Jul;21(7):1201–1213.

[44] Suzuki Y, Gojobori T. A method for detecting positive selection at single amino acid sites. Molecular Biology and Evolution. 1999 Oct;16(10):1315–1328.

[45] Suzuki Y. New methods for detecting positive selection at single amino acid sites. Journal of Molecular Evolution. 2004 Jul;59(1):11–19.

[46] Massingham T, Goldman N. Detecting amino acid sites under positive selection and purifying selection. Genetics. 2005 Mar;169(3):1753–1762.

[47] Kosakovsky Pond SL, Frost SDW. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Molecular Biology and Evolution. 2005 May;22(5):1208–1222.

[48] Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 1998 Mar;148(3):929–936.

[49] Yang Z, Nielsen R, Goldman N, Krabbe Pedersen AM. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 2000;155:431–449.

[50] Yang Z, Wong WSW, Nielsen R. Bayes empirical Bayes inference of amino acid sites under positive selection. Molecular Biology and Evolution. 2005;22(4):1107–1118.

[51] Anisimova M, Bielawski JP, Yang Z. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Molecular Biology and Evolution. 2002;18(6):1585–1592.

[52] Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution. 2002 Jun;19(6):908–917.

[53] Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Molecular Biology and Evolution. 2005 Dec;22(12):2472–2479.

[54] Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genetics. 2012 Jan;8(7):e1002764.

[55] Yang Z, dos Reis M. Statistical properties of the branch-site test of positive selection. Molecular Biology and Evolution. 2011 Mar;28(3):1217–1228.

[56] Maynard-Smith J, John Haigh. The hitch-hiking effect of a favourable gene. Genetical Research. 1974;23(1):23–35.

[57] Kaplan NL, Hudsont RR, Langle CH. The "Hitchhiking Effect" Revisited. Genetics. 1989;899:887–899.

[58] Barton N. The geometry of adaptation. Nature. 1998;395(6704):751–2.

[59] Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000 Jul;155(3):1405–13.

[60] Przeworski M. The Signature of Positive Selection at Randomly Chosen Loci. Genetics. 2002;1189:1179–1189.

[61] Kim Y. Allele Frequency Distribution Under Recurrent Selective Sweeps. Genetics. 2006;1978:1967–1978.

[62] Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 1995 Jun;140(2):783–96.

[63] Gillespie JH. Genetic Drift in an Infinite Population : The Pseudo-hitchhiking Model. Genetics. 2000;155:909–919.

[64] Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño Torres A, Pavlidis P, et al. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic acids research. 2014 Jan;42(Database issue):D903–9.

[65] Sabeti P, Usen S, Farhadian S, Jallow M, Doherty T, Newport M, et al. CD40L association with protection from severe malaria. Genes and immunity. 2002 Aug;3(5):286–91.

[66] Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989 Nov;123(3):585–95.

[67] Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Research. 2005;15:1566–1575.

[68] Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993 Mar;133(3):693–709.

[69] Ramos-Onsins SE, Rozas J. Statistical properties of new neutrality tests against population growth. Molecular biology and evolution. 2002 Dec;19(12):2092–100.

[70] Sabeti P, C, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449(7164):913–918.

[71] Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS biology. 2006;4(3):e72.

[72] Wall JD. Recombination and the power of statistical tests of neutrality. Genetical Research. 1999;74:65–79.

[73] Wall JD. A comparison of estimators of the population recombination rate. Molecular biology and evolution. 2000 Jan;17(1):156–63.

[74] Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics. 1997 Oct;147(2):915–25.

[75] Nei M. Molecular Evolutionary Genetics. New York, NY: Columbia University Press; 1987.

[76] Rozas J, Gullaud M, Blandin G, Aguadé M. DNA variation at the rp49 gene region of Drosophila simulans: evolutionary inferences from an unusual haplotype structure. Genetics. 2001 Jul;158(3):1147–55.

[77] Kelly JK. A Test of Neutrality Based on Interlocu Associations. Genetics. 1997;1206:1197–1206.

[78] Weir BS, Cockerham C. Estimating F-statistics for the analysis of population structure. Evolution. 1984;38:1358–1370.

[79] Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome Research. 2010 Mar;20(3):393–402.

[80] Hofer T, Ray N, Wegmann D, Excoffier L. Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. Annals of human genetics. 2009 Jan;73(1):95–108.

[81] Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences of the United States of America. 1979 Oct;76(10):5269–73.

[82] Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting FST. Nature Reviews Genetics. 2009;10(9):639–650.

[83] Teo YY, Small KS, Kwiatkowski DP. Methodological challenges of genome-wide association analysis in Africa. Nature reviews Genetics. 2010 Feb;11(2):149–60.

[84] Wall JD, Cox MP, Mendez FL, Woerner A, Severson T, Hammer MF. A novel DNA sequence database for analyzing human demographic history. Genome research. 2008 Aug;18(8):1354–61.

[85] Casals F, Bertranpetit J. Human genetic variation, shared and private. Science (New York, NY). 2012 Jul;337(6090):39–40.

[86] Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012 Nov;192(3):1065–93.

[87] Thornton KR, Jensen JD. Controlling the false-positive rate in multilocus genome scans for selection. Genetics. 2007;175(2):737–750.

[88] Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW. Limited evidence for classic selective sweeps in African populations. Genetics. 2012 Nov;192(3):1049–64.

[89] Alves I, Srámková Hanulová A, Foll M, Excoffier L. Genomic data reveal a complex making of humans. PLoS genetics. 2012 Jan;8(7):e1002837.

[90] Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993 Aug;134(4):1289–303.

[91] Enard D, Messer PW, Petrov Da. Genome-wide signals of positive selection in human evolution. Genome research. 2014 Mar;24(6):885–895.

[92] Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, Laval G. Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. Molecular biology and evolution. 2014 Jul;31(7):1850–68.

[93] Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science (New York, NY). 2012 May;336(6082):740–3.

[94] McEvedy C. The bubonic plague. Scientific American. 1988;258(2):117–23.

[95] Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, et al. Linkage disequilibrium in the human genome. Nature. 2001 May;411(6834):199–204.

[96] Hallatschek O, Nelson DR. Gene surfing in expanding populations. Theoretical population biology. 2008 Feb;73(1):158–70.

[97] Hudson RR. Oxford Surveys in Evolutionary Biology. Oxford University Press; 1991.

[98] Kingman JFC. The coalescent. Stochastic Processes and Their Applications. 1982;13:235–248.

[99] Wakeley J. Coalescent Theory: An Introduction. Greenwood Village, Colorado: Roberts & Company Publishers; 2008.

[100] Rosenberg Na, Nordborg M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature reviews Genetics. 2002 May;3(5):380–90.

[101] Marjoram P, Wall JD. Fast "coalescent" simulation. BMC genetics. 2006 Jan;7:16.

[102] Mailund T, Schierup MH, Pedersen CNS, Mechlenborg PJM, Madsen JN, Schauser L. CoaSim: a flexible environment for simulating genetic data under coalescent models. BMC bioinformatics. 2005 Jan;6:252.

[103] Spencer CCa, Coop G. SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics (Oxford, England). 2004 Dec;20(18):3673–5.

[104] Hudson RR. Bioinformatics applications note. Bioinformatics. 2002;18(2):337–338.

[105] Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences of the United States of America. 2011 Jul;108(29):11983–8.

[106] Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, et al. A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. Science. 2010 Jan;327(5967):883–886.

[107] Hernandez RD. A flexible forward simulator for populations subject to selection and demography. Bioinformatics (Oxford, England). 2008 Dec;24(23):2786–7.

[108] Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS genetics. 2009 Oct;5(10):e1000695.

[109] Uricchio LH, Hernandez RD. Robust Forward Simulations of Recurrent Hitchhiking. Genetics. 2014 Feb;p. 1–33.

[110] Nielsen R, Hubisz MJ, Clark AG. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics. 2004;168:2373–2382.

[111] Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, Nürnberg P, et al. Demographic history of Oceania inferred from genome-wide data. Current biology : CB. 2010 Nov;20(22):1983–92.

[112] Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, et al. The genetic prehistory of southern Africa. Nature communications. 2012 Jan;3:1143.

[113] Akey JM. Constructing genomic maps of positive selection in humans: where do we go from here? Genome research. 2009 May;19(5):711–22.

[114] Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome research. 2006 Aug;16(8):980–9.

[115] Teshima KM, Coop G, Przeworski M. How reliable are empirical genomic scans for selective sweeps? Genome research. 2006;16(6):702–712.

[116] Zeng K, Fu Yx, Shi S, Wu Ci. Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants. Genetics. 2006;174:1431–1439.

[117] Zeng K, Shi S, Wu CI. Compound tests for the detection of hitchhiking under positive selection. Molecular biology and evolution. 2007 Aug;24(8):1898–908.

[118] Watterson Ga. The homozygosity test of neutrality. Genetics. 1978 Feb;88(2):405–17.

[119] Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, et al. Identifying recent adaptations in large-scale genomic data. Cell. 2013 Feb;152(4):703–13.

[120] Lin K, Li H, Schlötterer C, Futschik A. Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics. 2011 Jan;187(1):229–44.

[121] Schapire RE. The strength of weak learnability. Machine Learning. 1990 Jun;5:197–227.

[122] Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current biology : CB. 2010 Feb;20(4):R208–15.

[123] Hamblin MT, Di Rienzo A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. American journal of human genetics. 2000 May;66(5):1669–79.

[124] Tishkoff Sa, Reed Fa, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nature genetics. 2007;39(1):31–40.

[125] Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Dickson M, Grimwood J, et al. Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science (New York, NY). 2005 Mar;307(5717):1928–33.

[126] Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, Carroll SB. The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. Cell. 2008 Mar;132(5):783–93.

[127] Scheinfeldt LB, Biswas S, Madeoy J, Connelly CF, Schadt EE, Akey JM. Population genomic analysis of ALMS1 in humans reveals a surprisingly complex evolutionary history. Molecular biology and evolution. 2009 Jun;26(6):1357–67.

[128] Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 2005 Apr;169(4):2335–52.

[129] Innan H, Kim Y. Pattern of polymorphism after strong artificial selection in a domestication event. Proceedings of the National Academy of Sciences of the United States of America. 2004 Jul;101(29):10667–72.

[130] Orr HA, Betancourt AJ. Haldane ' s Sieve and Adaptation From the Standing Genetic Variation. Genetics. 2001;157:875–884.

[131] Pennings PS, Hermisson J. Soft sweeps II–molecular population genetics of adaptation from recurrent mutation or migration. Molecular biology and evolution. 2006 May;23(5):1076–84.

[132] Przeworski M, Coop G, Wall JD. The signature of positive selection on standing genetic variation. Evolution; international journal of organic evolution. 2005 Nov;59(11):2312–23.

[133] Fu W, O'Connor TD, Akey JM. Genetic architecture of quantitative traits and complex diseases. Current opinion in genetics & development. 2013 Dec;23(6):678–83.

[134] Scheinfeldt LB, Tishkoff Sa. Recent human adaptation: genomic approaches, interpretation and insights. Nature reviews Genetics. 2013 Oct;14(10):692–702.

[135] Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On detecting incomplete soft or hard selective sweeps using haplotype structure. Molecular biology and evolution. 2014 May;31(5):1275–91.

[136] Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in Drosophila were abundant and primarily soft. arXiv. 2014;.

[137] Coop G, Witonsky D, Di Rienzo A, Pritchard JK. Using Environmental Correlations to Identify Loci Underlying Local Adaptation. Genetics. 2010 Jun;1423(August):1411–1423.

[138] Günther T, Coop G. Robust identification of local adaptation from allele frequencies. 2Genetics. 2013;195(1):205–20.

[139] Ralph PL, Coop G. Parallel Adaptation: One or Many Waves of Advance of an Advantageous Allele? Genetics. 2010 Jul;668(October):647–668.

[140] Hancock AM, Alkorta-Aranburu G, Witonsky DB, Di Rienzo A. Adaptations to new environments in humans: the role of subtle allele frequency shifts. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2010 Aug;365(1552):2459–68.

[141] Mendizabal I, Marigorta UM, Lao O, Comas D. Adaptive evolution of loci covarying with the human African Pygmy phenotype. Human genetics. 2012 Aug;131(8):1305–17.

[142] Orr HA. Testing Natural Selection vs. Genetic Drift in Phenotypic Evolution Using Quantitative Trait Locus Data. Genetics. 1998;149:2099–2104.

[143] Turchin M, Chiang CWK, Palmer CD, Sankararaman, Sriram Reich D, consortium G, Hirschhorn JN. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nature Genetics. 2012;44(9):1015–1019.

[144] Berg JJ, Coop G. The Population Genetic Signature of Polygenic Local Adaptation. arXiv. 2013;.

[145] Leinonen T, McCairns RJS, O'Hara RB, Merilä J. Q(ST)-F(ST) comparisons: evolutionary and ecological insights from genomic heterogeneity. Nature reviews Genetics. 2013 Mar;14(3):179–90.

[146] Barrett RDH, Hoekstra HE. Molecular spandrels: tests of adaptation at the genetic level. Nature reviews Genetics. 2011 Nov;12(11):767–80.

[147] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis Ca, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep;489(7414):57–74.

[148] Tishkoff Sa, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, et al. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science. 2001 Jul;293(5529):455–62.

[149] Ayodo G, Price AL, Keinan A, Ajwang A, Otieno MF, Orago ASS, et al. Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. American journal of human genetics. 2007 Aug;81(2):234–242.

[150] Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, et al. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. American journal of human genetics. 2004;74:1111–1120.

[151] Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, Kim Y, et al. Spread of an Inactive Form of Caspase-12 in Humans Is Due to Recent Positive Selection. American journal of human genetics. 2006;78:659–670.

[152] Carnero-Montoro E, Bonet L, Engelken J, Bielig T, Martínez-Florensa M, Lozano F, et al. Evolutionary and functional evidence

for positive selection at the human CD5 immune receptor gene. Molecular biology and evolution. 2012 Feb;29(2):811–23.

[153] Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a High-Density SNP Map for Signatures of Natural Selection. Genome research. 2002;12:1805–1814.

[154] Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome research. 2009;19(5):826–837.

[155] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct;467(7319):1061–1073.

[156] Hindorff La, Sethupathy P, Junkins Ha, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009 Jun;106(23):9362–7.

[157] Xue Y, Zhang X, Huang N, Daly A, Gillson CJ, Macarthur DG, et al. Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. Genetics. 2009 Nov;183(3):1065–77.

[158] Sabeti PC, Walsh E, Schaffner SF, Varilly P, Fry B, Hutcheson HB, et al. The case for selection at CCR5-Delta32. PLoS biology. 2005;3(11):e378.

[159] Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, et al. Genetic evidence for high-altitude adaptation in Tibet. Science. 2010 Jul;329(5987):72–5.

[160] Jeong C, Alkorta-Aranburu G, Basnyat B, Neupane M, Witonsky DB, Pritchard JK, et al. Admixture facilitates genetic adaptations to high altitude in Tibet. Nature communications. 2014 Jan;5:3281.

[161] Beall CM, Cavalleri GL, Deng L, Elston RC, Gao Y, Knight J, et al. Natural selection on EPAS1 (HIF2 α) associated with low hemoglobin concentration in Tibetan highlanders. Proc Natl Acad Sci U S A. 2010;107(25):11459–11464.

[162] Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, et al. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proceedings of the National Academy of Sciences of the United States of America. 2014 Apr;111(13):4832–7.

[163] Lamason RL, Mohideen MAPK, Mest JR, Wong AC, Norton HL, Aros MC, et al. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science (New York, NY). 2005 Dec;310(5755):1782–6.

[164] Engelken J, Carnero-Montoro E, Pybus M, Andrews GK, Lalueza-Fox C, Comas D, et al. Extreme population differences in the human zinc transporter ZIP4 (SLC39A4) are explained by positive selection in Sub-Saharan Africa. PLoS genetics. 2014 Feb;10(2):e1004128.

[165] The Gene Ontology Consortium. Gene Ontology : tool for the. Nature genetics. 2000;25:25–29.

[166] Mi H, Muruganujan A, Thomas PD. PANTHER in 2013 : modeling the evolution of gene function , and other gene attributes , in the context of phylogenetic trees. Nucleic acids research. 2013;41(Database issue):377–386.

[167] Kanehisa M, Goto S. KEGG : Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research. 2000;28(1):27–30.

[168] Croft D, Kelly GO, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome : a database of reactions , pathways and biological processes. Nucleic acids research. 2011;39(Database issue):691–697.

[169] Marques-Bonet T, Ryder Oa, Eichler EE. Sequencing primate genomes: what have we learned? Annual review of genomics and human genetics. 2009 Jan;10:355–86.

[170] Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, et al. Patterns of positive selection in six Mammalian genomes. PLoS genetics. 2008 Jan;4(8):e1000144.

[171] Serra F, Arbiza L, Dopazo J, Dopazo H. Natural selection on functional modules, a genome-wide analysis. PLoS Computational Biology. 2011;7(3):e10001093.

[172] Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-murci L, Robinson-rechavi M, et al. Evidence for Polygenic Adaptation to Pathogens in the Human Genome Article Fast Track. Molecular biology and evolution. 2013;30(7):1544–1558.

[173] Barreiro LB, Quintana-Murci L. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nature reviews Genetics. 2010;11(1):17–30.

[174] Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Ferrer-Admettla A, Pattini L, et al. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS genetics. 2011 Nov;7(11):e1002355.

[175] Hancock AM, Witonsky DB, Alkorta-aranburu G, Beall CM, Sukernik R, Utermann G, et al. Adaptations to Climate-Mediated Selective Pressures in Humans. PLoS genetics. 2011;7(4):e1001375.

[176] King Mc, Wilson AC. Humans and Chimpanze es. Science. 1975;188(4184):107–116.

[177] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research. 2005 Aug;15(8):1034–50.

[178] Fraser HB. Gene expression drives local adaptation in humans. Genome research. 2013 Jul;23(7):1089–96.

[179] Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray Ga. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nature genetics. 2007 Sep;39(9):1140–4.

[180] Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK. Gene expression levels are a target of recent natural selection in the human genome. Molecular biology and evolution. 2009 Mar;26(3):649–58.

[181] Dawkins R. The Selfish Gene. Oxford University Press; 1976.

[182] Lehner B. Molecular mechanisms of epistasis within and between genes. Trends in Genetics. 2011;27(8):323–331.

[183] Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. Epistasis as the primary factor in molecular evolution. Nature. 2012;490(7421):535–538.

[184] Kondrashov AS, Sunyaev S, Kondrashov FA. Dobzhansky – Muller incompatibilities in. Proceedings of the National Acadamy of Sciences. 2002;99(23):14878–14883.

[185] Lovell SC, Robertson DL. An integrated view of molecular coevolution in protein-protein interactions. Molecular biology and evolution. 2010 Nov;27(11):2567–75.

[186] Fryxell KJ. The coevolution of gene familv trees. Trends in Genetics. 1996;12(9):394–369.

[187] Doherty A, Alvarez-Ponce D, McInerney JO. Increased genome sampling reveals a dynamic relationship between gene duplicability and the structure of the primate protein-protein interaction network. Molecular biology and evolution. 2012 Nov;29(11):3563–73.

[188] Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary Rate in the Protein Interaction Network. Science. 2002;296(2002):750–752.

[189] Agrafioti I, Swire J, Abbott J, Huntley D, Butcher S, Stumpf MPH. Comparative analysis of the Saccharomyces cerevisiae and Caenorhabditis elegans protein interaction networks. BMC evolutionary biology. 2005 Jan;5:23.

[190] Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Molecular biology and evolution. 2005 May;22(5):1345–1354.

[191] Cui Q, Purisima E, Wang E. Protein evolution on a human signaling network. BMC Systems Biology. 2009;3(1):21.

[192] Clark NL, Aquadro CF. A novel method to detect proteins evolving at correlated rates: identifying new functional relationships between coevolving proteins. Molecular biology and evolution. 2010 May;27(5):1152–61.

[193] Alvarez-Ponce D, Fares Ma. Evolutionary rate and duplicability in the Arabidopsis thaliana protein-protein interaction network. Genome biology and evolution. 2012 Jan;4(12):1263–74.

[194] Wang GZ, Lercher MJ. The Effects of Network Neighbours on Protein Evolution. PLoS ONE. 2011 Apr;6(4):e18288.

[195] Clark NL, Alani E, Aquadro CF. Evolutionary rate covariation reveals shared functionality and coexpression of genes. Genome research. 2012 Apr;22(4):714–20.

[196] Freeman LC. A Set of Measures of Centrality Based on Betweenness. Sociometry. 1977;40(1):35–41.

[197] Qian W, He X, Chan E, Xu H, Zhang J. Measuring the evolutionary rate of protein-protein interaction. Proceedings of the National Academy of Sciences of the United States of America. 2011 May;108(21):8725–8730.

[198] Olson-Manning CF, Wagner MR, Mitchell-Olds T. Adaptive evolution: evaluating empirical support for theoretical predictions. Nature reviews Genetics. 2012 Dec;13(12):867–77.

[199] Orr HA. The genetic theory of adaptation: a brief history. Nature reviews Genetics. 2005 Mar;6(2):119–27.

[200] Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of Genomic Variants Using a Unified Biological Network Approach. PLoS Computational Biology. 2013;9(3).

[201] Rausher M, Miller RE, Tiffin P. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Molecular biology and evolution. 1999 Feb;16(2):266–274.

[202] Flowers JM, Sezgin E, Kumagai S, Duvernell DD, Matzkin LM, Schmidt PS, et al. Adaptive evolution of metabolic pathways in Drosophila. Molecular biology and evolution. 2007;24(6):1347–54.

[203] Livingstone K, Anderson S. Patterns of variation in the evolution of carotenoid biosynthetic pathway enzymes of higher plants. The Journal of heredity. 2009;100(6):754–761.

[204] Yang Yh, Zhang Fm, Ge S. Evolutionary rate patterns of the Gibberellin pathway genes. BMC Evolutionary Biology. 2009 Jan;9:206.

[205] Ramsay H, Rieseberg LH, Ritland K. The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. Molecular Biology and Evolution. 2009;26(5):1045–1053.

[206] Montanucci L, Laayouni H, Dall'Olio GM, Bertranpetit J. Molecular evolution and network-level analysis of the N-glycosylation metabolic pathway across primates. Molecular biology and evolution. 2011 Jan;28(1):813–823.

[207] Riley RM, Jin W, Gibson G. Contrasting selection pressures on components of the Ras-mediated signal transduction pathway in Drosophila. Molecular Ecology. 2003 May;12(5):1315–1323.

[208] Alvarez-Ponce D, Aguadé M, Rozas J. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. Genome Research. 2009;p. 234–242.

[209] Alvarez-Ponce D, Aguadé M, Rozas J. Comparative Genomics of the Vertebrate Insulin / TOR Signal Transduction Pathway : A Network-Level Analysis of. Genome biology and evolution. 2011;3:87–101.

[210] Invergo BM, Montanucci L, Laayouni H, Bertranpetit J. A system-level , molecular evolutionary analysis of mammalian phototransduction. BMC Evolutionary Biology. 2013;13:52.

[211] Lavagnino N, Serra F, Arbiza L, Dopazo H, Hasson E. Evolutionary Bioinformatics Evolutionary Genomics of Genes Involved in Olfactory Behavior in the Drosophila melanogaster Species Group. Evolutionary bioinformatics online. 2012;8:89–104.

[212] Fitzpatrick DA, Halloran DMO. Investigating the Relationship between Topology and Evolution in a Dynamic Nematode Odor Genetic Network. International Journal of Evolutionary Biology. 2012;2012.

[213] Lu Y, Rausher MD. Evolutionary rate variation in anthocyanin pathway genes. Molecular biology and evolution. 2003 Nov;20(11):1844–53.

[214] Rausher M, Lu Y, Meyer K. Variation in constraint versus positive selection as an explanation for evolutionary rate variation among anthocyanin genes. Journal of molecular evolution. 2008 Aug;67(2):137–144.

[215] Laportes DC, Walsh K, Koshland DE. The Branch Point Effect. Ultrasensitivity and subsensitivity to metabolic control. Journal of Biological Chemistry. 1984;259(22):14068–14075.

[216] Olson-Manning CF, Lee CR, Rausher MD, Mitchell-Olds T. Evolution of flux control in the glucosinolate pathway in Arabidopsis thaliana. Molecular Biology and Evolution. 2013 Jan;30(1):14–23.

[217] Eanes WF. Analysis of Selection on Enzyme Polymorphisms. Annual Review of Ecology and Systematics. 1999;30(65):301–26.

[218] Whitt SR, Wilson LM, Tenaillon MI, Gaut BS, Iv ESB. Genetic diversity and selection in the maize starch pathway. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(20):12959–12962.

[219] Olsen KM, Womack A, Garrett AR, Suddith JI, Purugganan MD. Contrasting Evolutionary Forces in the Arabidopsis thaliana Floral Developmental Pathway. Genetics. 2002;1:1641–1650.

[220] Alvarez-Ponce D, Guirao-Rico S, Orengo DJ, Segarra C, Rozas J, Aguadé M. Molecular population genetics of the insulin/TOR signal transduction pathway: a network-level analysis in Drosophila melanogaster. Molecular Biology and Evolution. 2012 Jan;29(1):123–132.

[221] Casals F, Sikora M, Laayouni H, Montanucci L, Muntasell A, Lazarus R, et al. Genetic adaptation of the antibacterial human innate immunity network. BMC evolutionary biology. 2011 Jan;11(1):202.

[222] Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nature reviews Genetics. 2004 Feb;5(2):101–13.

[223] Yamada T, Bork P. Evolution of biomolecular networks - lessons from metabolic and protein interactions. Nature Reviews Molecular Cell Biology. 2009;10(11):791–803.

[224] Jeong H, Albert R. The large-scale organization of metabolic networks. Nature. 2000;407(1990):651–654.

[225] Wuchty S. Evolution and topology in the yeast protein interaction network. Genome research. 2004 Jul;14(7):1310–4.

[226] Jordan IK, Wolf Y, Koonin E. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evolutionary Biology. 2003;3(1):1.

[227] Bloom JD, Adami C. Apparent dependence of protein evolutionary rate on number of sets. BMC Evolutionary Biology. 2003;3:21.

[228] Zhu X, Gerstein M, Snyder M. Getting connected: analysis and principles of biological networks. Genes & development. 2007 May;21(9):1010–24.

[229] Bloom JD, Adami C. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: response. BMC evolutionary biology. 2004 Jun;4:14.

[230] Drummond DA, Raval A, Wilke CO. A single determinant dominates the rate of yeast protein evolution. Molecular biology and evolution. 2006 Feb;23(2):327–37.

[231] Fraser H, Hirsh A. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. BMC Evolutionary Biology. 2004;4(1):13.

[232] Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Molecular biology and evolution. 2005 Apr;22(4):803–806.

[233] Fraser HB. Modularity and evolutionary constraint on proteins. Nature genetics. 2005 Apr;37(4):351–2.

[234] Kim PM, Korbel JO, Gerstein MB. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. Proceedings of the National Academy of Sciences. 2007 Dec;104(51):20274–20279.

[235] Förster J, Famili I, Fu P, Palsson BO, Nielsen J. Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic Network. Genome research. 2003;13:244–253.

[236] Duarte NC, Herrgå rd MJ, Palsson BO. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome research. 2004 Jul;14(7):1298–309.

[237] Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Acadamy of Sciences. 2007;104(6):1777–1782.

[238] Wagner A, Fell DA. The small world inside large metabolic networks. Proceedings of the Royal Society B: Biological Sciences. 2001;268:1803–1810.

[239] Ma H, Zeng Ap. genome data and analysis of their global. Bioinformatics. 2003;19(2):270–277.

[240] Ravasz E, Somera aL, Mongru Da, Oltvai ZN, Barabási aL. Hierarchical organization of modularity in metabolic networks. Science (New York, NY). 2002 Aug;297(5586):1551–5.

[241] Hahn MW, Conant GC, Wagner A. Molecular evolution in large genetic networks: does connectivity equal constraint? Journal of Molecular Evolution. 2004;58(2):203–211.

[242] Vitkup D, Kharchenko P, Wagner A. Influence of metabolic network structure and function on enzyme evolution. Genome biology. 2006;7(5):R39.

[243] Greenberg AJ, Stockwell SR, Clark AG. Evolutionary Constraint and Adaptation in the Metabolic Network of Drosophila. Molecular Biology and Evolution. 2008;25(12):2537–2546.

[244] Lu C, Zhang Z, Leach L, Kearsey MJ, Luo ZW. Impacts of yeast metabolic network structure on enzyme evolution. Genome biology. 2007 Jan;8(8):407.

[245] Papp B, Oliver S. Genome-wide analysis of the context-dependence of regulatory networks. Genome Biology. 2005;6:206.

[246] Babu MM, Ã NML, Aravind L, Gerstein M, Teichmann SA. Structure and evolution of transcriptional regulatory networks. Current Opinion in Structural Biology. 2004;14:283–291.

[247] Rodriguez-Caso C, Medina Ma, Solé RV. Topology, tinkering and evolution of the human transcription factor network. The FEBS journal. 2005 Dec;272(24):6423–34.

[248] Yu H, Gerstein M. Genomic analysis of the hierarchical structure of regulatory networks. Proceedings of the National Acadamy of Sciences. 2006;103:14724–14731.

[249] Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012 Sep;489(7414):91–100.

[250] Evangelisti AM, Wagner A. Molecular evolution in the yeast transcriptional regulation network. Journal of Experimental Zoology (Mol Dev Evol)2. 2004;302:392–411.

[251] Wang Y, Franzosa EA, Zhang Xs, Xia Y. Protein evolution in yeast transcription factor subnetworks. Nucleic Acids Research. 2010;38(18):5959–5969.

[252] Jovelin R, Phillips PC. Evolutionary rates and centrality in the yeast gene regulatory network. Genome Biology. 2009;10(4):R35.

[253] Jordan IK, Mariño Ramírez L, Wolf YI, Koonin EV. Conservation and coevolution in the scale-free human gene coexpression network. Molecular biology and evolution. 2004 Nov;21(11):2058–70.

[254] MacArthur DG, Balasubramanian S, Frankish A, Huang N, Walter K, Jostins L, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335(6070):823–828.

[255] Liao By, Zhang J. Null mutations in human and mouse orthologs frequently result in different phenotypes. Proceedings of the National Acadamy of Sciences. 2008;105(19):6987–6992.

[256] Limdi Na, Wadelius M, Cavallari L, Eriksson N, Crawford DC, Lee MTM, et al. Warfarin pharmacogenetics: a single VKORC1 polymorphism is predictive of dose across 3 racial groups. Blood. 2010 May;115(18):3827–34.

[257] Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, et al. Genomic regions exhibiting positive selection identified from dense genotype data Genomic regions exhibiting positive selection identified from dense genotype data. Genome Research. 2005;15:1553–1565.

[258] Wang ET, Kodama G, Baldi P, Moyzis RK. Global landscape of recent inferred Darwinian selection for Homo sapiens. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(1):13–140.

[259] Teo Yy, Sim X, Ong RTH, Tan AKS, Chen J, Tantoso E, et al. Singapore Genome Variation Project : A haplotype map of three Southeast Asian populations Singapore Genome Variation Project : A haplotype map of three Southeast Asian populations. Genome Research. 2009;19:2154–2162.

[260] Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. Nature genetics. 2008;40(3):340–345.

[261] Ross Ka, Bigham AW, Edwards M, Gozdzik A, Suarez-Kurtz G, Parra EJ. Worldwide allele frequency distribution of four polymorphisms associated with warfarin dose requirements. Journal of human genetics. 2010 Sep;55(9):582–9.

[262] Trynka G, Hunt Ka, Bockett Na, Romanos J, Mistry V, Szperl A, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nature genetics. 2011 Dec;43(12):1193–201.

[263] Patillon B, Luisi P, Blanché H, Patin E, Cann HM, Génin E, et al. Positive Selection in the Chromosome 16 VKORC1 Genomic Region Has Contributed to the Variability of Anticoagulant Response in Humans. PloS one. 2012 Jan;7(12):e53049.

[264] Laayouni H, Oosting M, Luisi P, Ioana M, Alonso S, Ricaño Ponce I, et al. Convergent evolution in European and Rroma populations reveals pressure exerted by plague on Toll-like receptors. Proceedings of the National Academy of Sciences of the United States of America. 2014 Mar;111(7):2668–73.

[265] Luisi P, Alvarez-Ponce D, Dall'Olio GM, Sikora M, Bertranpetit J, Laayouni H. Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations. Molecular biology and evolution. 2012 May;29(5):1379–92.

[266] Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, et al. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS genetics. 2009 Jul;5(7):e1000562.

[267] Medzhitov R. Toll-like receptors and innate immunity. Nature reviews Immunology. 2001 Nov;1(2):135–45.

[268] Dall'Olio GM, Jassal B, Montanucci L, Gagneux P, Bertranpetit J, Laayouni H. The annotation of the asparagine N-linked glycosylation pathway in the Reactome database. Glycobiology. 2011 Nov;21(11):1395–400.

[269] Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, et al. The BioGRID Interaction Database: 2011 update. Nucleic acids research. 2011 Jan;39(Database issue):D698–704.

[270] Sowa ME, Bennett EJ, Gygi SP, Harper JW. Defining the human deubiquitinating enzyme interaction landscape. Cell. 2009 Jul;138(2):389–403.

[271] Bandyopadhyay D, Huan J, Liu J, Prins J, Snoeyink J, Wang W, et al. Functional neighbors: inferring relationships between nonhomologous protein families using family-specific packing motifs. IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society. 2010 Sep;14(5):1137–43.

[272] von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, et al. Comparative assessment of large-scale data sets of protein-protein interactions. Nature. 2002 May;417(6887):399–403.

[273] Bader JS, Chaudhuri A, Rothberg JM, Chant J. Gaining confidence in high-throughput protein interaction networks. Nature biotechnology. 2004 Jan;22(1):78–85.

[274] Deeds EJ, Ashenberg O, Shakhnovich EI. A simple physical model for scaling in protein-protein interaction networks. Proceedings of the National Academy of Sciences of the United States of America. 2006 Jan;103(2):311–6.

[275] Kelly WP, Stumpf MPH. Assessing coverage of protein interaction data using capture-recapture models. Bulletin of mathematical biology. 2012 Feb;74(2):356–74.

[276] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database–2009 update. Nucleic acids research. 2009 Jan;37(Database issue):D767–72.

[277] Messer PW, Petrov Da. Frequent adaptation and the McDonald-Kreitman test. Proceedings of the National Academy of Sciences. 2013 May;110(21):8615–8620.

[278] Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. Nature. 2013 Jul;499(7459):471–5.

[279] Vidal M, Cusick ME, Barabási AL. Interactome Networks and Human Disease. Cell. 2011;144(6):986–998.

[280] Dall'Olio GM, Laayouni H, Luisi P, Sikora M, Montanucci L, Bertranpetit J. Distribution of events of positive selection and population differentiation in a metabolic pathway: the case of asparagine N-glycosylation. BMC evolutionary biology. 2012 Jan;12:98.

[281] Rausher MD. The evolution of genes in branched metabolic pathways. Evolution; international journal of organic evolution. 2012 Jan;67(1):34–48.

[282] Wright KM, Rausher MD. The evolution of control and distribution of adaptive mutations in a metabolic pathway. Genetics. 2010 Feb;184(2):483–502.

[283] Martin G. Fisher's Geometrical Model Emerges as a Property of Complex Integrated Phenotypic Networks. Genetics. 2014 May;197(1):237–55.

[284] Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, et al. Evolution of protein-coding genes in Drosophila. Trends in genetics. 2008 Mar;24(3):114–23.

[285] Morgan GJ. Emile Zuckerkandl, Linus Pauling, and the molecular evolutionary clock, 1959-1965. Journal of History of Biology. 1988;31(2):155–78.

[286] Zuckerkandl E. Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. Journal of Molecular Evolution. 1976;7(3):167–83.

[287] Pál C, Papp B, Hurst LD. Does the Recombination Rate Affect the Efficiency of Purifying Selection ? The Yeast. Molecular biology and evolution2. 2001;18:2323–2326.

[288] Rocha EPC, Danchin A. An analysis of determinants of amino acids substitution rates in bacterial proteins. Molecular biology and evolution. 2004 Jan;21(1):108–16.

[289] Wilson AC, Carlson SS, White TJ. Biochemical evolution. Annual Review of Biochemistry. 1977;46:573–639.

[290] Salathé M, Ackermann M, Bonhoeffer S. The effect of multifunctionality on the rate of evolution in yeast. Molecular biology and evolution. 2006 Apr;23(4):721–2.

[291] Podder S, Mukhopadhyay P, Ghosh TC. Multifunctionality dominantly determines the rate of human housekeeping and tissue specific interacting protein evolution. Gene. 2009 Jun;439(1-2):11–6.

[292] Hurst LD, Smith NGC. Do essential genes evolve slowly ? Current Biology. 1999;9(14):747–750.

[293] Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. Genome Research. 2002;12:962–968.

[294] Hirsh AE, Fraser HB. Protein dispensability and rate of evolution. Nature. 2001;411:1046–1049.

[295] Yang J, Gu Z, Li WH. Rate of protein evolution versus fitness effect of gene deletion. Molecular biology and evolution. 2003 May;20(5):772–4.

[296] Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, et al. Functional genomic analysis of the rates of protein evolution. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(5483-5488).

[297] Duret L, Mouchiroud D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Molecular biology and evolution. 2000 Jan;17(1):68–74.

[298] Pál C, Papp B, Hurst Lawrence D . Letter to the Editor. Genetics. 2001;158(1998).

[299] Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. Proceedings of the National Academy of Sciences of the United States of Americas. 2005;102(40):14338–14343.

[300] Liao BY, Scott NM, Zhang J. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Molecular biology and evolution. 2006 Nov;23(11):2072–80.

[301] Ingvarsson PK. Gene expression and protein length influence codon usage and rates of sequence evolution in Populus tremula. Molecular biology and evolution. 2007 Mar;24(3):836–44.

[302] Plotkin JB, Fraser HB. Assessing the determinants of evolutionary rates in the presence of noise. Molecular biology and evolution. 2007 May;24(5):1113–21.

[303] Alvarez-Ponce D. The relationship between the hierarchical position of proteins in the human signal transduction network and their rate of evolution. BMC Evolutionary Biology. 2012;12(1):192.

[304] Pál C, Papp B, Lercher MJ. An integrated view of protein evolution. Nature reviews Genetics. 2006 May;7(5):337–48.

[305] Subramanian S, Kumar S. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics. 2004 Sep;168(1):373–81.

[306] Quach H, Barreiro LB, Laval G, Zidane N, Patin E, Kidd KK, et al. Signatures of purifying and local positive selection in human miRNAs. American journal of human genetics. 2009 Mar;84(3):316–27.

[307] Smith JD, McManus KF, Fraser HB. A novel test for selection on cis-regulatory elements reveals positive and negative selection acting on mammalian transcriptional enhancers. Molecular biology and evolution. 2013 Nov;30(11):2509–18.

[308] Colombo M, Laayouni H, Invergo BM, Bertranpetit J, Montanucci L. Metabolic flux is a determinant of the evolutionary rates of enzyme-encoding genes. Evolution. 2013;68(2):605–613.

[309] Rhoné B, Brandenburg JT, Austerlitz F. Impact of selection on genes involved in regulatory network: a modelling study. Journal of evolutionary biology. 2011 Oct;24(10):2087–98.