SIGNATURES OF SELECTION IN THE INDEL-CONTAINING CODING SEQUENCES FROM HUMAN TO PRIMATE GENOME COMPARISONS

By

Wilfried Marie Francis Guiblet

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master in Science

In

Biology

University of Puerto Rico

Mayaguez

2013

Approved by:

Juan C. Martínez Cruzado Member, Graduate Committee	Date
Carlos Ríos Velázquez Member, Graduate Committee	Date
Steven E. Massey Member, Graduate Committee	Date
Taras K. Oleksyk President, Graduate Committee	Date
Juan A. Ortíz Representative of Graduate Studies	Date
Nanette Diffoot Chairperson of the Department	Date

ABSTRACT

Gene sequences are relatively conserved, and usually show few differences in comparisons between closely related species, such as between humans and non-human primates. In this study, we focused on >10 bp insertions and deletions (Indels) found in the alignments between the human, chimpanzee, gorilla, orangutan, and rhesus macaque reference genome sequences and examined these regions in order to characterize signatures of adaptive and non-adaptive evolution in the phylogenetic lineage leading to our own species. A public data set of 36,422 Indels identified by comparing the reference genomes was filtered to set aside 146 Indels within coding sequences (with a potentially high impact on proteins). Among these, 80 fragments were successfully amplified by PCR and visualized on electrophoresis gels to distinguish real features from the computational artifacts. Only 22 Indels could be related to specific feature in the sequence alignment using the reference genomes. These Indel-containing genes were interrogated for the signatures of selection with *PAML* package by producing pairwise Ka/Ks ratios in all species comparisons. The significance of this approach was evaluated by a resampling method, where exactly the same procedures and tests were performed with a dataset of randomly created Indels (simdels) matched by size and distributed across the reference genomes. Indels showed significantly higher Ka/Ks ratios indicating that they were located in less constrained sequences, and a trend was observed with first exons showing the largest difference between the observed Indels and simdels. We also searched for more recent signatures of selection by searching for the chromosomal regions demonstrating diminished multilocus heterozygosity and high population divergence (F_{ST}) by comparing dense genotyping data in the moving

windows along the chromosomes between populations of the Human Genome Diversity Project (HGDP). We identified several genes by comparing the observed distribution to a distribution of simdels and discussed our results in from the prospective of relevant evolutionary history during major human migrations.

RESUMEN

Las secuencias de los genes son relativamente conservadas, y usualmente demuestran pocas diferencias entre especies estrechamente relacionadas, tal como humanos y primates. En este estudio, nos enfocamos en las inserciones y deleciones (Indels) > 10 pb que hemos descubierto comparando las secuencias de genomas de referencia de humanos, chimpancés, gorilas, orangutanes y macacos, y examinando estas regiones para caracterizar huellas de evolución adaptativa y no adaptativa en el linaje que compartimos los humanos y primates. Un banco de datos públicos de 36,422 Indels que fueron identificados comparando los genomas de referencia fue filtrado para extraer 146 Indels predichos a estar dentro de secuencias codificantes (con un potencial de impacto fuerte sobre las proteínas). De éstos, 80 fragmentos fueron amplificados con éxito por PCR y corridos en geles de electroforesis para distinguir los Indels reales de aquellos creados por sesgos en los métodos computacionales. Sin embargo, solamente 22 pudieron ser relacionados a Indels específicos en un alineamiento de genes usando los genomas de referencia. Para buscar huellas de selección, usé el paquete PAML para producir razones de Ka/Ks entre especies para los 22 genes que contienen un Indel validado. El significado de la prueba fue evaluado por un método de re-muestreo, donde los mismos procedimientos y pruebas fueron llevadas a cabo sobre un banco de datos de Indels creados y distribuidos a lo largo de los genomas de referencia al azar (simdels), y que eran iguales en tamaño a los Indels reales. Los Indels demostraron unas razones significativamente mayores de Ka/Ks, indicando que están localizados en secuencias bajo menores restricciones selectivas, y una tendencia fue observada en la que los primeros exones demostraron una diferencia mayor entre los Indels

observados y los simdels. También buscamos huellas de selección más recientes, buscando regiones cromosómicas demostrando una disminución en varios lugares de heterocigosidad y un aumento en divergencia poblacional (FST) al comparar datos genotípicos densos entre las poblaciones del Human Genome Diversity Project (HGDP). Identificamos varios genes comparando la distribución observada a la distribución de los simdels y discutimos nuestros resultados desde el punto de vista de la historia evolutiva de las grandes migraciones humanas.

© Wilfried Marie Francis Guiblet, 2013

To my family and friends I left to live a dream.

To my dad, I hope he would have been proud.

ACKNOWLEDGEMENTS

Frist, I would like to thank my committee of professors for their support since the beginning of my studies in the University of Puerto Rico at Mayagüez. Dr. Carlos Rios Velazquez was the first professor I met in this university and followed my progresses ever since. I took almost every class from Dr. Juan Carlos Martinez Cruzado and learned a lot about genetics under his supervision. Dr. Steven Massey introduced me to the world of bioinformatics, which would have been much more frightening without him. And last but not least, Dr. Taras Oleksyk, my chairman, taught me much more than science. Again, I want to thank them so much.

I address my special thanks to everyone that was involved in this project: Dr. Natalia Volfovsky and Dr. Robert Stephens from the NCI-Frederick for providing the databases; Anyi Mazo Vargas, Daysha Ferrer Torres, Christina Ruiz Rodriguez, Angeliz Caro Monroig and Angelia Caro Monroig for the many hours spent validating the Indels by PCR; and finally Dr. Luis Figueroa from the Puerto Rico Zoo for providing rhesus macaque samples.

Also, I want to say that this work would have never been possible without the support of our National Science Foundation grant: MCB-1019454 RIG: Comparative Genomics of Indels in Primate Lineages 2010-2012.

I want to thank my colleagues and friends from the Laboratory of Genomics Diversity: Yashira Afanador, Israel Rivera, Stefanie Cosme, Ingrid Rivera, Frances Marin and Priscila Rodriguez. It was a real pleasure to work with them all this time.

Finally, I am very thankful for the welcoming of the Puerto Rican people. Their kindness was really welcome so far from my home. I met great friends here and I will always feel attached to this place.

Table of Contents

List of Tables

List of Figures

Introduction

In the last two decades, the life sciences have seen a revolution with the development of new technologies enabling whole new disciplines like genomics, the detailed study of the entire genetic information contained in organisms. It started with the amazing international effort called the Human Genome Project (HGP), which brought insights on the genetic information and developed the technologies to bring biology to a totally new scale (Lander et al. 2001; Venter et al. 2001; Naidoo et al. 2011). After the HGP, effort has been shifted to describe the diversity of human genomes. First, the International HapMap Project (Altshuler et al. 2010) and the Human Genome Diversity Panel (Cavalli-Sforza 2005) genotyped human populations worldwide. As sequencing costs continued to fall, population-wide resequencing has become feasible: the 1,000 Genomes Project has been able to sequence 1,092 human genomes in 14 populations, and project this number to increase to 2,500 by the end of the study (Abecasis et al. 2012). In addition, technologies developed from the Human Genome Project and the human reference genome enabled several primates' genome projects: chimpanzee (Consortium 2005), gorilla (Scally et al. 2012), orangutan (Locke et al. 2011), rhesus macaque (Gibbs et al. 2007). As the technologies continue to improve, more and more data is becoming publicly available, and more sequencing is being done every day. It is becoming common for a country to have its own human genome diversity project, Iceland being one example (Palsson et al. 1999). Recently, the United Kingdom announced a project aiming to sequence 100,000 patient genomes to adapt genomic medicine (www.genomeweb.com, December 10 2012). Meanwhile, genomes of other species continue to accumulate giving insight to the structural and functional

differences on the evolutionary scale: the Genome 10K Project intends to resequence more than 10,000 vertebrate species (2009; Haussler 2009). Whole genome sequencing has fallen in price so rapidly that has become available even for local community to fund a project and to perform advanced genomic studies in an undergraduate setting (Oleksyk et al. 2012).

For the first time in science history, molecular and informatics technologies made the data faster (and cheaper) to produce than process. We entered in a new area where the challenge of new discoveries reside less in obtaining the information than making sense of it. The amount of data provided by Next Generation Sequencing (NGS) and Genome-Wide Association Studies (GWAS) is the key for major breakthrough in health care (personalized medicine) and in our understanding of evolution. The new generation of scientists needs to adapt their approaches and techniques: it is now necessary for a biologist to increase their skills in statistics and informatics, as the next decades might bring profound changes in our global knowledge, and even modify our lifestyles (Collins et al. 2003).

This study aims to bring some insights about a particular kind of mutations: insertions and deletions (Indels) that occurred in the evolutionary lineage leading to our own species. First, the computationally predicted Indels from the reference genomes comparisons were validated by molecular methods. Then, their evolutionary impact was described in interspecific (between hominids) and intraspecific (between human populations) comparisons. Finally, the resulting information brings new leads on hominids and human divergence and proposed several genes as strong candidates for adaptation.

Literature Review

Correlation between genetic events and natural selection

The changes in phenotypic characters can largely be traced to genetic mutations such as single nucleotide substitutions (SNPs), insertions and deletions, segment duplications, chromosomal rearrangements, inversions, and translocations. Most of these mutations will be removed by genetic drift, at a rate proportional to the effective population size (N_e) (Charlesworth 2009). At the same time, most mutations with any impact on fitness are deleterious or slightly deleterious, and get quickly removed from the genetic pool by the evolutionary force of negative (or purifying) selection. Approximately 5.5% of the human genome has undergone purifying selection (Lindblad-Toh et al. 2011). Once in a while, new mutations rise to fixation by chance and, given sufficient time, many such changes with no or little effect on fitness accumulate across genomes contributing to neutral evolution. Occasionally, a new mutation (or a set of mutations) will improve the individual fitness and rise in frequency under the effect of directional selection. Once fixed, this mutation continues to be maintained by purifying selection, because most new mutations will have lower fitness values. In some rare cases, both the mutant and the ancestral alleles are maintained in a population, so the polymorphism can persist for many generations maintained by balancing selection (Hurst 2009). Directional selection increases frequency of the variant of interest much faster than drift, but mutations are not evolving independently from other variants in its neighborhood. A selection event actually raises the frequency of the selected haplotype around the mutation. Thus, when a haplotype reaches fixation by directional selection, variation around a

selected mutation disappears. This is called a selective sweep (Lewontin et al. 1973). Several generations after the sweep occurrence, a combination of mutation and recombination introduces new variations around the selected allele, and the sweep will progressively disappear (Sabeti et al. 2006; Oleksyk et al. 2010).

The accumulation of variants in isolated populations eventually leads to a speciation event (Wu et al. 2004). Thus, studying the evolutionary history of genetic variants helps us understand why and how our species evolved, and might explain the appearance of the key human traits (Lorente-Galdos et al. 2013). The molecular similarities and differences between modern humans and chimpanzees, our closest related species, have been increasingly studied since the beginning of the genomics era and led to the insights about human uniqueness (Varki et al. 2008), and how humans evolved to their current state. Natural selection pressures on humans have been related to the nervous, sensory, musculoskeletal, reproductive and immune systems, as to the skin and appendages (Nielsen et al. 2005; Walsh et al. 2005; Izagirre et al. 2006; Sabeti et al. 2006; Voight et al. 2006; Wang et al. 2006). However, excepting rare examples, there are still many discussions about which genes, and what kind of evolutionary mechanisms are involved in making humans unique (Varki et al. 2008).

Insertions and Deletions in evolutionary history of Humans and Primates

The data provided by the human and primate genetic studies is crucial to understand the divergence of the lineages and identify specific loci responsible for the speciation events. Indeed, due to the phylogenetic proximity of human to primates, and primates to each other, genomic comparisons can help to improve the

annotation of all the genomes. However, most of the studies have been focusing on SNPs, which are easier to validate than structural variants, such as Indels, with next-generation sequencing platforms (Dalca et al. 2010). They are challenging to discover and validate, and receive less attention (Mullaney et al. 2010). This is not well justified, since Indels (defined as sequences missing in comparisons of individuals or closely related species), are second most numerous class of polymorphisms in human genomes, and account for a large impact in our evolution (Wetterbom et al. 2006),

Considering more base pairs in human genomes are altered as result of structural variation than SNPs (Mullaney et al. 2010), many discoveries might rely on studying Indels. However, events leading to Indels arising are not easy to identify. Short Indels might be caused by polymerase slippage, bigger ones by various kinds of transposable elements. It has also been suggested that recombination events may lead to Indels appearance (Sjodin et al. 2010).

Due to the short fragment length used in next generation sequencing, the consecutive assembly process creates a noticeable amount of artificial gaps because of repeat sequences, resulting in a high ratio of false positive Indel detections (Albers et al. 2011). Experimental validating the presence of Indels, although difficult, is therefore necessary to remove (Volfovsky et al. 2009).

```
Denisovan    AAAAGAGATC CAGGGCTTCC TCGATTGTGC CGCGAGKKYT CAGGWNGCCC
Macaca       AAAAGAGATC CAGGGCTTCC TCGATTGTGC CGCGAAGGCT CAGGAAGCCC
Gorilla      AAAAGAGATC CAGGGCTTCC TCGATTGTGC CGCGAGGGCT CAGGAAGCCC
Homo         AAAAGAGATC CAGGGCTTCC TCGATTGTGC CGCGAGGGCT CAGGAAGCCC
Pan          AAAAGAGATC CAGGGCTTCC TCGATTGTGC CGCGAGGGCT CAGGAAGCCC

Denisovan    GAAAGATGAG ATCAATACAG GAAACCCTGG GAGAGTCTGG GAGTTTACTT
Macaca       GAAAGATGAG ATCAATAC.. .......... GAGAGTCTGG GAGTTTACCT
Gorilla      GAAAGATGAG ATCAATACAG GAAACCCTGG GAGAGTCTGG GAGTTTACTT
Homo         GAAAGATGAG ATCAATACAG GAAACCCTGG GAGAGTCTGG GAGTTTACTT
Pan          GAAAGATGAG ATCAATACAG GAAACCCTGG GAGAGTCTGG GAGTTTACTT

Denisovan    CCAAATAAAT TGAATAAGTT GTTACAGAGG TTTCCTAACA AACCTTACCT
Macaca       CCAAATAAAT TGAATAAATT GTTACAGAGG TTTCCTAACA AACCTCACCT
Gorilla      CCAAATAAAT TGAATAAGTT GTTACAGAGG TTTCCTAACA AACCTCACCT
Homo         CCAAATAAAT TGAATAAGTT GTTACAGAGG TTTCCTAACA AACCTTACCT
Pan          CCAAATAAAT TGAATAAGTT GTTACAGAGG TTTCCTAACA AACCTTACCT
```

Figure 1. Example of an Indel. Indels are gaps observed when aligning homologous sequences. As the alignment by itself does not provide enough information to state if the mutation was an insertion or a deletion, the gap is defined with the generic name of “Indel”.

In an earlier study evaluating Indels in different functional elements across the genome, fewer Indels were found in genes than expected after comparing it to a simulation under neutral (or random) evolution (Volfovsky et al. 2009). Since insertions or deletions of sequences bring major changes into the locus where they appear, they seem to be removed from coding sequences by strong purifying selection. Indeed, we expect most of these high impact Indels (i.e. able to change the coding sequence) to be deleterious, as many of them have been related to diseases like cystic fibrosis, fragil X syndrome, Huntingtons disease and many cancers (Sjodin et al. 2010). The size of the structure of the Indel influences strongly its impact: when its length in nucleotide is not divisible by 3, an exonic Indel (start and end contained inside an exon sequence) causes a frameshift. The location of the Indels also matters: if overlapping a splicing site, the mature mRNA can be significantly different. The higher the impact of the Indel, the higher we expect the selective pressures to be.

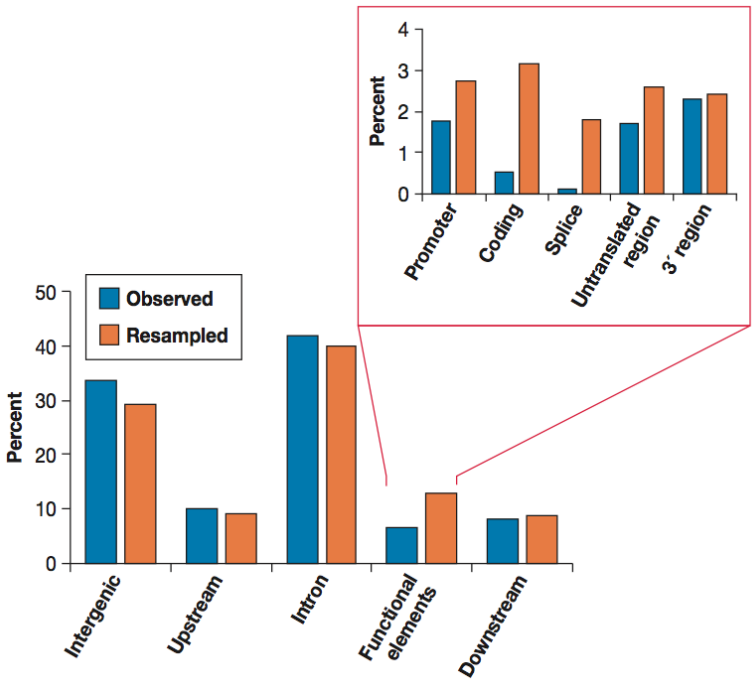

Figure 2. Distribution of Indels classified by their location relative to gene elements (figure from Volfovsky 2009). Real and randomized Indels show similar distributions in most locations, suggesting a neutral evolution of Indels in the studied genomes. However, less Indels were found in coding and splicing sequences than expected.

Considering that most of the high impact Indels to have appeared in human and primate evolutions might have been removed by drift and purifying selection, those variants that reached fixation for different alleles in recently diverged primate lineages are of particular interest. I predicted three different historical scenarios for the remaining high-impact Indels: the coding-sequence could be under (i) relaxed or neutral selection (no significant impact on gene function or fitness), (ii) directional selection (positive effect on fitness), or (iii) balancing selection (the presence of both allele in the population is the fittest). These scenarios must have have been of key importance in human and primates' divergence. In order to describe the potential selective pressures upon Indels in hominids, we validated high-impact Indels from predicted set and search for selection footprints in the genes they appeared in. Since selection-detecting algorithms are proficient for limited time scales, defined by

the underlying model assumptions (Sabeti et al. 2006; Oleksyk et al. 2010), two different approaches were used (presented below) to cover the evolution history of the Indels starting from their appearance in the common ancestors of humans and primates, and until the recent major population divergence among modern human populations.

Ratios of Synonymous and Non-Synonymous substitutions

Evolution of Indels has not yet been well described in the hominids lineage. In this study, we searched for the insights on the selective pressures around the Indels comparing closely related primate species. In inter-specific comparisons, such as between human and gorilla or human and chimpanzee, time scales are usually over several millions years (Wu et al. 2004). To detect selective sweeps that old, an efficient strategy is to observe the rates of synonymous and non-synonymous substitutions in coding sequences. Indeed, most changes in protein sequence tend to lower fitness, making non-synonymous mutations likely to be removed by purifying selection. In comparison, synonymous substitutions have far less impact (if any) on the protein sequence and can freely accumulate. Thus, even if the genetic code provides more opportunities for substitutions to be non synonymous, coding sequences proved to be much more permissive to synonymous substitutions (Nei et al. 1986). Comparing human to chimpanzee genomes showed that coding sequences are mainly under purifying selection (Bustamante et al. 2005).

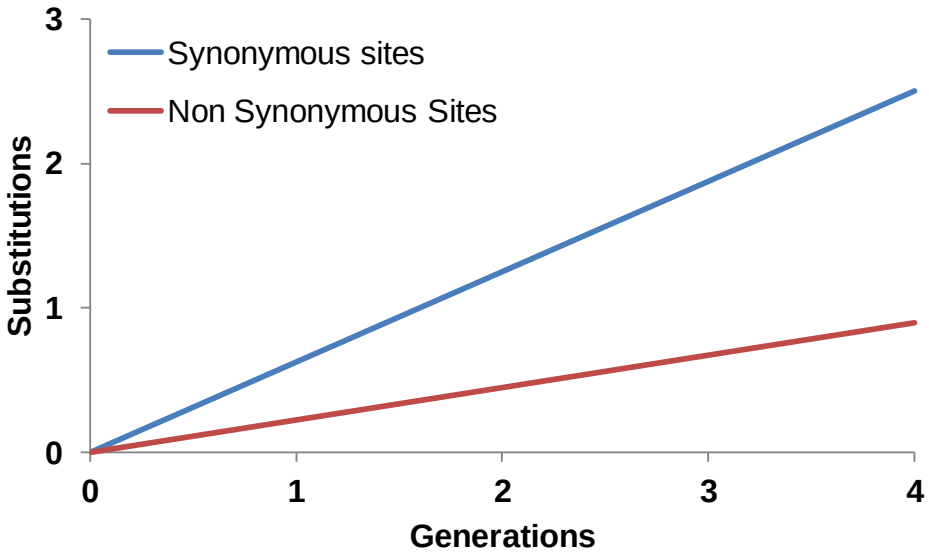

Figure 3. Schematic Representation of rates of synonymous and non synonymous substitutions in coding sequences (adopted and modified from (Li 1998)). Experimental observations have shown fewer substitutions in Non Synonymous Sites than Synonymous sites.

I chose to use the references of *Homo sapiens* and the great apes *Pan trolgodytes, Gorilla gorilla and Pongo abelii* to cover the hominids lineage, the best databases available for this lineage. In this study, *Macaca mulatta* was used as outgroup. Also, the recent publication of the Denisovan human genome made available a high coverage data of a hominid (Meyer et al. 2012), whose ancestor diverged from modern humans' about 350,000 years ago (Stoneking et al. 2011), which is more recent than human and chimpanzee. This data was included in order to represent a wider spectrum of evolutionary time in the interspecific comparisons.

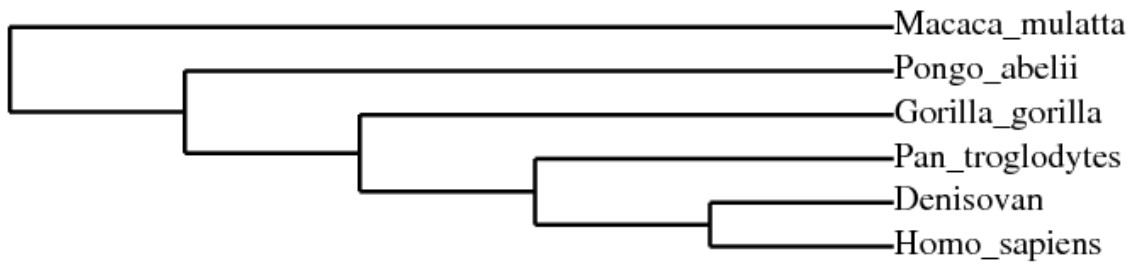

Figure 4. Phylogenetic tree of the hominids and out-group species presented in this study.

Variations of F_{ST} and Homozygosity

Human populations have been largely studied after the initial Human Genome Project. This provides a unique opportunity to study a potential correlation between Indels and selective sweeps between different populations. A previously published study (Oleksyk et al. 2008) showed an original strategy to discover selective sweeps, which has been then successfully used to find signatures of selection in a human population study (Zhao et al. 2012). It compares variation in F_{ST} and Homozygosity in actual genome regions against randomly sampled locations across the same genomes (simulating neutral scenario (Oleksyk et al. 2008)). Both F_{ST} and Homozygosity values rise to extreme in selective sweeps due to the fixation of alternatives alleles between populations. This method calculates how likely are the values of Homozygosity and F_{ST} at a certain locus to be have appeared by neutral evolution. When both F_{ST} and Homozygosity at a locus are both high, it indicates that a selective sweep may have occurred in this particular region (Oleksyk et al. 2008). It has been argued previously, that the multilocus F_{ST} variance (S^2F_{ST}) is more useful for the detecting signatures of selection, as F_{ST} mean or median may decrease when high and low values for alternatively fixed alleles across the window are added up

(Oleksyk et al. 2008). Therefore, in this sudy we used S^2F_{ST} estimates for the multilocus windows in our comparisons.

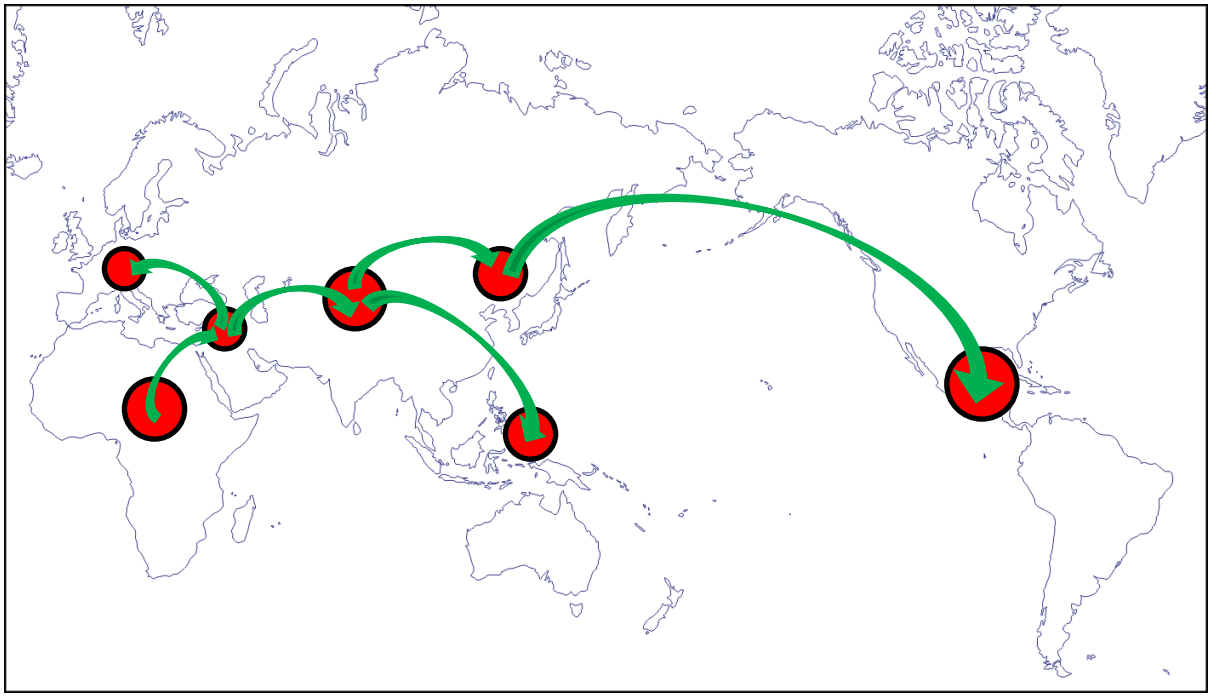

Figure 5. World map representing the continental comparisons following the great human migration (modified from: (Henn et al. 2012)). Populations (red dots) and their pairwise comparisons (green arrows) are consistent with the models describing the great human migration.

While there are other worldwide databases of genetic variation (HapMap, 1000Genomes, etc.), in this study I used the Human Genome Diversity Panel (HGDP) data, as it contains genotype information from the most represented ethnic groups worldwide (Cavalli-Sforza 2005). To reduce the number of comparisons, the number of comparisons was limited to reflect historic migrations by testing populations in the context of the global human migration history. The populations from HGDP where grouped in the following categories: Africa, Middle East, Europe, South Asia (which also contains Central Asia), Eastern Asia, Oceania and Americas. The populations were tested pairwise, and were assumed to have split from a single ancestral population in order to avoid artificial signal (mainly affecting the S^2F_{ST}).

From the 49 possible pairwise comparisons, I tested the six that followed the major routes of human migration history (Figure 5).

Objectives

1. To validate the predicted Indels using both informatics and molecular data.
2. To test coding sequences surrounding the observed and validated Indels for signatures of selection along the primate phylogeny.
3. To test genome regions (genes) containing the observed and valididated Indels for signatures of selection along the historic splits between the populations during the major human expansions.

Materials and Methods

Indel detection

We utilized a database of pairwise comparisons between chimpanzee (*Pan troglodytes*), gorillas (*Gorilla gorilla*), orangutans (*Pongo pygmaeus*), and rhesus macaque (*Macaca mulata*) genomes mapped to the reference human genome (Tolstorukov et al. 2012). Indels were identified with a procedure similar to that used for the published study characterizing and validating Indels discovered along a singe chromosome in human to chimpanzee comparison (Volfovsky et al. 2009). The database contained SNPs and Indels supported by at least three occurrences identified in traces downloaded from the NCBI Trace Archive and filtered from Simple repeats. The original database contains a total 1,059,367 primate Indels (Tolstorukov et al. 2012). We extracted a subset of this database containing insertions and deletions (from 10 bp to 10Kbp) surrounded by at least 10 bp of perfectly aligned sequence containing no more than 50% undetermined bases (Ns). In total, 36,422 Indels satisfied these criteria, of which 24,229 came from human-chimpanzee, 245 from human-gorilla, 8,895 from human-macaque and 3,053 came from the human-orangutan comparison. The current study focused on the Indels contained entirely within an exon (exonic) or overlapping a splice site (either donor or acceptor). Only 146 of these Indels were discovered in the human to four primate species comparisons. These Indels were also verified in Denisovans (see *Obtaining gene sequences*).

Indel validation

We designed universal primers and amplified these fragments in laboratory validation by PCR and electrophoresis. We used 400 bp of conserved sequence in the flanking regions of the selected Indel to identify primer pairs, and checked them for their uniqueness in relation to the copies on the chromosomes and in the rest of the genome. To validate the existence of the Indels, we first designed primers for each locus using NCBI Primer Blast. The primers were chosen to be at least 24 base pair long in flanking regions (5' upstream and 3' downstream in the fragment of interest). These conservative primers were optimized using different temperatures for the best PCR amplification for the fragments of interest. One of the primer sequences had to be unique in each genome and the other to have no more than 10 copies per genome. In some cases, primer pairs had to be selected manually by trial and error. The primer pairs were tested on a set of primate and human samples. Some samples were purchased from either the Integrated Primate Biomaterials and Information Resource or the Coriell Institute for Medical Research, Camden, NJ, and others obtained through collaborations with the NCI-Frederick and Puerto Rico Zoo Juan A. Rivero. The initial set of species included at least two unrelated samples of each chimpanzee, gorillas, orangutans, and rhesus macaques.

At each locus, 5ng of genomic DNA was amplified with AmpliTaq Gold (Applied Biosystems) with touchdown PCR protocol: 5 min heating at 94°C, 5 cycles of 30 sec at 94°C, 30 sec at 65°C, and 30 sec at 72°C, and 21 cycles at the same conditions, except lowering the annealing temperature by 0.5°C at each cycle (to 55°C), continued by 15 cycles of 30 sec annealing at 94°C, 30 sec at 55°C, and 30 sec at 72°C, and finished by 10 min of final extension at 72°C. Indels will be initially detected by PCR amplification, visualized and analyzed on 3% agarose gels.

Next, optimized PCR products were ran by electrophoresis on 3% agarose gels. The visualization of the fragments was used to validate the existence of the Indel. The Indel was considered "validated" if different sizes of fragments were detected in one of the five species. An Indel was considered not validated when there was no difference in fragments sizes between the five species (not shown). A database with the results was created. In the database we identified the primers that did amplify in the correct region in all of the five species. The Indels that did not amplify were re-optimized several times by changing the annealing temperatures.

A total of 146 Indels with highest potential effect on protein sequence were tested for validity by developing sets of universal primers and amplifying their regions in all five species of primates and examining resulting differences in length on electrophoresis gels. At this time, 80 Indels have been validated, and 18 have not been validated (Table 1).

Obtaining gene sequences

The sequences of each gene containing a validated exonic or splicing Indel were retrieved on the ENSEMBL website. The Indels were uplifted from hg 18 to hg19/GRCh37 (*Homo sapiens*), while CHIMP2.1.4 (*Pan troglodytes*), gorGor3.1 (*Gorilla gorilla*), PPYG2 (*Pongo abelii*) and MMUL_1 (*Macaca mulatta)* were used for the other four species of primates. I extracted the coding sequence (CDS) by looking up gene names in the database for each of the five species. If several transcripts were identified, I chose one according to these priorities: (1) presence of the consensus coding sequence (Pruitt et al. 2009), (2) largest amino acids chain length, (3) largest nucleotide sequence length. In some cases, the gene had not

been annotated in some species, leading to an incompletely covered lineage. In order to rebuild the Denisovan's CDSs of each gene of interest, I obtained the sequences of this recently discovered hominin species aligned to the human hg19/GRCh37 from the Max Plank Institute's website (http://www.eva.mpg.de/denisova). We performed an assembly of the sequences against the CDSs of interests from *Homo sapiens* using the Geneious software (Geneious v.5.5.6, created by Biomatters Inc. available from http://www.geneious.com) using the following parameters: medium sensitivity and no fine-tuning. Once the sequences were mapped, I built a consensus sequence for the Denisovan using a 75% threshold: the base at each position is found in at least 75% of the sequences used in the alignment to be consensual., If the sequence failed to map to the human genome, the base was not defined (and annotated as N).

Generating a simulated set of Indels (simdels)

As in the previous study (Volfovsky et al.,, 2009), a set of simulated Indels was produced by randomly placing insertions and deletions genome-wide *in silico*. These simulated Indels (or simdels) were produced in the same human genome reference sequence (hg18) as the observed set according to the following rules: (i) human genomic sequence was selected as the source of all sequences; (ii) coordinates of beginnings of randomly chosen sequences were selected from the range and frequency distribution of the analyzed human chromosomal fragments; (iii) the number of generated resampled sequences was 10 times larger than the number of Indels in the original data set. This new simulated data was filtered using the same criteria as the original set of observed Indels. Furthermore, the distributions of Indels

and simdels were matched by size: from the resulting database of simdels, a subset was chosen to match the sequence length distribution of the observed set (Figure 6) exactly the same way as described previously (Volfovsky et al.,, 2009). Local structures and overlapping annotated features of simdel regions were determined using exactly the same procedures as with the original Indels. Finally, in order to limit computational time, I parsed this simulated data set to assign each of the valid Indel a simdel of the same size and type of local structure.

Interspecific Alignments

Aligned fragments from all the six species in this study (modern and Denisovan humans, chimpanzee, gorilla, orangutan and rhesus macaque) were trimmed to coding sequence (from AUG to STOP codons) using the *http://insilico.ehu.es/translate/*. As a result, the genes annotated on these databases represent unique Open Reading Frames (ORFs) for the all sequences. Some of the rebuilt CDSs for the Denisovan presented shattered ORFs possibly due to the incomplete or uneven coverage (Meyer et al. 2012), and were discarded. Nucleotide sequences were converted into amino acids using self coded Python (www.python.org) software with BioPython (Cock et al. 2009) dependencies. The Python script then aligned the resulting polypeptide sequences (or indirectly the codons) calling for the MUSCLE (Edgar 2004) software. Then, the aligned polypeptides were reverse-translated into nucleotide sequences using the amino acids alignments and the original (not aligned) CDS by Pal2Nal (Suyama et al. 2006). The resulting alignments were parsed to observe consistency between the gaps observed and the list of valid Indels (observed after PCR). Only 22 genes were

matched between the *in silico* sequence alignment database and the validated CDS Indels by PCR amplifications (Table 1). Consequently, the original database of 152 Indels was classified into three categories (1) null/not-validated – where no evidence of Indel discovered by reference genome alignment (*in silico*) was found by PCR amplification (*in vitro*), (2) Indel/not-validated – where there the evidence for Indel did not match between *in-vitro* and *in-silico,* and (3) Indel/validated where the presence of the same Indel *in-silico* was validated *in-vitro*, In this study, I proceeded with the analysis of the 22 Indels in the latter category (Indel/validated). All the resulting alignments of the entire CDS in all species showing Indelsca were saved in the PHYLIP interleaved format (available online from *genomes.uprm.edu/Indels*).

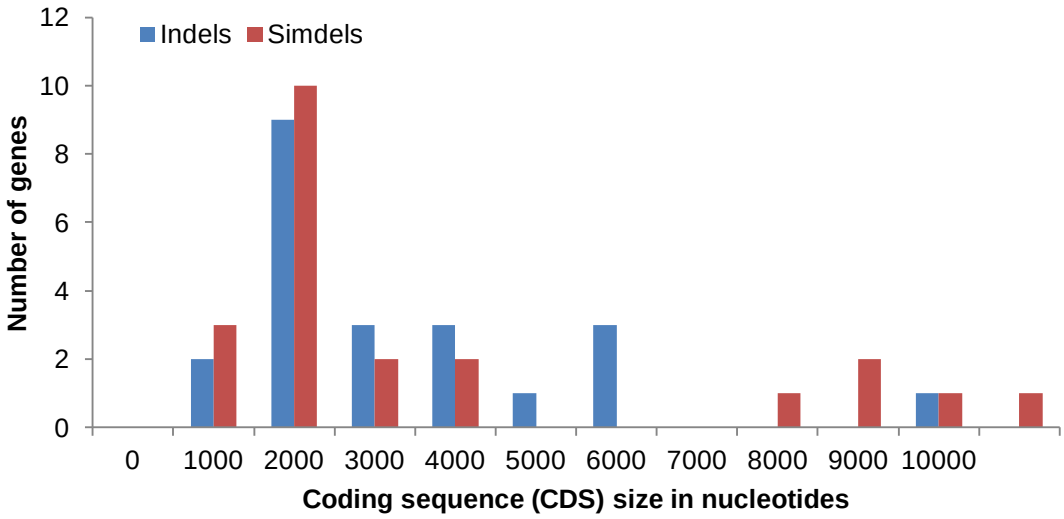

Figure 6. Size of the coding sequences containing the Indels and simdels

Selection test: changes in altering mutation rate (Ka/Ks)

The Ka/Ks ratio was calculated using the *yn00* program from the *PAML* package with the following parameters: verbose = 0, icode = 0, weighting = 0, commonf3x4 = 0, ndata = 1 (Yang 1997). This program produces pairwise comparison of the submitted sequences and calculates the Ka/Ks ratios for each with 5 different corrections of the substitution bias.

Diversity within and between human populations

To study the effect of evolution at the population level, I focused on *Homo sapiens*, currently the only species with enough publicly available data on several distinct populations to perform intra-specific tests. A large database featuring 52 distinct human populations (Human Genome Diversity Panel (HGDP)) is hosted by a Stanford University dedicated website (*http://hagsc.org/hgdp/*) and is extensively used for population-wide research on human diversity (Cavalli-Sforza 2005). I designed a custom Python script able to calculate the homozygosity and F_{ST} values of each variable locus (Single Nucleotide Polymorphism or SNP) from HGDP for two populations using an approach previously published (Oleksyk et al.,, 2008) and used successfully on a population comparison to find signatures of selection (Zhao et al.,, 2012). This approach uses a genome-wide comparison across populations using multilocus heterozygosity and F_{ST} variance in variable windows, and compares it to the randomly resampled windows representing a genome-wide reference. In this study, I ran a custom version of this program in six different comparisons selected based on the history of human migrations (Figure 5; (Henn et al. 2012): (i) Africa versus Middle East, (ii) Middle East versus Europe, (iii) Middle East versus South

Asia, (iv) South Asia versus Oceania, (v) South Asia versus Eastern Asia, (vi) Eastern Asia versus Americas. To compensate for the differences in sample sizes only 15 individuals were selected from each population randomly. Markers located on the mitochondrial DNA (mtDNA) and Y-chromosomes were removed as homozygosity calculations cannot be calculated for haploid chromosomes.

Variations of Heterozigosity and F_{ST}

The HGDP database covers the entire human genome, with a total of 660,756 SNPs (excluding mtDNA). The custom PYTHON script designed for this study samples 30 sliding windows of incremental sizes (5, 7, 9, etc. until 63 SNPs) along each chromosome and the SNPs in the center of the windows is assigned the average values of homozygosities (for population 1 and 2) and F_{ST} for the entire window. Thus, in the end, each SNPs was assigned 30 values for homozygosity for each of the two population and 30 values of F_{ST}, for each population comparison for each of the 30 window sizes. This data can be accessed online at *www.genomes.uprm.edu/Indels*.

In addition to the sequentially moving windows, and in accordance with the approach described previously (Oleksyk 2008; Zhao 2012), the PYTHON script also produced 30 distributions with windows of the same size as before, but filled by 1,000,000 SNPs picked by chance and with replacement at random location across the chromosome. From now on I will refer to these as the resampled distributions. . Using the resampled datasets, I built reference distribution and used them to calculate the percentiles, in order to evaluate homozygosity for each population and F_{ST}, for each population comparison for each moving window for each of the 30

window sizes. The program then stored the maximum value among the 30 assigned percentiles for each central SNP in each window, (highest homozygosity and F_{ST} variance (S^2F_{ST}), Figure 7). S^2F_{ST} is used to compensate for averaging of alternatively fixed alleles following an earlier described approach (Oleksyk 2008). Finally, the maximum percentiles for each SNP were plotted along the chromosomal positions for the region containing the gene of interest. Positions of SNPs showing the highest values of homozygosities and S^2F_{ST} were retained for each gene containing the Indel of interest and plotted (Figure 11).

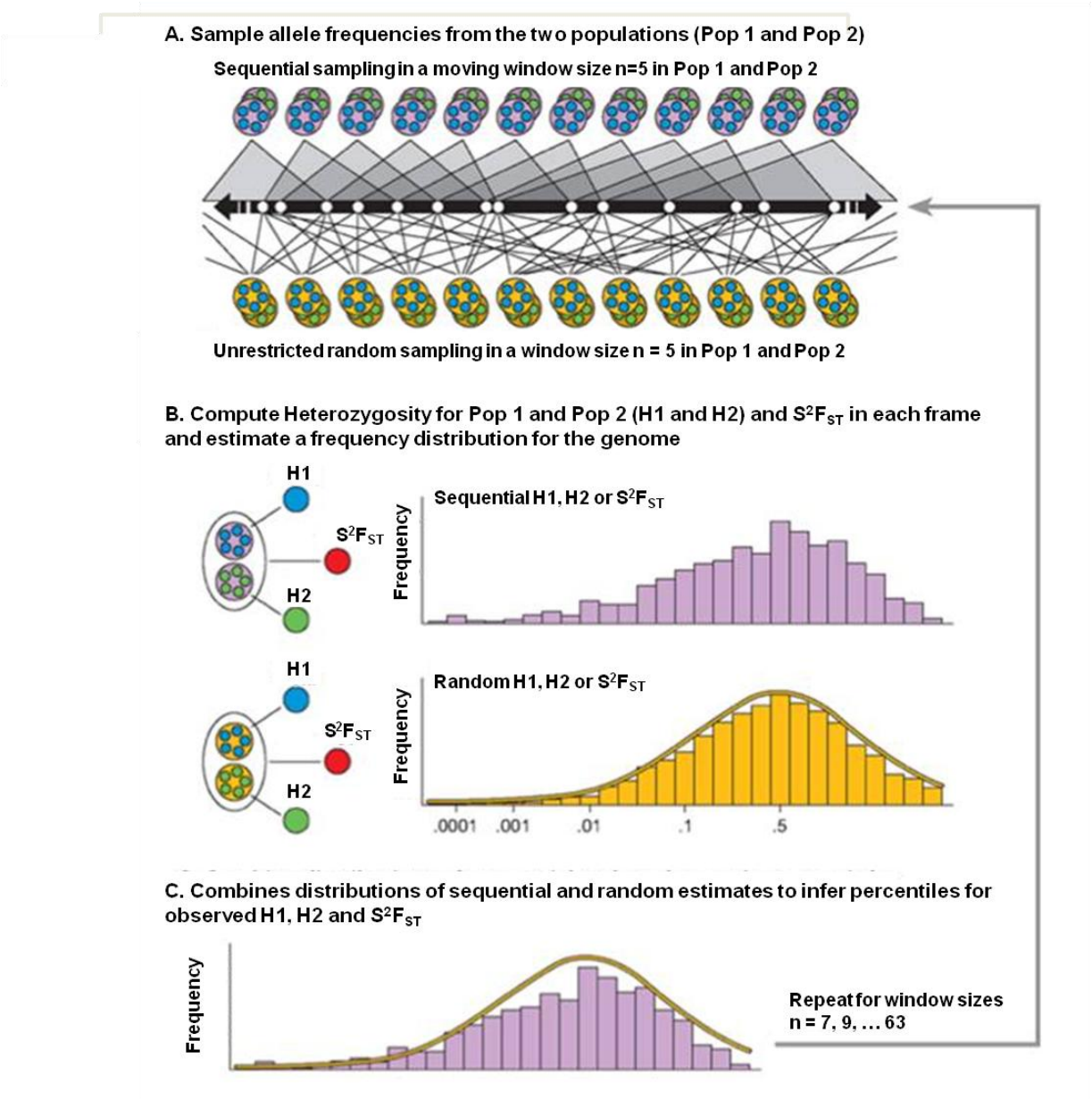

Figure 7. A flow chart for analyzing regions for local heterozygosity in both populations along with the variance of F_{ST} (S^2F_{ST}) to infer the most extreme percentile value for each SNP. (Figure modified from Oleksyk et al.,, 2008). (A) The windows are filled sequentially for the observed values and randomly for the neutral expectations. Real windows are done once for each position and random ones a total 1,000,000 times. (B) Heterozygosities and F_{ST} values are calculated for each position. (C) Real values from the windows at a certain position are compared to the 1,000,000 random values from the same windows sizes to obtain a percentile value.

Results

Part 1. Ratios of Synonymous and Non-Synonymous substitutions (Ka/Ks)

Table 1. Results of Indel validation in five primate species.

Species	Total Number of Indels	Status		
		Validated	Proved Artifacts	No evidence
Pan troglodytes	66	31	9	26
Gorilla gorilla	4	-	2	2
Pongo abelii	6	6	-	-
Macaca mulatta	70	43	7	10

(i) Distributions of Ka/Ks values for Indels and simdels

The overall distribution of Ka/Ks ratios in genes containing a real Indel and the matched simulated Indels (simdels) shows significant differences in the variance distribution among groups (d.f.=18, F=4.2, $p<0.001$, ANOVA, GLM, SAS 9.1 (2011)). These differences can be attributed to the overall difference between Indels and simdels, as seen in the figure below (Figure 8).

(ii) Ka/Ks values in pairwise comparisons

In addition, each of the groups also contain significant variance differences (Indels (d.f.=17. F=3.94, p<0.001); simdels (d.f.=17, F=2.13, p=0.0072)). These can be attributed primarily to the differences between the species comparisons (Figure 8). In addition, there is a visible trend for the Indel values to be higher than the simdel Ka/Ks values for all observed Indel comparisons except Human to Denisovan (HuDe), where the trend is reversed (Table 2, Figure 9). This may be due to the lack of statistical power, as only 3 of out of the 22 Indel-containing genes in the recently published Denisovan were successfully aligned to the human reference genome. There was a significant variation between the pairs, with the group including Denisovans (De)* showing values closest to 1 (neutrality), while other groups demonstrating values approximately equal or less than 0.5 (indicating purifying selection (Nickel et al. 2008)). The apparent difference can be in part contributed to the discovery bias, and partly to the small numbers of the Denisovan Indels.

Overall, Indel comparisons had higher values of Ka/Ks (closer to 1) than the randomly selected simdels, indicating a move away from purifying selection to selective neutrality, or that Indels occurred in less constrained regions. However, the proportion of the comparisons between genes with Ka/Ks above 1 possibly indicates the presence of positive selection (Figure 8).

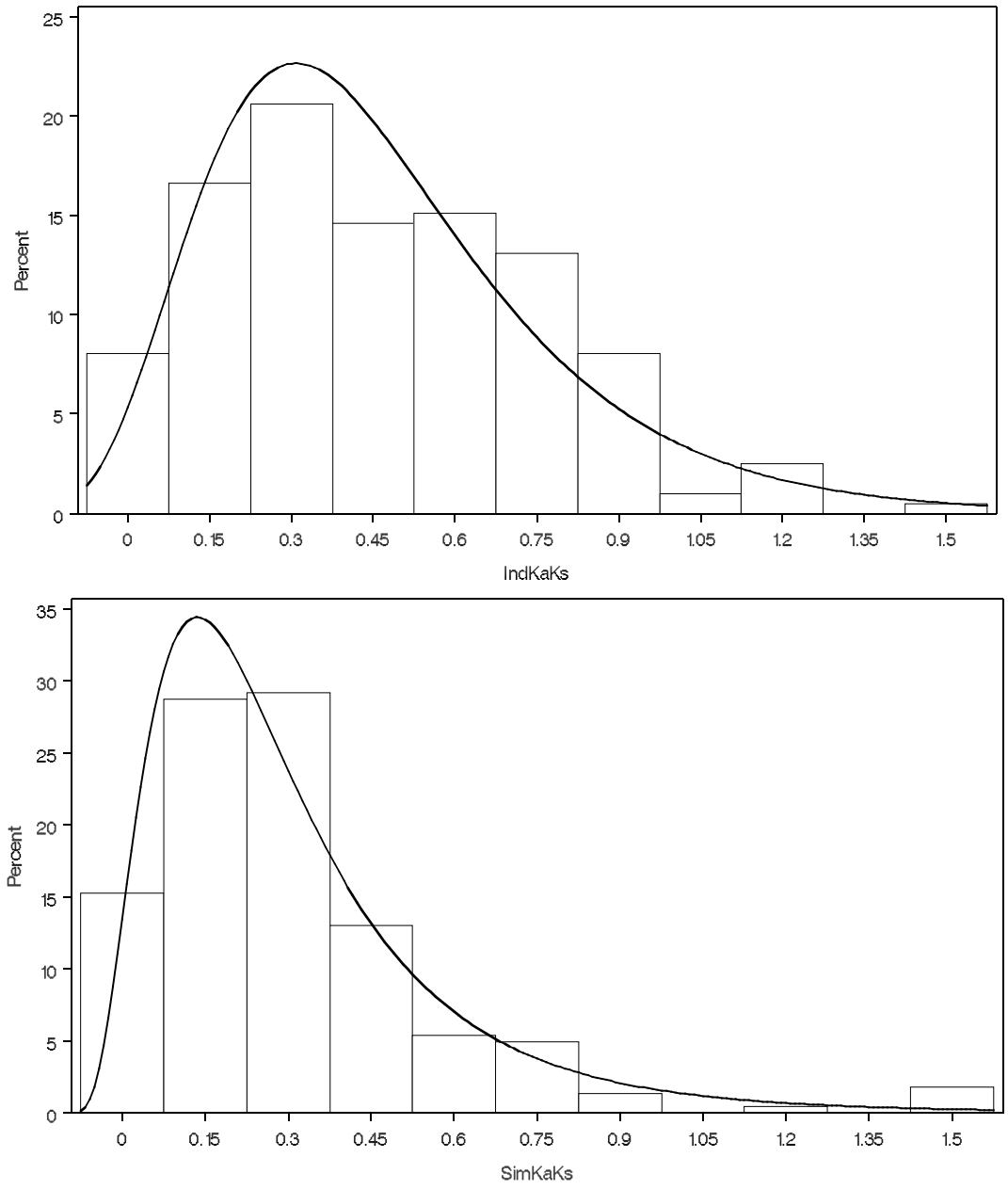

Figure 8. Ka/Ks values in the pairwise comparisons between 22 Indel (top) containing genes is higher than in simdel containing gene comparisons (bottom) between five primate and 2 human species. There were 199 Indel comparisons with the average Ka/Ks value 0.46, and 223 comparisons for simdel-containing genes with an average of 0.31 ($p<0.001$ after the Bonferroni correction, see the ANOVA analysis above). Ka/Ks values equal to one indicate selective neutrality, below 1 indicate purifying selection, while values >1 indicate positive (Darwinian) selection (Nickel et al. 2008).

Table 2. Ka/Ks values in the comparisons between species. Real Indel containing gene comparisons are contrasted with the comparisons with the simulated genes in the same species pair.

	Indel Ka/Ks			Simdel Ka/Ks		
Comparison*	**N**	**Mean**	**St. Dev.**	**N**	**Mean**	**St. Dev.**
Ch-Go	18	0.423	0.212	9	0.381	0.418
Ch-Ma	19	0.415	0.240	7	0.384	0.273
Ch-Or	12	0.449	0.354	8	0.245	0.138
De-Ch	7	0.682	0.238	17	0.259	0.198
De-Go	7	0.616	0.341	19	0.284	0.201
De-Ma	8	0.522	0.203	8	0.273	0.186
De-Or	5	0.707	0.338	18	0.342	0.252
Go-Ma	20	0.427	0.238	22	0.299	0.299
Go-Or	13	0.480	0.428	20	0.262	0.199
Hu-Ch	18	0.499	0.298	19	0.366	0.352
Hu-De	3	0.524	0.455	4	0.961	0.525
Hu-Go	20	0.437	0.354	20	0.251	0.160
Hu-Ma	21	0.434	0.276	17	0.293	0.210
Hu-Or	14	0.458	0.368	17	0.264	0.215
Or-Ma	14	0.348	0.229	18	0.290	0.251

* Abbreviations are for *H. sapiens* (Hu), *H. denisova* (De), *P. troglodytes* (Ch), *G. gorilla* (Go), *P. abelii* (Or), and *M. mulatta* (Ma).

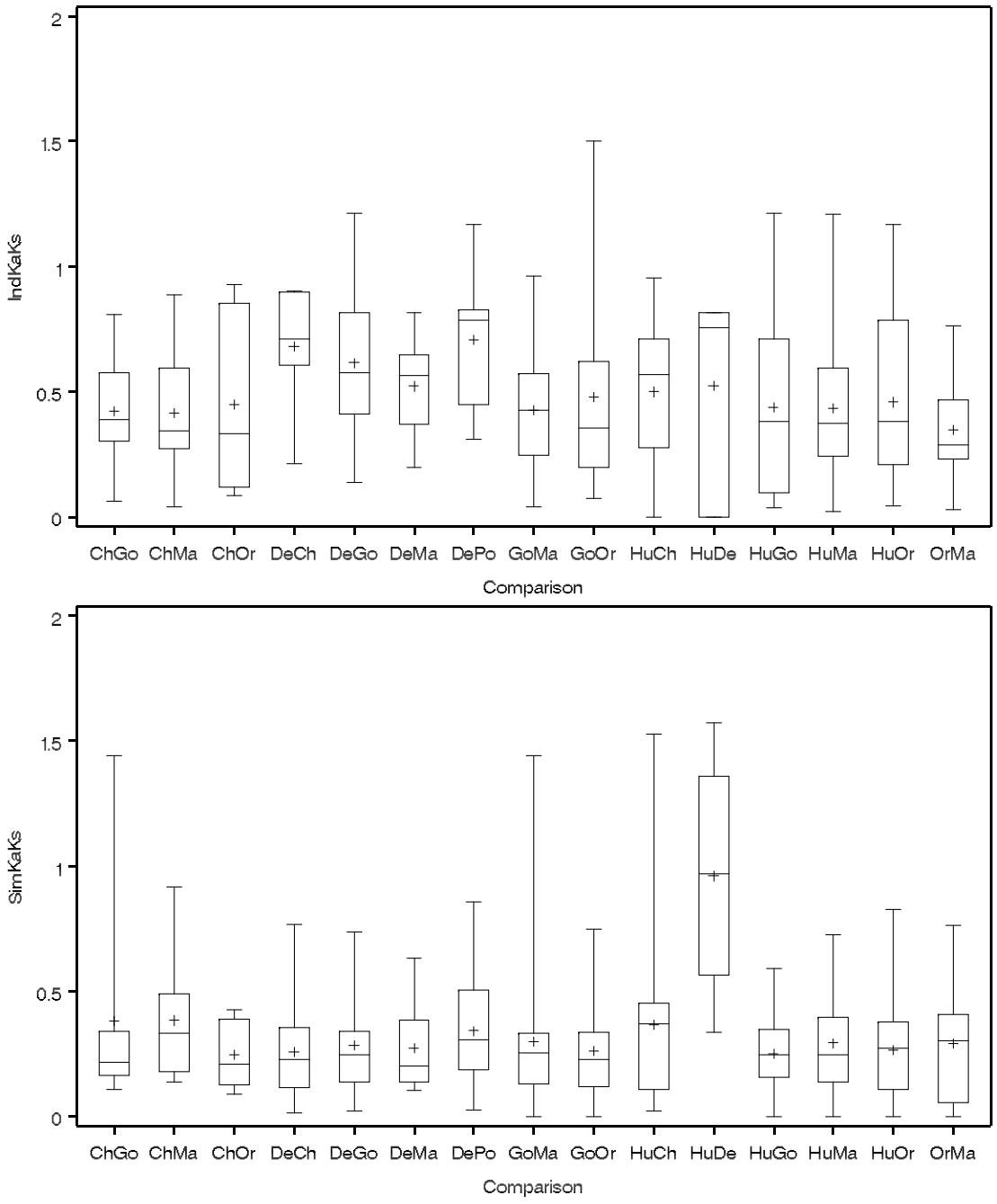

Figure 9. The Ka/Ks ratios in the pairwise species comparisons for Indels (top) and simdels (bottom). There is a significant variation between the pairs, with the group including Denisovans (De)* showing values closest to 1 (neutrality). Overall, Indel comparisons had higher values of Ka/Ks (closer to 1) than the randomly selected simdels.

* For the rest of abbreviations, please see Table 2.

(iii) Ka/Ks ratios and Indel position

Ka/Ks ratios were contrasted between those located on the first exon to other locations in the coding sequence (middle or the last exon). There were significant differences between the Ka/Ks values in the first exons in Indels vs. simdel comparisons (Table 3), while middle or final exons do not show such differences with the expected values. In addition, while Indels in first exons show a significant difference in the comparisons to the other classes of Indels, the simdel comparisons do not show such a difference (Table 4, Figure 10).

Table 3. Differences between Indel and simdel Ka/Ks ratios in the different part of the coding sequence: first, middle or last exons.

Exon #	**Indels**			**Simdels**			
	N	**Mean**	**St.Dev**	**N**	**Mean**	**St.Dev.**	*p*
First	53	**0.703**	0.281	67	**0.279**	0.261	**>0.0001**
Middle	115	0.377	0.277	139	0.324	0.263	*ns*
Last	31	0.372	0.150	17	0.282	0.285	*ns*

Table 4. Differences in Ka/Ks values between classes of Indels by position with a gene (first, middle, last), separately for Indels and simdels as shown in the figure (next page). The table shows the difference and the confidence limits (CL) of each comparison.

Class	**Comparison**	**Lower CL**	**Difference**	**Upper CL**	**Significance**
Indels	**First vs. Middle**	0.218	0.325	0.432	***
	First vs. Last	0.185	0.330	0.476	***
	Middle vs. First	-0.432	-0.325	-0.218	***
	Middle vs. Last	-0.125	0.005	0.136	*ns*
	Last vs. First	-0.476	-0.330	-0.185	***
	Last vs. Middle	-0.136	-0.005	0.125	*ns*
Simdels	**First vs. Middle**	-0.114	0.043	0.200	*ns*
	First vs. Last	-0.045	0.045	0.136	*ns*
	Middle vs. First	-0.200	-0.043	0.114	*ns*
	Middle vs. Last	-0.163	0.003	0.169	*ns*
	Last vs. First	-0.136	-0.045	0.045	*ns*
	Last vs. Middle	-0.169	-0.003	0.163	*ns*

*** ***p*<0.0001**

While there was no significant difference between the numbers of Indels and simdels expected and observed in the different classes of exons, a trend may be present with less Indels observed in the first exons (d.f.=2, X^2=4.36, p=0.11, Table 5).

Table 5. Numbers of Indels and simdels expected and observed Ka/Ks ratios in different parts of coding sequence: first, middle and last exons.

		Indel		simdel	
Exon #	**Expected**	**Observed**	**%**	**Observed**	**%**
1. First	82.5	75	22.7	90	27.3
2. Middle	210	210	63.6	210	63.6
3. Last	45	75	13.6	30	9.1

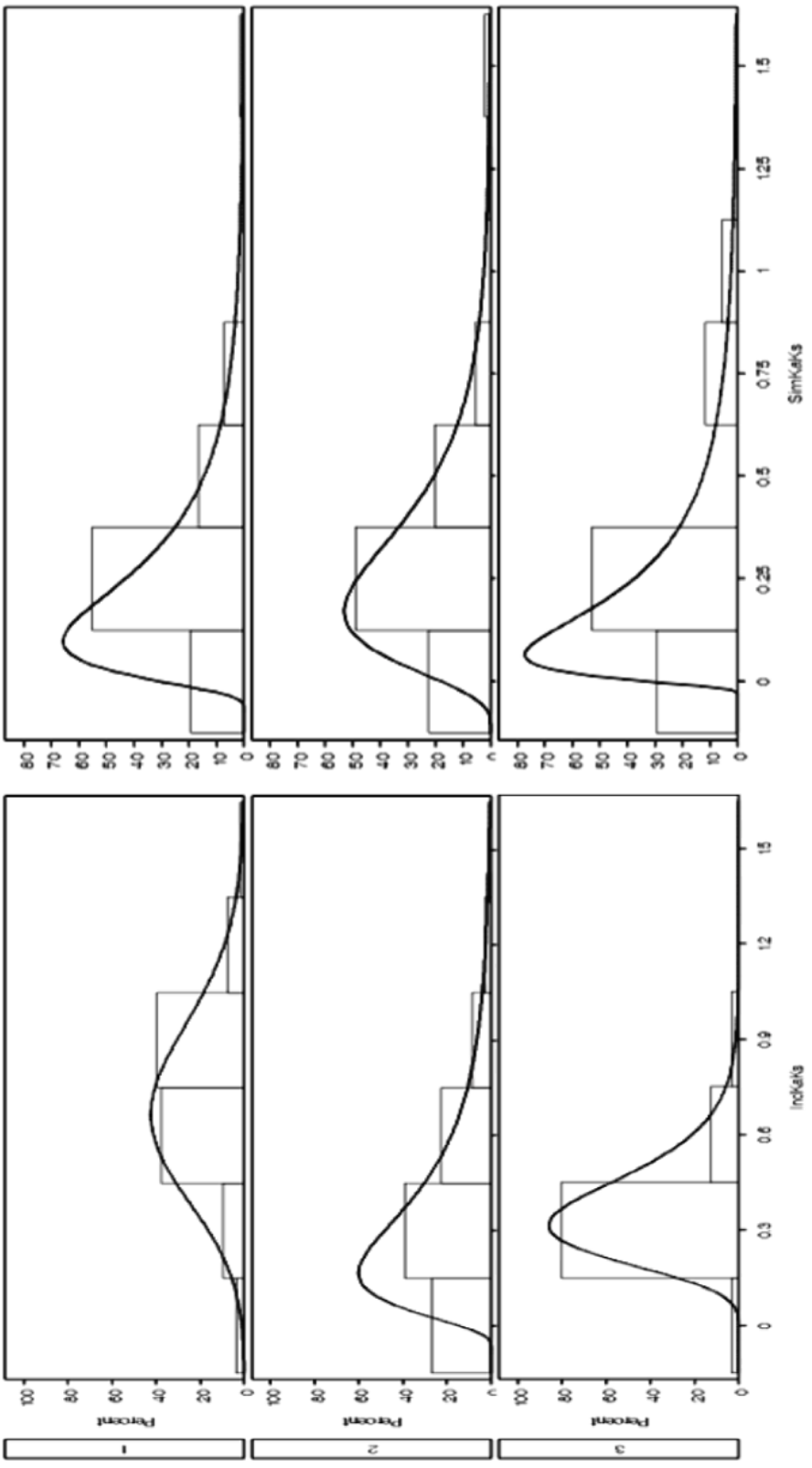

Figure 10. Distribution of Ka/Ks ratios in Indel (left) and simdel (right) comparisons in first (top), middle (middle) and last (bottom) exons of the Indel (simdel) containing genes. Indels in first exons (top left) have rations much closer to 1 (selective neutrality) than expected (top right). Other locations (middle and last exons) do not show this trend.

Part 2. Variations of Heterozygosity and F_{ST}

(i) Quantitative Observations

The Indel-containing regions were compared between human populations from HGDP (Cavalli-Sforza 2005) pairwise along the major routes of human migrations (Figure 5). Overall in the population comparisons, genes containing Indels show more extereme values, both for homozygosity and S^2F_{ST}, than the ones containing simdels (14 for Indels, six for simdels, Figure 11). Additionally, the Indels showed more extreme values across all the comparisons, whereas simdels show more of them in the last ones (including the Americas and Oceania) which are more likely to be affected by genetic drift founder effects. In addition, it seems that the extreme values show redundancy across the comparisons, indicating that some genes are good candidate to contain signatures of positive selection. The most often found outliers were located in and around *CENPN*, *CELSR1* and *GJA8* (Figure 11, Table 6). We did not include the two genes (one with Indel, one with simdel) located on chromosome X from the figures as their values of homozygosity and S^2F_{ST} were showing them to be outliers in all comparisons, appearing to be a artifact that is likely due to the reduced rate of recombination in this chromosome.

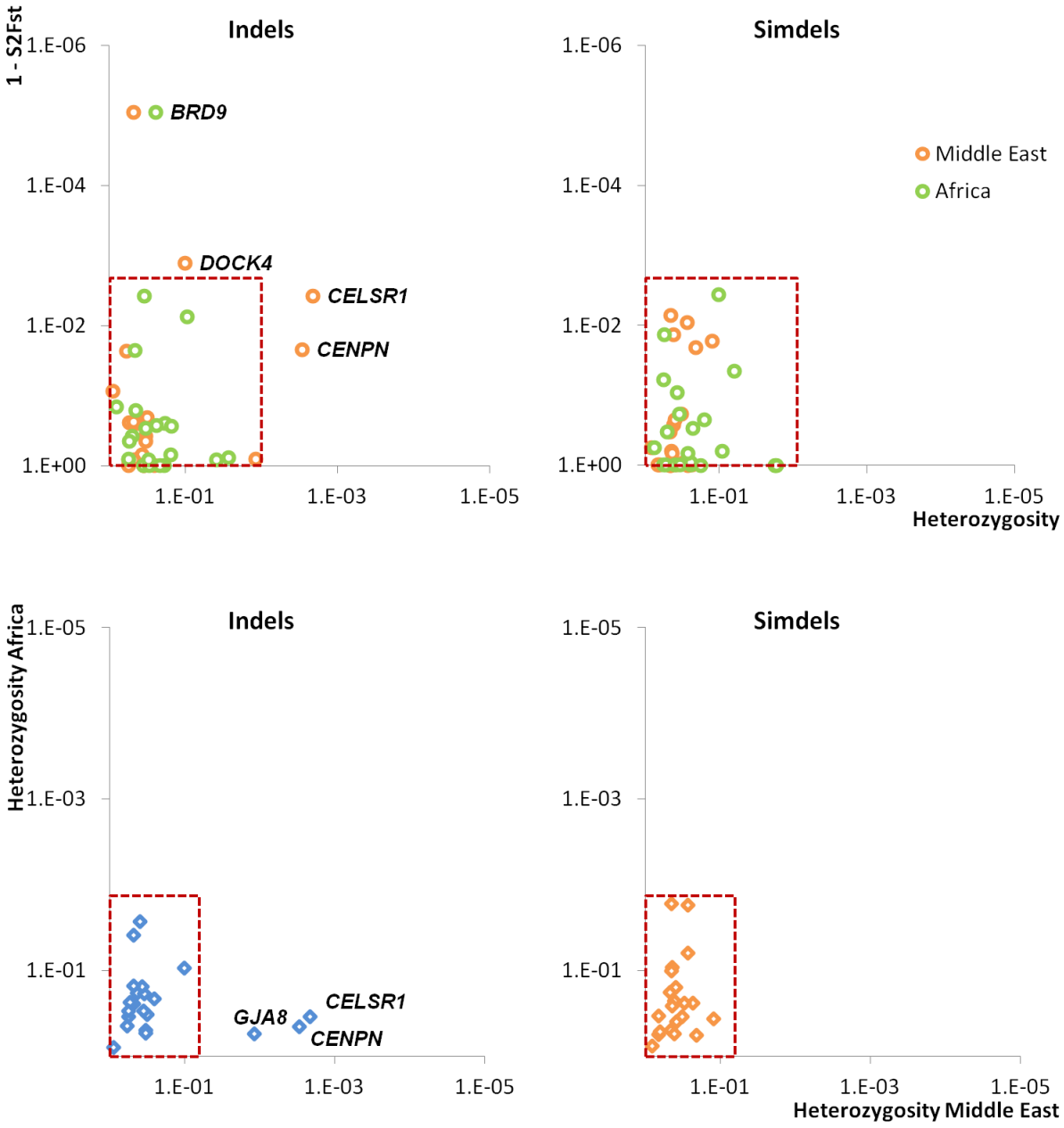

Figure 11.A Comparison of the F_{ST} variance (S^2F_{ST}) and Heterozygosity percentiles found in the chromosomal regions containing Indels and simdels (calculated using resampling approach described in Oleksyk et al.,, 2008). (Top left) S^2F_{ST} vs Heterozygocity percentiles in Indels for African vs Middle Eastern popuation comparison. (Top right) S^2F_{ST} vs Heterozygocity in simdels. (Bottom left) Heterozygocity percentiles in Indels for African vs Middle Eastern population comparison. (Bottom right) Heterozygocity percentiles in simdels. Red box indicates a trend from the simdels distribution, but is not based on calculations.

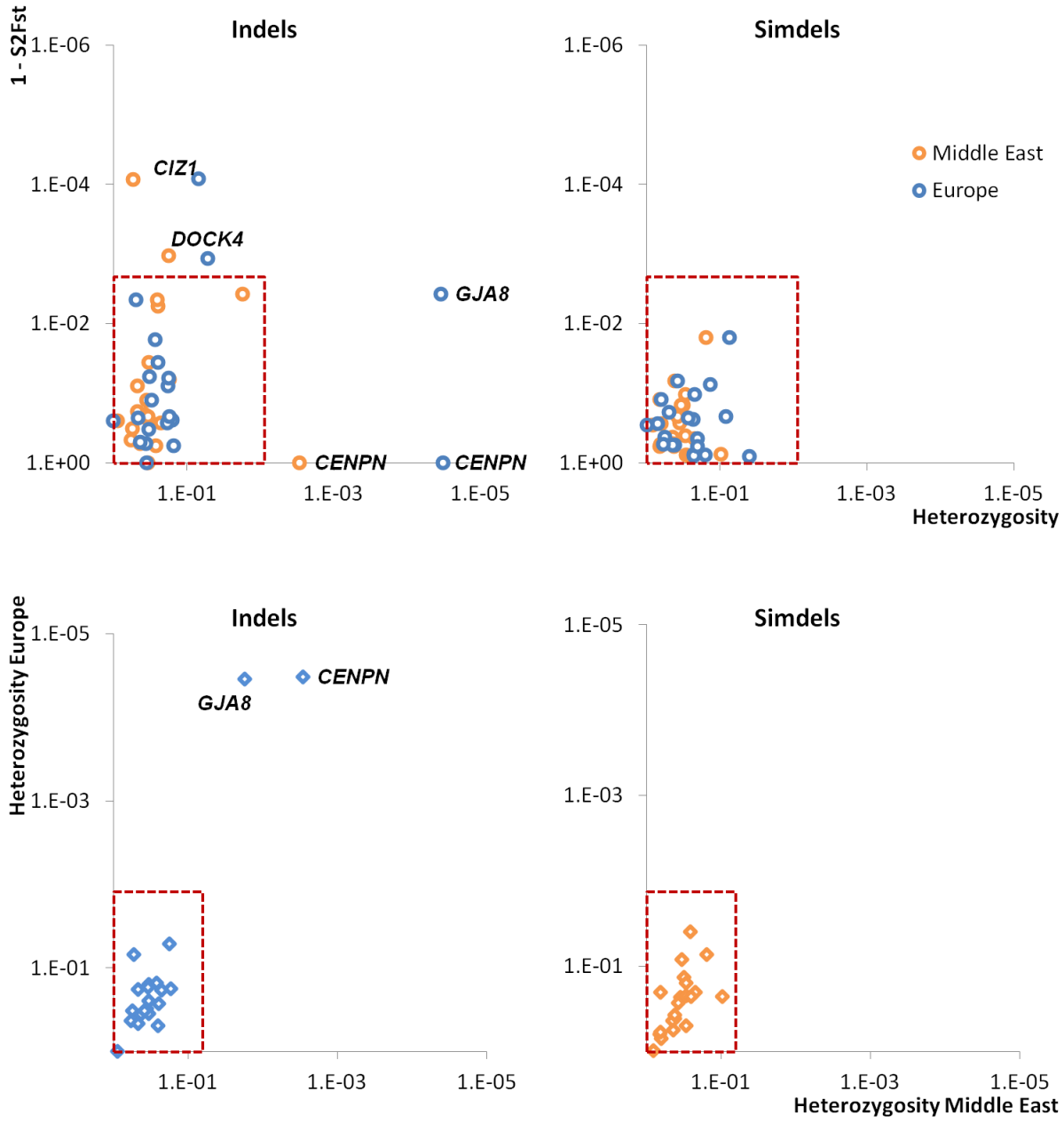

Figure 11.B Comparison of the F_{ST} variance (S^2F_{ST}) and Heterozygosity percentiles found in the chromosomal regions containing Indels and simdels (calculated using resampling approach described in Oleksyk et al.,, 2008). (Top left) S^2F_{ST} vs Heterozygocity percentiles in Indels for Middle East and Europe popuation comparison. (Top right) S^2F_{ST} vs Heterozygocity in simdels. (Bottom left) Heterozygocity percentiles in Indels for Middle East and Europe population comparison. (Bottom right) Heterozygocity percentiles in simdels. Red box indicates a trend from the simdels distribution but is not based on calculations.

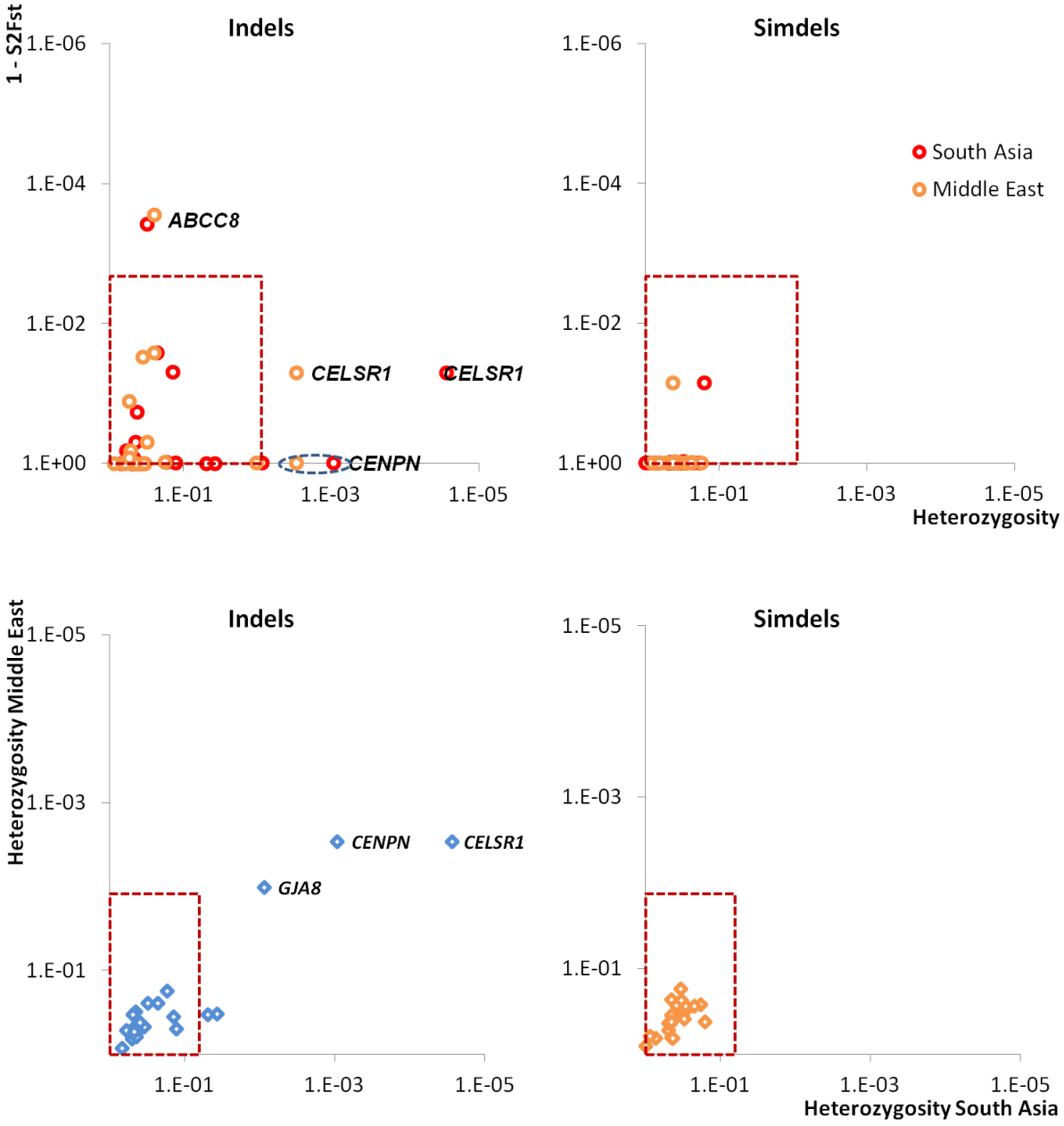

Figure 11.C Comparison of the F_{ST} variance (S^2F_{ST}) and Heterozygosity percentiles found in the chromosomal regions containing Indels and simdels (calculated using resampling approach described in Oleksyk et al.,, 2008). (Top left) S^2F_{ST} vs Heterozygocity percentiles in Indels for Middle East and Asia popuation comparison. (Top right) S^2F_{ST} vs Heterozygocity in simdels. (Bottom left) Heterozygocity percentiles in Indels for Middle East and Asia population comparison. (Bottom right) Heterozygocity percentiles in simdels. Red box indicates a trend from the simdels distribution but is not based on calculations.

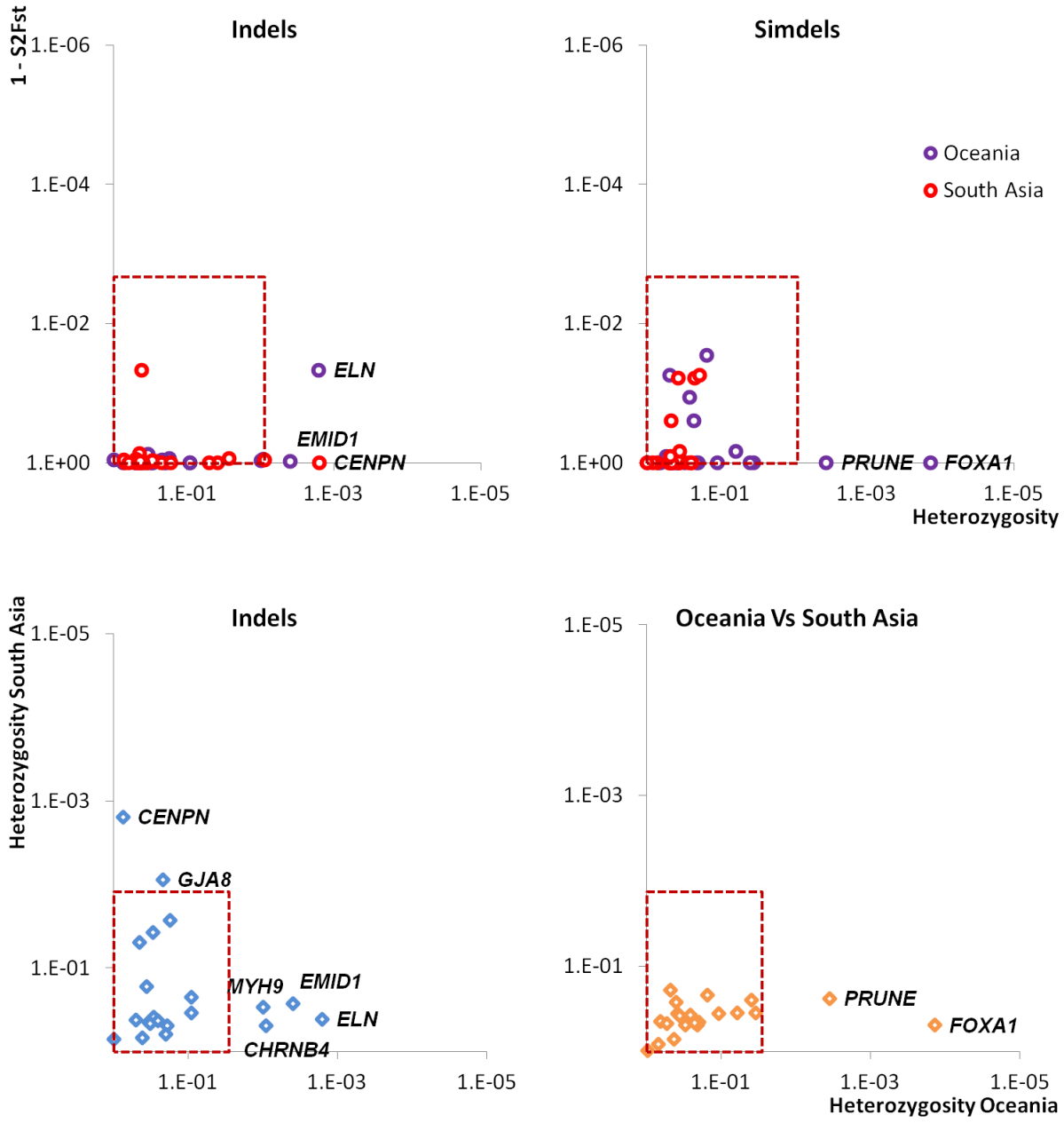

Figure 11.D Comparison of the F_{ST} variance (S^2F_{ST}) and Heterozygosity percentiles found in the chromosomal regions containing Indels and simdels (calculated using resampling approach described in Oleksyk et al.,, 2008). (Top left) S^2F_{ST} vs Heterozygocity percentiles in Indels for South Asia and Oceania popuation comparison. (Top right) S^2F_{ST} vs Heterozygocity in simdels. (Bottom left) Heterozygocity percentiles in Indels for South Asia and Oceania popu;ation comparison. (Bottom right) Heterozygocity percentiles in simdels. Red box indicates a trend from the simdels distribution but is not based on calculations.

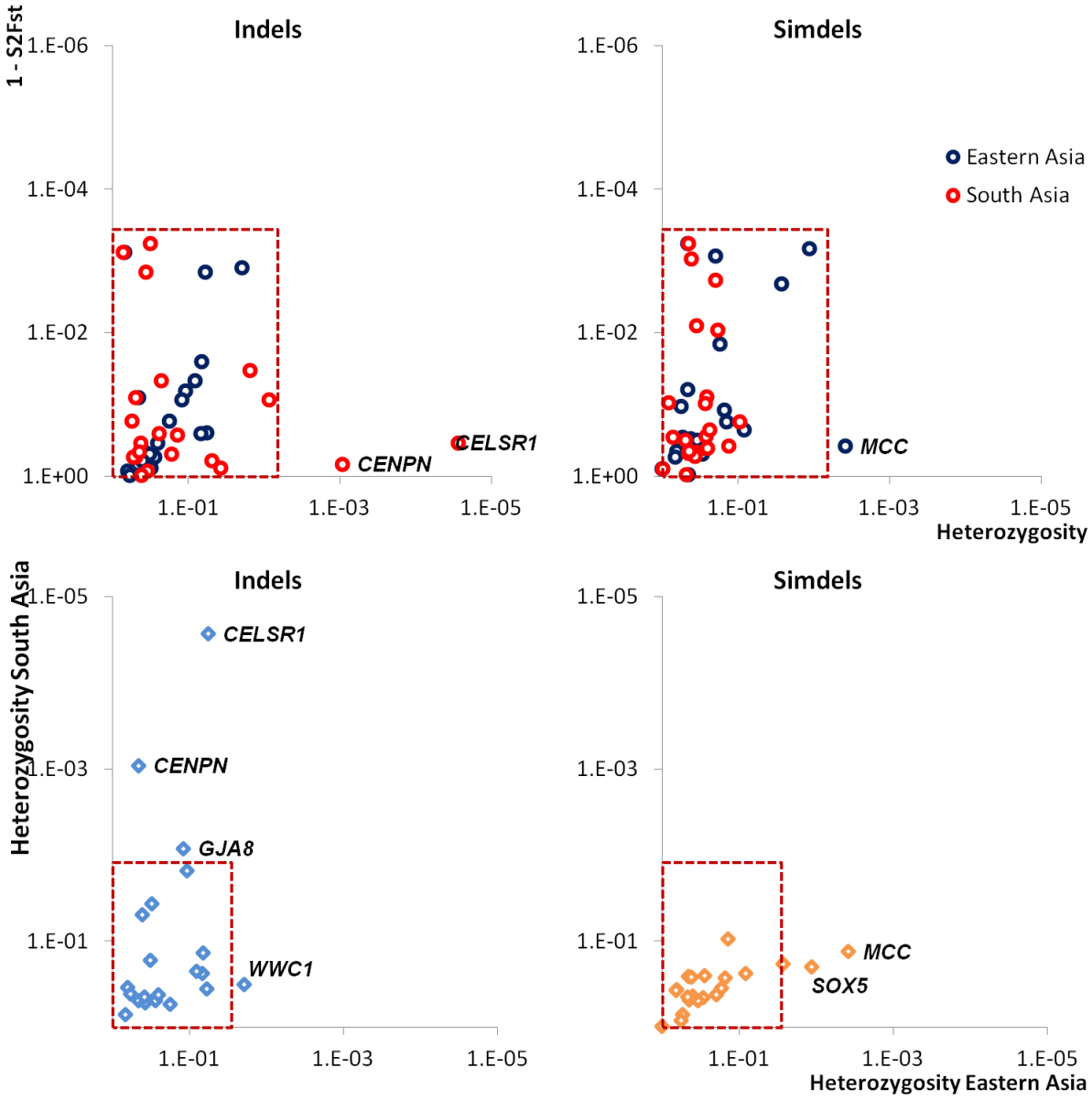

Figure 11.E Comparison of the F_{ST} variance (S^2F_{ST}) and Heterozygosity percentiles found in the chromosomal regions containing Indels and simdels (calculated using resampling approach described in Oleksyk et al.,, 2008). (Top left) S^2F_{ST} vs Heterozygocity percentiles in Indels for South Asia and East Asia popuation comparison. (Top right) S^2F_{ST} vs Heterozygocity in simdels. (Bottom left) Heterozygocity percentiles in Indels for South Asia and East Asia population comparison. (Bottom right) Heterozygocity percentiles in simdels. Red box indicates a trend from the simdels distribution but is not based on calculations.

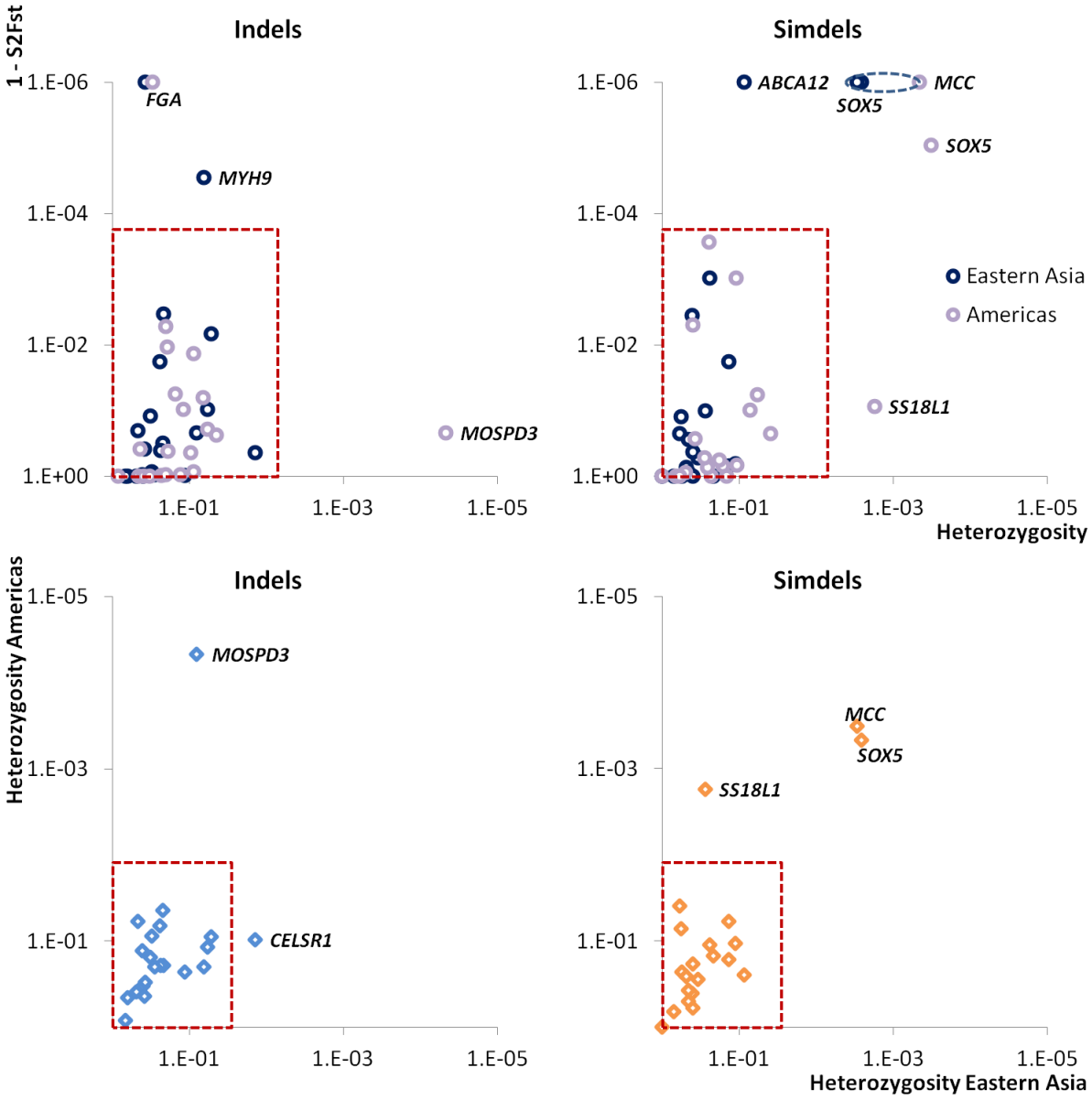

Figure 11.F Comparison of the F_{ST} variance (S^2F_{ST}) and Heterozygosity percentiles found in the chromosomal regions containing Indels and simdels (calculated using resampling approach described in Oleksyk et al.,, 2008). (Top left) S^2F_{ST} vs Heterozygocity percentiles in Indels for East Asia the Americas popuation comparison. (Top right) S^2F_{ST} vs Heterozygocity in simdels. (Bottom left) Heterozygocity percentiles in Indels for East Asia and the Americas population comparison. (Bottom right) Heterozygocity percentiles in simdels. Red box indicates a trend from the simdels distribution but is not based on calculations.

(ii) Qualitative observations

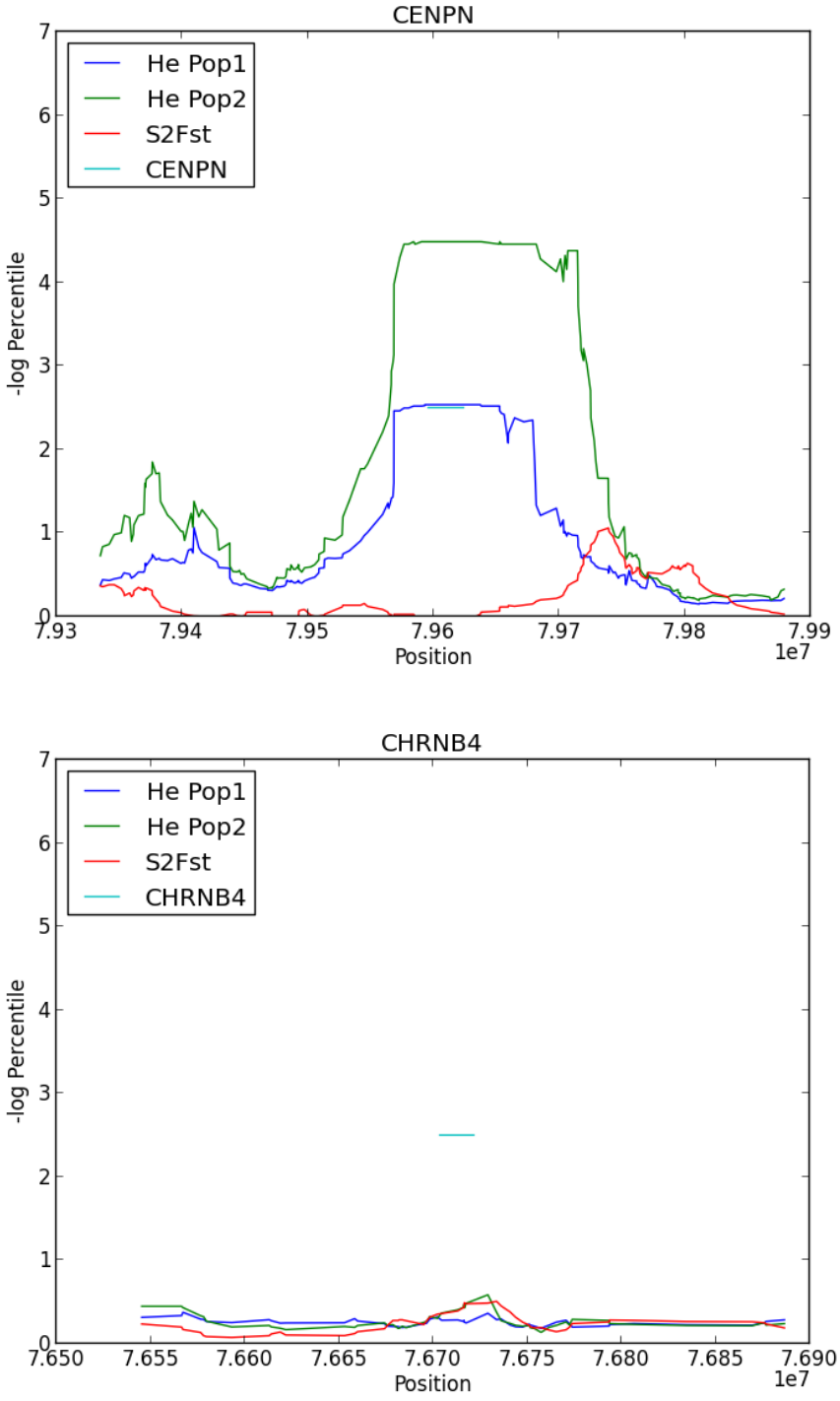

Figure 12. Examples of the output from the PYTHON script for signatures of selection based on Oleksyk et al.,, 2008. (Top) An example of a visually predicted selective sweep (here on gene *CENPN*). To the contrary, (Bottom) gene *CHRNB4* did not show any sign of the extreme values relative to selective sweeps.

Table 6. Qualitative observations of sweeps in Intra and Interspecific comparisons.

Indel Containing Genes	Intraspecific	Interspecific
ABCC8	x	x
AUNIP	x	Selection in Homo/Gorilla, Homo/Pongo and Gorilla/Pongo
BRD9	x	x
CELSR1	Selected since MiddleEast(recent) until Americas (all comparisons - Ancestral)	x
CENPN	Ancestral selection in MiddleEastVsEuropean and MiddleEastVsSouthAsia	x
CHCHD1	x	x
CHRNB4	x	x
CIZ1	x	x
DOCK4	Selected since MiddleEast(recent) until Americas (all comparisons - Ancestral)	x
ELN	x	x
EMID1	x	x
FGA	x	x
GJA8	missing SNPs resolution around	x
LTBP2	x	x
MOSPD3	x	x
MYH9	x	x
PPP1R3A	x	x
RGL4	x	Selection in Homo/Pongo, Homo/Macaca and Gorilla/Pongo
SEPN1	x	x
SLC43A1	x	x
TBC1D8B	Selected since Africa (recent) until Americas (all comparisons - Ancestral) - Chr X	x
WWC1	x	x

Table 6 (continued). Qualitative observations of sweeps in Intra and Interspecific comparisons.

Simdel Containing Genes	Intraspecific	Interspecific
A4GNT	x	x
ABCA12	x	x
ACSM4	x	x
AFTPH	x	Selection in Homo/Pongo
AGTRAP	x	x
ANKRD24	x	x
CTBP1	x	x
DEK	x	x
FAM184B	x	x
FOXA1	x	x
FRY	x	x
HHLA3	x	Selection in Homo/Pan
MCC	Ancestral selection in EasternAsiaVsAmericas	x
MXRA5	Selected before Africa (Ancestral) until Americas (all comparisons minus EasternAsiaVsAmericas - Ancestral) - Chr X	Selection in Homo/Denisovan
NDUFS2	x	x
PLD4	x	x
PRUNE	x	x
SOX5	Ancestral selection in OceaniaVsSouthAsia, Recent in EasternAsiaVsSouthAsia and EasternAsiaVsAmericas	Selection in Homo/Denisovan
SS18L1	x	x
UQCRC1	Recent selection in EasternAsiaVsAmericas	x
ZC3H12D	x	x
ZNF133	x	x

This table is a resume of the sweeps visualized in the graphics like the example shown above (Intraspecific) and the Ka/Ks ratios inferring positive selection (Interspecific). In intraspecific comparisons, it is possible to infer if the selection occurred before or after the population split, as described in Oleksyk et. al, 2008.

Discussion

The purpose of this study was to validate and study the evolutionary impact of predicted high impacting (changes on protein sequence) Indels. The ones proved to be real by PCR where used in two types of comparisons, searching for footprints of natural selection: (i) interspecific, comparing the Indel containing gene to its homologous in the hominids lineage; (ii) intraspecific, comparing the Indel containing gene in different human populations. The set of validated Indels was compared to a set of simulated random Indels (simdels), which is used as null hypothesis (no particular effect related to the presence of Indels).

Indel Detection and Validation

In this study a database of insertions/deletions (Indels) from four different primate species and humans (Human/Chimpanzee, Human/Gorilla, Human/Orangutan, Human/Macaque) was examined (Tolstorukov et al., 2012). Among the 36,422 Indels, in the database, 146 Indels from this set were in coding sequences. We attempted to validate these discovered Indels using molecular techniques, and then interrogate the genes containing the validated Indels and their chromosomal neighborhoods for the signatures of selective sweeps in pairwise comparisons between two humans (modern and Denisovan) and four primates species.

In the validation process, genomics regions in the modern human and four primates, only 22 Indels unequivocally corresponded to the predicted amplicon sizes and gene alignments. Other Indels were either of the wrong size, or the sequences

were not different. A possible reason for this is the poor quality of the public reference sequences for primate species. However, it is possible that some of these Indels could exist in the reference individual, but could not be amplified in the samples we used in this study. For these reasons, not every Indel predicted by the alignment of the reference sequences can be found in vitro by PCR validation.

Ratios of Synonymous and Non-Synonymous substitutions (Ka/Ks)

Genes containing Indels in the coding sequence should have the highest impact on the protein structure, and thus are likely to contribute to the differences between the species. If this difference is advantageous, positive selection would impact the gene and its flanking sequence. On the other hand, a large change is likely to be possible because of the relaxed selective constraint. In our case, Ka/Ks values in the Indel-containing genes were found to be higher than the same values calculated from coding sequences of the same species comparison drawn at random. Specifically, the overall distribution for Indel containing was higher (closer to 1) suggesting neutral variation, but also contained more values inferring positive selection (Ka/Ks > 1). The former trend can likely be explained by the lack of selective constraint (impact on fitness) in the genes containing observed Indels: these fragments are under relaxed purifying selection. The latter group of the Indels (with Ka/Ks > 1) may be under positive selection. This is more difficult to show, as positive selection can be localized, thus the Ka/Ks ratio on the complete CDS can be higher than expected but lower than 1 due to the dilution of the signal by the surrounding purifying selection (Hurst 2009). It is likely that either or both trends are

affecting different genes in our set, explaining the observed distribution Ka/Ks (Figure 8).

The relaxation of purifying selection could be affecting different parts of the sequence to a different degree. I expected the regions with less impact on the protein function (and therefore fitness) to be more permissive to non-synonymous mutations, thus to have higher Ka/Ks. For example, regions coding for loops in protein sequence might be more permissive to non-synonymous mutations and Indels. However, this has not been proved yet. Additionally, the Ka/Ks ratios in this study were calculated for entire gene sequences, reducing potential local bias due to the local relaxation of purifying selection in regions such as loops.

It also appears that the first exons may be under more constrained selective pressures compared to the rest of the gene (Figure 9, Table 5), since the values of Ka/Ks observed in genes containing an Indel in the first exon are significantly higher. In addition, there is an indication that Indels located in the first exon of genes are underrepresented in the database. This effect should be further validated by increasing a sample size of simulated Indels (simdels), but, if true, would support the conclusion of stronger selective pressures at the first exon of the gene, as the larger Indels in this region are more likely to be eliminated by purifying selection. I did not find any reference supporting such phenomena.

Finally, the coding sequences showing a Ka/Ks higher than 1 (considering the dilution effect mentioned earlier) are strong candidates for targets of positive selection. *AUNIP* and *RGL4* are the two Indel containing genes that showed Ka/Ks ratios higher than 1. *AUNIP* produces a protein related to centrosomes, and is involved in cell cycle in heart, skeletal muscles, testis and placenta. *RGL4* has been

related to the regulation of the Ras system (controlling many aspects of eukaryotic cell homeostasis), and has been related to speciation events in eukaryote evolution (Diez et al. 2011). These genes might have had a key role in human to primate, or primate to primate, divergence and if confirmed by future studies, may help understand functional evolution of our own species.

Variations of Heterozygosity and F_{ST}

As human populations diverged, they encountered different sets of environmental conditions and diseases, which reflected in diverse spectra of selective pressures impacting genetic pools of each separate group. Selection acts to reduce genetic variation in the gene and its genome neighborhood, leaving behind a characteristic footprint in the loci linked to the advantageous trait: selective sweep (Sabeti et al.,, 2006). At the same time, local genome divergence (measured as F_{ST}) between the two separated populations will increase (Oleksyk et al.,, 2010). The challenge is to evaluate such regions against the variation in genome diversity that comes as a consequence of demographic factors or genetic drift. In this study, a comparative approach was taken to evaluate the reduction of genetic diversity (multilocus heterozygosity) and population divergence (multilocus variance of F_{ST} or S^2F_{ST}) in the flanking regions of observed Indels versus the randomly assigned locations in genes (simdels). In this analysis, comparisons between continental populations showed more genes with extreme values of homozygosity and S^2F_{ST} in the genes of the Indel set than the simdel set (Figure 11 A-F). These particular trend of values is found in all population comparisons along the human great migrations history (Figure 5, Figure 11 A-F), while the extreme values in the simdel set only

appeared after the known founder effects characterizing populations of the Americas and Oceania, and can probably be attributed to the action of genetic drift. Therefore, the present results indicate that at least some Indels that appeared during the divergence of human and primates' ancestors might have been affected by positive selection later, during the great human migrations.

Several genes can be traced along the routed of human migrations (Figure 11). Table 6 is a resume of the sweeps visualized in the graphics like the example shown above (Intraspecific) and the Ka/Ks ratios inferring positive selection (Interspecific). In intraspecific comparisons, it is possible to infer if the selection occurred before or after the population split, as described in Oleksyk et. al, 2008. For example, PCR amplicons suggest that *CELSR1* is a gene that had an insertion in the human species. Interspecific comparisons show this gene did not present any sign of ancient selection in the human to primate comparison. However this gene contains an extreme value suggesting on-going selection footprints in all populations since the split between Africa and Middle East (Figure 11 A-F). *CELSR1* encodes for a protein of the cadherin superfamily. It has been mentioned in numerous studies and related to (the following list is non-exhaustive): cell polarity (Qu et al. 2010; Tissir et al. 2013), lung morphogenesis (Yates et al. 2013), cancers (Liao et al. 2012; Kaucka et al. 2013), neural development (Zhou et al. 2007; Boutin et al. 2012) and disorders (Juriloff et al. 2012), craniofacial phenotypes (Mukhopadhyay et al. 2012), ischemic stroke (Yamada et al. 2009; Gouveia et al. 2011), etc. Similarly, *CENPN* also present an insertion in human, with no selection footprint in interspecific comparisons, but shows evidence for selective sweep between Middle East, European and Central/South Asian populations (Figure 11). *CENPN* encodes for a protein part of the nucleosome-associated complex and is important for kinetochore

assembly, thus involved in the cell cycle. It has been related (the following list is non-exhaustive) to neural disorders (Chen et al. 2013), genomic machinery (Saltzman et al. 2011) and instabilities (Reinhold et al. 2011), craniosynostotic conditions (Fanganiello et al. 2007), cancers (Liang et al.), etc. *DOCK4* seems to have suffered either an insertion in Macaque Rhesus or a deletion in the hominid lineage. Like *CELSR1*, no sign of ancient selection were found in interspecific comparisons, but selection footprints appeared since the split between Africa and Middle East (Figure 11 and Table 6). *DOCK4* encodes for a membrane-associated, cytoplasmic protein involved in regulation of adherens junctions between cells. It has been related to (the following list is non-exhaustive) cancers (Gadd et al. 2010; LaFramboise et al. 2010), dendritic development (Ueda et al. 2008; Ueda et al. 2013), and autism (Pagnamenta et al. 2010; Kalkman 2012).

It is important to state that the challenges in Indel detection and validation reduced noticeably our data sets. As a result, the observed trends might be biased. However, a more consistent list of Indels may reveal more candidates related to species or population adaptations.

Conclusions

In summary, this study brought several insights about the Indels found in the hominid lineage. First, many predicted Indels seem to be artifacts due to inconsistencies in the reference genomes. Also, to validate Indels merging molecular and informatics technologies appears to be a tedious but necessary process. Second, interspecific comparisons suggest that the appearance of Indel is manly related to low selective constraint, which could be the cause and/or the effect of the fixation of the mutation. Finally, some Indels show strong evidence of positive selection and might have had a particular effect on the divergence of the human species and the diversity in modern populations.

Recommendations

On a short term scale, I recommend to add several steps to this study. First, as the public databases might not always be accurate about orthologous genes, running a software like *Orthomcl* may be necessary to prevent bias by using wrong sequences in the interspecific alignments. To obtain a better statistical consistency, adding more Indels containing genes is necessary. Using the Indels partially validated (present different size of amplicons between at least two species) in a different category may reinforce our conclusions. Also, using more simdels (at least 3 simdels for each Indels) would reinforce the null hypothesis (Indels not having special effect on evolution of the hominids lineage). For the interspecific comparisons, adding a branch-test (*codeml* from the *PAML* package) would reinforce the approach of pairwise Ka/Ks ratios. Finally, testing our approach on population comparison on simulated selective sweeps would allow us to calculate the genetic distance between the sweep and the location of the signal detected. This would make the search for extreme values in the gene locus more accurate.

In the long term scale, validating more Indels with newer dataset would increase the resolution of our approach. Also, resequencing the flanking regions of Indels in each species would be an excellent way to validate the Indels. Advanced statistical tests on synonymous and non-synonymous substitutions (like McDonald-Kreitman and Hudson-Kreitman-Aguade tests) would allow more accurate conclusions on the effect of Indels on the hominids evolution.

References

Abecasis, G. R., A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth and G. A. McVean (2012). "An integrated map of genetic variation from 1,092 human genomes." Nature **491**(7422): 56-65.

Albers, C. A., G. Lunter, D. G. MacArthur, G. McVean, W. H. Ouwehand and R. Durbin (2011). "DIndel: accurate Indel calls from short-read data." Genome Res **21**(6): 961-73.

Altshuler, D. M., R. A. Gibbs, L. Peltonen, D. M. Altshuler, R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner, F. Yu, L. Peltonen, E. Dermitzakis, P. E. Bonnen, D. M. Altshuler, R. A. Gibbs, P. I. de Bakker, P. Deloukas, S. B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X. Jia, A. Palotie, M. Parkin, P. Whittaker, F. Yu, K. Chang, A. Hawes, L. R. Lewis, Y. Ren, D. Wheeler, R. A. Gibbs, D. M. Muzny, C. Barnes, K. Darvishi, M. Hurles, J. M. Korn, K. Kristiansson, C. Lee, S. A. McCarrol, J. Nemesh, E. Dermitzakis, A. Keinan, S. B. Montgomery, S. Pollack, A. L. Price, N. Soranzo, P. E. Bonnen, R. A. Gibbs, C. Gonzaga-Jauregui, A. Keinan, A. L. Price, F. Yu, V. Anttila, W. Brodeur, M. J. Daly, S. Leslie, G. McVean, L. Moutsianas, H. Nguyen, S. F. Schaffner, Q. Zhang, M. J. Ghori, R. McGinnis, W. McLaren, S. Pollack, A. L. Price, S. F. Schaffner, F. Takeuchi, S. R. Grossman, I. Shlyakhter, E. B. Hostetter, P. C. Sabeti, C. A. Adebamowo, M. W. Foster, D. R. Gordon, J. Licinio, M. C. Manca, P. A. Marshall, I. Matsuda, D. Ngare, V. O. Wang, D. Reddy, C. N. Rotimi, C. D. Royal, R. R. Sharp, C. Zeng, L. D. Brooks and J. E. McEwen (2010). "Integrating common and rare genetic variation in diverse human populations." Nature **467**(7311): 52-8.

Boutin, C., A. M. Goffinet and F. Tissir (2012). "Celsr1-3 cadherins in PCP and brain development." Curr Top Dev Biol **101**: 161-83.

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz, S. Glanowski, D. M. Tanenbaum, T. J. White, J. J. Sninsky, R. D. Hernandez, D. Civello, M. D. Adams, M. Cargill and A. G. Clark (2005). "Natural selection on protein-coding genes in the human genome." Nature **437**(7062): 1153-7.

Cavalli-Sforza, L. L. (2005). "The Human Genome Diversity Project: past, present and future." Nat Rev Genet **6**(4): 333-40.

Charlesworth, B. (2009). "Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation." Nat Rev Genet **10**(3): 195-205.

Chen, X., Y. Shen, Y. Gao, H. Zhao, X. Sheng, J. Zou, V. Lip, H. Xie, J. Guo, H. Shao, Y. Bao, J. Shen, B. Niu, J. F. Gusella, B. L. Wu and T. Zhang (2013). "Detection of copy number variants reveals association of cilia genes with neural tube defects." PLoS One **8**(1): e54492.

Cock, P. J., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski and M. J. de Hoon (2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics." Bioinformatics **25**(11): 1422-3.

Collins, F. S., M. Morgan and A. Patrinos (2003). "The Human Genome Project: lessons from large-scale biology." Science **300**(5617): 286-90.

Consortium, C. S. a. A. (2005). "Initial sequence of the chimpanzee genome and comparison with the human genome." Nature **437**(7055): 69-87.

Dalca, A. V. and M. Brudno (2010). "Genome variation discovery with high-throughput sequencing data." Brief Bioinform **11**(1): 3-14.

Diez, D., F. Sanchez-Jimenez and J. A. Ranea (2011). "Evolutionary expansion of the Ras switch regulatory module in eukaryotes." Nucleic Acids Res **39**(13): 5526-37.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-7.

Fanganiello, R. D., A. L. Sertie, E. M. Reis, E. Yeh, N. A. Oliveira, D. F. Bueno, I. Kerkis, N. Alonso, S. Cavalheiro, H. Matsushita, R. Freitas, S. Verjovski-Almeida and M. R. Passos-Bueno (2007). "Apert p.Ser252Trp mutation in FGFR2 alters osteogenic potential and gene expression of cranial periosteal cells." Mol Med **13**(7-8): 422-42.

Gadd, S., S. T. Sredni, C. C. Huang and E. J. Perlman (2010). "Rhabdoid tumor: gene expression clues to pathogenesis and potential therapeutic targets." Lab Invest **90**(5): 724-38.

Gibbs, R. A., J. Rogers, M. G. Katze, R. Bumgarner, G. M. Weinstock, E. R. Mardis, K. A. Remington, R. L. Strausberg, J. C. Venter, R. K. Wilson, M. A. Batzer, C. D. Bustamante, E. E. Eichler, M. W. Hahn, R. C. Hardison, K. D. Makova, W. Miller, A. Milosavljevic, R. E. Palermo, A. Siepel, J. M. Sikela, T. Attaway, S. Bell, K. E. Bernard, C. J. Buhay, M. N. Chandrabose, M. Dao, C. Davis, K. D. Delehaunty, Y. Ding, H. H. Dinh, S. Dugan-Rocha, L. A. Fulton, R. A. Gabisi, T. T. Garner, J. Godfrey, A. C. Hawes, J. Hernandez, S. Hines, M. Holder, J. Hume, S. N. Jhangiani, V. Joshi, Z. M. Khan, E. F. Kirkness, A. Cree, R. G. Fowler, S. Lee, L. R. Lewis, Z. Li, Y. S. Liu, S. M. Moore, D. Muzny, L. V. Nazareth, D. N. Ngo, G. O. Okwuonu, G. Pai, D. Parker, H. A. Paul, C. Pfannkoch, C. S. Pohl, Y. H. Rogers, S. J. Ruiz, A. Sabo, J. Santibanez, B. W. Schneider, S. M. Smith, E. Sodergren, A. F. Svatek, T. R. Utterback, S. Vattathil, W. Warren, C. S. White, A. T. Chinwalla, Y. Feng, A. L. Halpern, L. W. Hillier, X. Huang, P. Minx, J. O. Nelson, K. H. Pepin, X. Qin, G. G. Sutton, E. Venter, B. P. Walenz, J. W. Wallis, K. C. Worley, S. P. Yang, S. M. Jones, M. A. Marra, M. Rocchi, J. E. Schein, R. Baertsch, L. Clarke, M. Csuros, J. Glasscock, R. A. Harris, P. Havlak, A. R. Jackson, H. Jiang, Y. Liu, D. N. Messina, Y. Shen, H. X. Song, T. Wylie, L. Zhang, E. Birney, K. Han, M. K. Konkel, J. Lee, A. F. Smit, B. Ullmer, H. Wang, J. Xing, R. Burhans, Z. Cheng, J. E. Karro, J. Ma, B. Raney, X. She, M. J. Cox, J. P. Demuth, L. J. Dumas, S. G. Han, J. Hopkins, A. Karimpour-Fard, Y. H. Kim, J. R. Pollack, T. Vinar, C. Addo-Quaye, J. Degenhardt, A. Denby, M. J. Hubisz, A. Indap, C. Kosiol, B. T. Lahn, H. A. Lawson, A. Marklein, R. Nielsen, E. J. Vallender, A. G. Clark, B. Ferguson, R. D. Hernandez, K. Hirani, H. Kehrer-Sawatzki, J. Kolb, S. Patil, L. L. Pu, Y. Ren, D. G. Smith, D. A. Wheeler, I. Schenck, E. V. Ball, R. Chen, D. N. Cooper, B. Giardine, F. d. Hsu, W. J. Kent, A. Lesk, D. L. Nelson, E. O'Brien W, K. Prufer, P. D. Stenson, J. C. Wallace, H. Ke, X. M. Liu, P. Wang, A. P. Xiang, F. Yang, G. P. Barber, D. Haussler, D. Karolchik, A. D. Kern, R. M. Kuhn, K. E. Smith and A. S. Zwieg (2007). "Evolutionary and biomedical insights from the rhesus macaque genome." Science **316**(5822): 222-34.

Gouveia, L. O., J. Sobral, A. M. Vicente, J. M. Ferro and S. A. Oliveira (2011). "Replication of the CELSR1 association with ischemic stroke in a Portuguese case-control cohort." Atherosclerosis **217**(1): 260-2.

Haussler, D. (2009). "Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species." J Hered **100**(6): 659-74.

Henn, B. M., L. L. Cavalli-Sforza and M. W. Feldman (2012). "The great human expansion." Proc Natl Acad Sci U S A **109**(44): 17758-64.

Hurst, L. D. (2009). "Fundamental concepts in genetics: genetics and the understanding of selection." Nat Rev Genet **10**(2): 83-93.

Izagirre, N., I. Garcia, C. Junquera, R. C. de la and S. Alonso (2006). "A scan for signatures of positive selection in candidate loci for skin pigmentation in humans." Mol Biol Evol **23**: 1697-1706.

Juriloff, D. M. and M. J. Harris (2012). "A consideration of the evidence that genetic defects in planar cell polarity contribute to the etiology of human neural tube defects." Birth Defects Res A Clin Mol Teratol **94**(10): 824-40.

Kalkman, H. O. (2012). "A review of the evidence for the canonical Wnt pathway in autism spectrum disorders." Mol Autism **3**(1): 10.

Kaucka, M., K. Plevova, S. Pavlova, P. Janovska, A. Mishra, J. Verner, J. Prochazkova, P. Krejci, J. Kotaskova, P. Ovesna, B. Tichy, Y. Brychtova, M. Doubek, A. Kozubik, J. Mayer, S. Pospisilova and V. Bryja (2013). "The Planar Cell Polarity Pathway Drives Pathogenesis of Chronic Lymphocytic Leukemia by the Regulation of B-Lymphocyte Migration." Cancer Res **73**(5): 1491-1501.

LaFramboise, T., N. Dewal, K. Wilkins, I. Pe'er and M. L. Freedman (2010). "Allelic selection of amplicons in glioblastoma revealed by combining somatic and germline analysis." PLoS Genet **6**(9).

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen and C. International Human Genome Sequencing (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lewontin, R. C. and J. Krakauer (1973). "Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms." Genetics **74**(1): 175-95.

Li, W. H. (1998). Molecular Evolution, Sinauer Associates.

Liang, W. S., D. W. Craig, J. Carpten, M. J. Borad, M. J. Demeure, G. J. Weiss, T. Izatt, S. Sinari, A. Christoforides, J. Aldrich, A. Kurdoglu, M. Barrett, L. Phillips, H. Benson, W. Tembe, E. Braggio, J. A. Kiefer, C. Legendre, R. Posner, G. H. Hostetter, A.

Baker, J. B. Egan, H. Han, D. Lake, E. C. Stites, R. K. Ramanathan, R. Fonseca, A. K. Stewart and D. Von Hoff "Genome-wide characterization of pancreatic adenocarcinoma patients using next generation sequencing." PLoS One **7**(10): e43192.

Liao, S., M. M. Desouki, D. P. Gaile, L. Shepherd, N. J. Nowak, J. Conroy, W. T. Barry and J. Geradts (2012). "Differential copy number aberrations in novel candidate genes associated with progression from in situ to invasive ductal carcinoma of the breast." Genes Chromosomes Cancer **51**(12): 1067-78.

Lindblad-Toh, K., M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, L. D. Ward, C. B. Lowe, A. K. Holloway, M. Clamp, S. Gnerre, J. Alfoldi, K. Beal, J. Chang, H. Clawson, J. Cuff, F. Di Palma, S. Fitzgerald, P. Flicek, M. Guttman, M. J. Hubisz, D. B. Jaffe, I. Jungreis, W. J. Kent, D. Kostka, M. Lara, A. L. Martins, T. Massingham, I. Moltke, B. J. Raney, M. D. Rasmussen, J. Robinson, A. Stark, A. J. Vilella, J. Wen, X. Xie, M. C. Zody, J. Baldwin, T. Bloom, C. W. Chin, D. Heiman, R. Nicol, C. Nusbaum, S. Young, J. Wilkinson, K. C. Worley, C. L. Kovar, D. M. Muzny, R. A. Gibbs, A. Cree, H. H. Dihn, G. Fowler, S. Jhangiani, V. Joshi, S. Lee, L. R. Lewis, L. V. Nazareth, G. Okwuonu, J. Santibanez, W. C. Warren, E. R. Mardis, G. M. Weinstock, R. K. Wilson, K. Delehaunty, D. Dooling, C. Fronik, L. Fulton, B. Fulton, T. Graves, P. Minx, E. Sodergren, E. Birney, E. H. Margulies, J. Herrero, E. D. Green, D. Haussler, A. Siepel, N. Goldman, K. S. Pollard, J. S. Pedersen, E. S. Lander and M. Kellis (2011). "A high-resolution map of human evolutionary constraint using 29 mammals." Nature **478**(7370): 476-82.

Locke, D. P., L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth, D. M. Muzny, S. P. Yang, Z. Wang, A. T. Chinwalla, P. Minx, M. Mitreva, L. Cook, K. D. Delehaunty, C. Fronick, H. Schmidt, L. A. Fulton, R. S. Fulton, J. O. Nelson, V. Magrini, C. Pohl, T. A. Graves, C. Markovic, A. Cree, H. H. Dinh, J. Hume, C. L. Kovar, G. R. Fowler, G. Lunter, S. Meader, A. Heger, C. P. Ponting, T. Marques-Bonet, C. Alkan, L. Chen, Z. Cheng, J. M. Kidd, E. E. Eichler, S. White, S. Searle, A. J. Vilella, Y. Chen, P. Flicek, J. Ma, B. Raney, B. Suh, R. Burhans, J. Herrero, D. Haussler, R. Faria, O. Fernando, F. Darre, D. Farre, E. Gazave, M. Oliva, A. Navarro, R. Roberto, O. Capozzi, N. Archidiacono, G. Della Valle, S. Purgato, M. Rocchi, M. K. Konkel, J. A. Walker, B. Ullmer, M. A. Batzer, A. F. Smit, R. Hubley, C. Casola, D. R. Schrider, M. W. Hahn, V. Quesada, X. S. Puente, G. R. Ordonez, C. Lopez-Otin, T. Vinar, B. Brejova, A. Ratan, R. S. Harris, W. Miller, C. Kosiol, H. A. Lawson, V. Taliwal, A. L. Martins, A. Siepel, A. Roychoudhury, X. Ma, J. Degenhardt, C. D. Bustamante, R. N. Gutenkunst, T. Mailund, J. Y. Dutheil, A. Hobolth, M. H. Schierup, O. A. Ryder, Y. Yoshinaga, P. J. de Jong, G. M. Weinstock, J. Rogers, E. R. Mardis, R. A. Gibbs and R. K. Wilson (2011). "Comparative and demographic analysis of orang-utan genomes." Nature **469**(7331): 529-33.

Lorente-Galdos, B., J. Bleyhl, G. Santpere, L. Vives, O. Ramirez, J. Hernandez, R. Anglada, G. M. Cooper, A. Navarro, E. E. Eichler and T. Marques-Bonet (2013). "Accelerated exon evolution within primate segmental duplications." Genome Biol **14**(1): R9.

Meyer, M., M. Kircher, M. T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prufer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andres, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso and S. Paabo (2012). "A high-coverage genome sequence from an archaic Denisovan individual." Science **338**(6104): 222-6.

Mukhopadhyay, P., G. Brock, C. Webb, M. M. Pisano and R. M. Greene (2012). "Strain-specific modifier genes governing craniofacial phenotypes." Birth Defects Res A Clin Mol Teratol **94**(3): 162-75.

Mullaney, J. M., R. E. Mills, W. S. Pittard and S. E. Devine (2010). "Small insertions and deletions (INDELs) in human genomes." Hum Mol Genet **19**(R2): R131-6.

Naidoo, N., Y. Pawitan, R. Soong, D. N. Cooper and C. S. Ku (2011). "Human genetics and genomics a decade after the release of the draft sequence of the human genome." Hum Genomics **5**(6): 577-622.

Nei, M. and T. Gojobori (1986). "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions." Mol Biol Evol **3**(5): 418-26.

Nickel, G. C., D. L. Tefft, K. Goglin and M. D. Adams (2008). "An empirical test for branch-specific positive selection." Genetics **179**(4): 2183-93.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. S. J, M. D. Adams and M. Cargill (2005). "A scan for positively selected genes in the genomes of humans and chimpanzees." PLoS Biol **3**(6): e170.

Oleksyk, T. K., J. F. Pombert, D. Siu, A. Mazo-Vargas, B. Ramos, W. Guiblet, Y. Afanador, C. T. Ruiz-Rodriguez, M. L. Nickerson, D. M. Logue, M. Dean, L. Figueroa, R. Valentin and J. C. Martinez-Cruzado (2012). "A locally funded Puerto Rican parrot (*Amazona vittata*) genome sequencing project increases avian data and advances youg researcher education." Gigascience **1**(14).

Oleksyk, T. K., M. W. Smith and S. J. O'Brien (2010). "Genome-wide scans for footprints of natural selection." Philos Trans R Soc Lond B Biol Sci **365**(1537): 185-205.

Oleksyk, T. K., K. Zhao, F. M. De La Vega, D. A. Gilbert, S. J. O'Brien and M. W. Smith (2008). "Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations." PLoS One **3**(3): e1712.

Pagnamenta, A. T., E. Bacchelli, M. V. de Jonge, G. Mirza, T. S. Scerri, F. Minopoli, A. Chiocchetti, K. U. Ludwig, P. Hoffmann, S. Paracchini, E. Lowy, D. H. Harold, J. A. Chapman, S. M. Klauck, F. Poustka, R. H. Houben, W. G. Staal, R. A. Ophoff, M. C. O'Donovan, J. Williams, M. M. Nothen, G. Schulte-Korne, P. Deloukas, J. Ragoussis, A. J. Bailey, E. Maestrini and A. P. Monaco (2010). "Characterization of a family with rare deletions in CNTNAP5 and DOCK4 suggests novel risk loci for autism and dyslexia." Biol Psychiatry **68**(4): 320-8.

Palsson, G. and P. Rabinow (1999). "Iceland: the case of a national human genome project." Anthropol Today **15**(5): 14-8.

Pruitt, K. D., J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M. M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. Dicuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B. L. Maidak, J. Mudge, M. R. Murphy, T. Murphy, J. Rajan, B. Rajput, L. D. Riddick, C. Snow, C. Steward, D. Webb, J. A. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin and D. Lipman (2009). "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes." Genome Res **19**(7): 1316-23.

Qu, Y., D. M. Glasco, L. Zhou, A. Sawant, A. Ravni, B. Fritzsch, C. Damrau, J. N. Murdoch, S. Evans, S. L. Pfaff, C. Formstone, A. M. Goffinet, A. Chandrasekhar and F. Tissir (2010). "Atypical cadherins Celsr1-3 differentially regulate migration of facial branchiomotor neurons in mice." J Neurosci **30**(28): 9392-401.

Reinhold, W. C., I. Erliandri, H. Liu, G. Zoppoli, Y. Pommier and V. Larionov (2011). "Identification of a predominant co-regulation among kinetochore genes, prospective regulatory elements, and association with genomic instability." PLoS One **6**(10): e25991.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler and E. S. Lander (2006). "Positive natural selection in the human lineage." Science **312**(5780): 1614-20.

Saltzman, A. L., Q. Pan and B. J. Blencowe (2011). "Regulation of alternative splicing by the core spliceosomal machinery." Genes Dev **25**(4): 373-84.

Scally, A., J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, J. Herrero, A. Hobolth, T. Lappalainen, T. Mailund, T. Marques-Bonet, S. McCarthy, S. H. Montgomery, P. C.

Schwalie, Y. A. Tang, M. C. Ward, Y. Xue, B. Yngvadottir, C. Alkan, L. N. Andersen, Q. Ayub, E. V. Ball, K. Beal, B. J. Bradley, Y. Chen, C. M. Clee, S. Fitzgerald, T. A. Graves, Y. Gu, P. Heath, A. Heger, E. Karakoc, A. Kolb-Kokocinski, G. K. Laird, G. Lunter, S. Meader, M. Mort, J. C. Mullikin, K. Munch, T. D. O'Connor, A. D. Phillips, J. Prado-Martinez, A. S. Rogers, S. Sajjadian, D. Schmidt, K. Shaw, J. T. Simpson, P. D. Stenson, D. J. Turner, L. Vigilant, A. J. Vilella, W. Whitener, B. Zhu, D. N. Cooper, P. de Jong, E. T. Dermitzakis, E. E. Eichler, P. Flicek, N. Goldman, N. I. Mundy, Z. Ning, D. T. Odom, C. P. Ponting, M. A. Quail, O. A. Ryder, S. M. Searle, W. C. Warren, R. K. Wilson, M. H. Schierup, J. Rogers, C. Tyler-Smith and R. Durbin (2012). "Insights into hominid evolution from the gorilla genome sequence." Nature **483**(7388): 169-75.

Sjodin, P., T. Bataillon and M. H. Schierup (2010). "Insertion and deletion processes in recent human history." PLoS One **5**(1): e8650.

Stoneking, M. and J. Krause (2011). "Learning about human population history from ancient and modern genomes." Nat Rev Genet **12**(9): 603-14.

Suyama, M., D. Torrents and P. Bork (2006). "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments." Nucleic Acids Res **34**(Web Server issue): W609-12.

Tissir, F. and A. M. Goffinet (2013). "Atypical cadherins celsr1-3 and planar cell polarity in vertebrates." Prog Mol Biol Transl Sci **116**: 193-214.

Tolstorukov, M. Y., N. Volfovsky, R. M. Stephens and P. J. Park (2012). "Impact of chromatin structure on sequence variability in the human genome." Nat Struct Mol Biol **18**(4): 510-5.

Ueda, S., S. Fujimoto, K. Hiramoto, M. Negishi and H. Katoh (2008). "Dock4 regulates dendritic development in hippocampal neurons." J Neurosci Res **86**(14): 3052-61.

Ueda, S., M. Negishi and H. Katoh (2013). "Rac GEF Dock4 interacts with cortactin to regulate dendritic spine formation." Mol Biol Cell.

Varki, A., D. H. Geschwind and E. E. Eichler (2008). "Explaining human uniqueness: genome interactions with environment, behaviour and culture." Nat Rev Genet **9**(10): 749-63.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C.

Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Voight, B. F., S. Kudaravalli, X. Wen and J. K. Pritchard (2006). "A map of recent positive selection in the human genome." PLoS Biol **4**(3): e72.

Volfovsky, N., T. K. Oleksyk, K. C. Cruz, A. L. Truelove, R. M. Stephens and M. W. Smith (2009). "Genome and gene alterations by insertions and deletions in the evolution of human and chimpanzee chromosome 22." BMC Genomics **10**: 51.

Walsh, E. C., P. Sabeti, H. B. Hutcheson, B. Fry, S. F. Schaffner, P. I. de Bakker, P. Varilly, A. A. Palma, J. Roy, R. Cooper, C. Winkler, Y. Zeng, G. de The, E. S. Lander, S. O'Brien and D. Altshuler (2005). "Searching for signals of evolutionary selection in 168 genes related to immune function." Hum Genet: 1-11.

Wang, E. T., G. Kodama, P. Baldi and R. K. Moyzis (2006). "Global landscape of recent inferred Darwinian selection for Homo sapiens." Proc Natl Acad Sci U S A **103**(1): 135-40.

Wetterbom, A., M. Sevov, L. Cavelier and T. F. Bergstrom (2006). "Comparative genomic analysis of human and chimpanzee indicates a key role for Indels in primate evolution." J Mol Evol **63**(5): 682-90.

Wu, C. I. and C. T. Ting (2004). "Genes and speciation." Nat Rev Genet **5**(2): 114-22.

Yamada, Y., N. Fuku, M. Tanaka, Y. Aoyagi, M. Sawabe, N. Metoki, H. Yoshida, K. Satoh, K. Kato, S. Watanabe, Y. Nozawa, A. Hasegawa and T. Kojima (2009). "Identification of CELSR1 as a susceptibility gene for ischemic stroke in Japanese individuals by a genome-wide association study." Atherosclerosis **207**(1): 144-9.

Yang, Z. (1997). "PAML: a program package for phylogenetic analysis by maximum likelihood." Comput Appl Biosci **13**(5): 555-6.

Yates, L. L., C. Schnatwinkel, L. Hazelwood, L. Chessum, A. Paudyal, H. Hilton, M. R. Romero, J. Wilde, D. Bogani, J. Sanderson, C. Formstone, J. N. Murdoch, L. A. Niswander, A. Greenfield and C. H. Dean (2013). "Scribble is required for normal epithelial cell-cell contacts and lumen morphogenesis in the mammalian lung." Dev Biol **373**(2): 267-80.

Zhao, K., Y. Ishida, T. K. Oleksyk, C. A. Winkler and A. L. Roca (2012). "Evidence for selection at HIV host susceptibility genes in a West Central African human population." BMC Evol Biol **12**: 237.

Zhou, L., F. Tissir and A. M. Goffinet (2007). "The atypical cadherin Celsr3 regulates the development of the axonal blueprint." Novartis Found Symp **288**: 130-4; discussion 134-40, 276-81.