Genomic and Functional Approaches to Genetic Adaptation

Elena Carnero Montoro

TESI DOCTORAL UPF / 2013

Thesis Director

Dra. ELENA BOSCH

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

UNIVERSITAT
POMPEU FABRA

Fitxer PDF de la tesi dividit en 7 parts

Part 1 de 7	pàg 0 - 17	Introduction Cap. 1 – Cap 2
Part 2 de 7	pàg 18 - 20	Introduction Cap. 3 : 3.1, 3.2
Part 3 de 7	pàg. 20 – 25	Introduction Cap. 3 : 3.2.1
Part 4 de 7	pàg. 25 – 28	Introduction Cap. 3: 3.2.2 – 3.2.3
Part 5 de 7	pàg. 29 - 64	Introduction Cap. 4 – Cap. 6
Part 6 de 7	pàg. 65 - 183	Objectives, Results, Discussion, Concluding remarks
Part 7 de 7	pàg. 184 – 233	References, Annexes

Part 6 de 7

Objectives

This PhD thesis project has the main purpose of deepen in the understanding of specific processes of genetic adaptation by overcoming some of the limitations of genome-wide studies of positive selection.

Three studies have been performed to achieve the main objective, in:

1) **The human immune receptor *CD5* gene.** In this case, our specific aims were:

    a. to confirm previous SNP-based signatures of selection by generating and analyzing resequencing data in four human populations of African, European and Asian origin using classical tests of selection;
    b. to detect all possible adaptive variants in the gene region;
    c. to functionally assay their possible implication in immune adaptation and other phenotypic traits.

2) **The human zinc transporter *SLC39A4* gene**. Here, our specific aims were:

    a. to investigate whether a highly differentiated non-synonymous variant in this zinc transporter could have reached extreme frequencies in West Africans as result of positive selection even though no other classical signals of a selective sweep are present along the surrounding genomic region;
    b. to evaluate the functional impact of this non-synonymous variant regarding the cellular zinc uptake mediated by the transporter; as well as,
    c. to discuss its possible evolutionary relevance.

3) **In diverse functional pathways**. The objectives of this last work will be:

   a. to combine and compare human and chimpanzee divergence data with polymorphism in chimpanzees in two pre-defined reference data sets for acceleration and constrained evolutionary patterns in the chimpanzee lineage such as the complement and the actin pathways;
   b. to interrogate the specific patterns of evolution of three pathways related to neurological function (Parkinson, amiloid and presenilin);
   c. to compare possible differential evolutionary trends among the coding DNA sequences (CDS), introns and regulatory regions of all these pathways.

Results

Carnero-Montoro E, Bonet L, Engelken J, Bielig T, Martínez-Florensa M, Lozano F, Bosch E. Evolutionary and functional evidence for positive selection at the human CD5 immune receptor gene. Mol Biol Evol. 2012 Feb;29(2):811-23. doi: 10.1093/molbev/msr251

Chapter 2

Extreme Population Differences in the Human Zinc Transporter ZIP4 (SLC39A4) are explained by Positive Selection in Sub-Saharan Africa

Submitted to PLOS Genetics.

Authors and Affiliations

Johannes Engelken[1,2,☯], Elena Carnero-Montoro[1,☯], Marc Pybus[1], Glen K. Andrews[3], Carles Lalueza-Fox[1], David Comas[1], Israel Sekler[4], Marco de la Rasilla[5], Antonio Rosas[6], Mark Stoneking[2], Miguel A. Valverde[7], Rubén Vicente[7,¶,*], Elena Bosch[1,¶,*]

[1]Institute of Evolutionary Biology (CSIC-UPF), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, 08003 Barcelona, Spain
[2]Department of Evolutionary Genetics, Max-Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany
[3] Department of Biochemistry and Molecular Biology, University of Kansas Medical Center, Kansas City, KS 66160-7421, USA
[4] Department of Physiology, Ben-Gurion University, 84105 Beer-Sheva, Israel
[5] Área de Prehistoria, Departamento de Historia, Universidad de Oviedo, 33011 Oviedo, Spain
[6] Group of Paleoanthropology MNCN-CSIC, Department of Paleobiology, National Museum of Natural Sciences, CSIC, 28006 Madrid, Spain
[7] Laboratory of Molecular Physiology and Channelopathies, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, 08003 Barcelona, Spain

☯ These authors contributed equally to this work
¶ These authors are joint senior authors on this work.
*E-mail: ruben.vicente@upf.edu (RV), elena.bosch@upf.edu (EB)

Running head: **Selection on a human zinc transporter gene in Africa**

ABSTRACT

Extreme differences in allele frequency between West Africans and Eurasians are observed for a leucine to valine substitution (Leu372Val) in the human intestinal zinc uptake transporter, ZIP4. However, there is no further evidence for a selective sweep around the *ZIP4* gene (*SLC39A4*). By interrogating allele frequencies in more than 100 diverse human populations and resequencing Neanderthal DNA, we have confirmed the ancestral state of this locus and found a strong geographical gradient for the derived allele (Val372), with near fixation in West Africa. In extensive coalescent simulations, we show that the extreme allele frequency differences yet absence of a classical sweep signature can be explained by the effect of a local recombination hotspot, together with directional selection favoring the Val372 allele in Sub-Saharan Africans. The possible functional effect of the Leu372Val substitution together with two pathological mutations at the same codon (Leu372Pro and Leu372Arg) that cause acrodermatitis enteropathica (a disease phenotype characterized by extreme zinc deficiency) was investigated by transient overexpression of human ZIP4 protein in HeLa cells. Both acrodermatitis mutations cause absence of the ZIP4 transporter cell surface expression and nearly absent zinc uptake, while the Val372 variant displayed significantly reduced surface protein expression, reduced basal levels of intracellular zinc, and reduced zinc uptake when compared with the Leu372 variant. We speculate that reduced zinc uptake by the *ZIP4* derived Val372 allele may act to starve certain pathogens from zinc, and hence may have been advantageous in Sub-Saharan Africa. Moreover, these functional results may indicate the existence of zinc homeostasis differences among modern human populations with possible relevance for disease risk.

AUTHOR SUMMARY

Zinc is an essential trace element with many biological functions in our body and whose concentrations are tightly regulated by different membrane transporters. Here we report an unusual case of positive natural selection for an amino acid replacement in the human intestinal zinc uptake transporter ZIP4. This substitution had been previously recognized as one of the most strongly genome-wide differentiated polymorphisms between different human populations. However, since the extreme population differentiation of this non-synonymous site was not accompanied by additional signatures of natural selection, it was unclear whether this resulted from genetic adaptation. Using computer simulations we demonstrate that such an unusual pattern can be explained by the effect of the local recombination, together with positive selection in Sub-Saharan Africa. Moreover, we provide evidence to suggest functional differences between the two ZIP4 isoforms in terms of the transporter cell surface expression and zinc uptake. This result is the first genetic indication that zinc regulation differs among modern human populations, a finding that may have implications for health research. Further, we speculate that reduced zinc uptake mediated by the derived variant may have been advantageous in Sub-Saharan Africa, possibly by restricting access of this micronutrient to a geographically restricted pathogen.

INTRODUCTION

Zinc homeostasis is critically important for human health. Similar to iron, zinc has manifold functions in the body, such as in the immune system (Rink and Haase 2007), aging (Swindell 2011), DNA repair (Ho and Ames 2002), signaling (Haase and Rink 2009) and in diseases such as diabetes (Jansen, Karges, and Rink 2009) and cancer (Alam and Kelleher 2012). On the molecular level, zinc acts as a co-factor in hundreds of metallo-enzymes as well as in hundreds of DNA-binding proteins (e.g. zinc finger proteins). Zinc homeostasis is tightly regulated by 10 zinc efflux transporters and 14 zinc influx transporters (encoded by the *SLC30A* and *SLC39A* gene families, respectively). ZIP4 (SLC39A4) is the most important intestinal zinc uptake transporter and is expressed at the apical membrane of enterocytes (Dufner-Beattie et al. 2003; K. Wang et al. 2002). Loss-of-function mutations in *ZIP4* cause acrodermatitis enteropathica (Küry et al. 2002; F. Wang et al. 2004) [MIM 201100], a congenital disease characterized by extreme zinc deficiency if left untreated without supplemental zinc (Moynahan 1974; Neldner and Hambidge 1975). Fittingly, a recent report of a *ZIP4* intestine-specific knockout mouse showed that loss of expression of this gene causes systemic zinc deficiency, leading to disruption of the intestine stem cell niche and loss of intestine integrity (Geiser et al. 2012b).

The single nucleotide polymorphism (SNP) c.1114C>G (rs1871534) in the *ZIP4* gene (*SLC39A4*; NM_130849.2) results in the substitution of leucine for valine at amino acid 372 (Leu372Val) in the human ZIP4 transporter. This non-synonymous SNP is one of the most markedly differentiated genetic variants in the genome in terms of allele frequency differences between populations (Barreiro et al. 2008; Xue et al. 2009; The 1000 Genomes Project Consortium 2012) using data from HapMap (Frazer et al. 2007), the Human Genome Diversity Panel (HGDP) (Cann et al. 2002) and the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012). Extreme population differentiation is one signature of local positive selection (Tamara Hofer, Foll, and Excoffier 2012; T Hofer et al. 2009; Gardner et al. 2007; Xue et al. 2009), but genomic scans for targets of natural selection based on other criteria, such as extended long haplotypes (Pickrell et al. 2009; Voight et al. 2006; Tang, Thornton, and Stoneking 2007) or selective signatures in the allele frequency spectrum (Carlson et al. 2005), have failed to identify *ZIP4* as a candidate gene of positive selection. To date, whether this variant has evolved under positive selection or neutrality, and its potential functional significance, has not been examined.

In this paper, we have three main objectives: (i) to investigate evolutionary explanations for the extreme population differentiation of the ZIP4 Leu372Val polymorphism by use of coalescent simulations; (ii) to test for functional differences in cellular zinc transport between the alleles at the Leu372Val polymorphism using an heterologous expression system; as well as (iii) to discuss potential selective forces behind this possibly adaptive event and their implications for zinc homeostasis in modern humans. We have extensively characterized the extreme geographical differentiation of the Leu372Val substitution and provide evidence that it has been subject to a nearly complete but mild selective sweep in Sub-Saharan Africa. Our simulations show how the extreme pattern of population differentiation but absence of other classical signatures of positive selection can be explained by directional selection accompanied by the effects of a recombination hotspot near the polymorphic adaptive site. Additionally, our data demonstrate *in vitro* functional differences between the two human polymorphic alleles at codon 372 of the human ZIP4 transporter regarding surface protein expression, basal intracellular levels of zinc and zinc uptake. We hypothesize that the reduction in intracellular zinc levels mediated by the Val372 allele may have been advantageous in Sub-Saharan Africa, possibly by restricting access of this micronutrient to a geographically restricted pathogen, and that other possible secondary consequences on disease risk and health may result from the differential activity of the ZIP4 alleles.

RESULTS

Worldwide allele frequencies

Five common non-synonymous SNPs are known in the human *ZIP4* gene (Table 1); these are Glu10Ala (rs2280839), Ala58Thr (rs2280838), Ala114Thr (rs17855765), Thr357Ala (rs2272662) and Leu372Val (rs1871534). However, only the latter two SNPs show elevated levels of population differentiation in the 1000 Genomes Phase1 sequencing data when comparing the Yoruba from Ibadan, Nigeria (YRI), with either the Han Chinese from Beijing, China (CHB) or the Utah residents with Northern and Western European origin (CEU). As shown in Figures 1A and 1B, their F_{ST} values fall above the 99th percentile of the genome-wide F_{ST} distributions between CEU-YRI (with F_{ST} values for rs2272662 and rs1871534 of 0.48 and 0.98, respectively) and between CHB-YRI (with F_{ST} values of 0.51 and 0.98, respectively). We therefore verify that the Leu372Val substitution encoded by SNP rs1871534 is the most extreme non-synonymous polymorphism within the human *ZIP4* gene. Next, we genotyped the 51 populations from the Human Genome Diversity Panel (HGDP) and compiled additional allele frequencies for this position in worldwide populations from the Alfred database (Osier et al. 2002; Rajeevan et al. 2003). Further we have provided new data from a Pygmy population from Gabon and North African populations from Western Sahara, Morocco, and Libya. These new data confirm that the Leu372 variant is the most common allele outside of Africa, and provide a more detailed picture of the geographical frequency differentiation of this non-synonymous polymorphism (Figure 1C and Supplemental Table S1). Overall, the Val372 variant showed the highest frequencies in Sub-Saharan Africa with populations such as the Ibo or the Yoruban people showing the most extreme derived allele frequencies worldwide (0.99 and 0.96, respectively). Interestingly, two presumably early-branching groups in Sub-Saharan Africa, the Pygmy and the San people, showed opposite trends in the derived allele frequency

(0.94 and 0.0, respectively). However, the absence of the Val372 allele in San could simply reflect their small sample size (only 6 individuals). Given the elevated levels of population differentiation of the SNP rs2272662 we also genotyped the HGDP panel for the Thr357Ala polymorphism. However, this non-synonymous SNP displayed intermediate frequencies worldwide (Supplemental Figure S1 and Supplemental Table S1) and less extreme allele frequency differences between populations compared with the Leu372Val substitution.

Identification of Leu372 as the ancestral variant by resequencing in a Neanderthal

Given the allele frequency differences observed for the Leu372Val polymorphism between the two early human branches in Africa and the uncertainty associated with the low coverage of the Neanderthal genome draft (Green et al. 2010), we resequenced the corresponding orthologous positions for rs1871534 and rs2272662 in an additional Neanderthal sample, labeled SD1253 and excavated at El Sidrón site in Spain (Rosas et al. 2006). The two positions were amplified in a multiplexed reaction along with a diagnostic Neanderthal mitochondrial DNA (mtDNA) fragment to monitor contamination in the PCR reaction. For the L16230-H16262 diagnostic mtDNA fragment, 64 clones were generated (Supplemental Figure S2), all of which show the Neanderthal-specific 16234T-16244A-16256A-16258G haplotype (Green et al. 2010). This again supports the very low level of contamination in this particular sample. For the orthologous positions of the human rs1871534 and rs2272662 SNPs, 19 and 14 sequences were successfully obtained, respectively. With the exception of one clone in the second position, all sequences showed the previously inferred ancestral alleles, in agreement with the reads present for the Vindija 33.16 (one read for each position), 33.25 (two for rs1871534 and none for rs2272662) and 33.26 (two and one, respectively) individuals (Figure 2). The successful resequencing of this Neanderthal individual together with published reads from additional Neanderthals (Green et al. 2010) and from the Denisovan individual (Meyer et al. 2012) strongly suggest that the Leu372 variant (encoded by the C allele in rs1871534) is the human ancestral form, which is also in agreement with the chimpanzee state (Figure 2). Together with the extreme population differentiation pattern, these results suggest that a selective sweep may have taken place in Sub-Saharan Africa, where the derived variant is nearly fixed.

Extreme population differentiation explained by selection and a recombination hotspot

Next we examined the complete genomic region around *ZIP4* (Figure 3) in the 1000 Genomes sequencing data. Whereas we found a cluster of three strongly elevated F_{ST} scores between CEU and YRI in the neighboring SNPs rs1871535 (intronic), rs1871534 and rs2272662 (further suggesting directional selection in a specific geographical region), in both populations there was a clear absence of extreme values in neutrality statistics, such as Tajima's D or Fay and Wu's H (Supplemental Figure S3). Notably, no other polymorphism in the flanking region of the human ZIP4 displays the high levels of population differentiation of the Leu372Val substitution. Interestingly, in African and non-African populations there is a recombination hotspot in the *ZIP4* gene, which could have reduced any signature of selection on the surrounding linked variation, thereby explaining the apparent lack of significant departures from neutrality. To further investigate this possibility, we carried out coalescence simulations under a variety of recombination and selection scenarios using a well-established demography (Schaffner et al. 2005). As shown in Figure 3D, the observed empirical value for F_{ST} and for most of the different neutrality statistics explored cannot be entirely explained either by neutral evolution or by positive selection with a constant recombination rate. Instead, this atypical pattern of extreme population differentiation but seemingly neutral Tajima´s D and other neutrality statistics could be fully recovered in simulations with directional selection on the derived allele in Sub-Saharan African populations in the context of the observed recombination landscape, including the hotspot (Figures 3D and 3E). Therefore, a scenario that can explain the results is that positive selection has indeed acted upon the Val372 allele in Sub-Saharan African populations and that recombination has subsequently erased further accompanying signatures of the selective sweep.

Functional effect of Leu372Val

We observed that the Leu372Val polymorphism affects a highly conserved amino acid (Figure 4) and that the same codon position has been altered in acrodermatitis patients carrying missense mutations Leu372Arg (Li et al. 2010) and Leu372Pro (K. Wang et al. 2002). Moreover, both PolyPhen (Adzhubei et al. 2010) and SIFT (Ng 2003) algorithms predict functional effects for the Leu372Val substitution (see Table 1). These observations led us to test the Leu372Val polymorphism for a possible functional change in the ZIP4 transporter, using transiently transfected HeLa cells. To be able to control for possible haplotypic effects between the two most highly differentiated non-synonymous SNPs in the ZIP4 transporter, we also considered variation at the Thr357Ala polymorphism in the functional analyses. Furthermore, we introduced the pathological mutations Leu372Arg and Leu372Pro in the Ala357 background of the human *ZIP4* gene and analyzed them as well. The pathological impact of the Leu372Pro mutation on ZIP4 protein biology and function has already been evaluated in the mouse ZIP4 protein (F. Wang et al. 2004), but not the Leu372Arg mutation. Besides providing confirmation of their impact in the human gene context, the use of these pathological mutations provides an extreme phenotype to which to compare the phenotypic relevance of the *ZIP4* non-synonymous polymorphisms. In all cases, functional analyses were carried out for effects on expression, subcellular localization, and for zinc transport.

As shown in Figure 5, human ZIP4 proteins carrying the Leu372Pro and Leu372Arg mutations showed absence of surface protein expression ($P<0.001$, one way ANOVA versus the Ala357-Leu372 isoform), consistent with the known causal role of these variants in the zinc deficiency disorder acrodermatitis enteropathica. Interestingly, the derived Val372 variant also showed significantly decreased surface expression, but to a much lesser extent, and independently of the Thr357Ala substitution ($P<0.05$ in both Ala357 and Thr357 backgrounds; one way ANOVA versus the Ala357-Leu372 isoform). Overall, the Leu372Val substitution had a highly significant effect on surface expression (ANOVA, $p= 0.00021$), while there was no effect due to the Thr357Ala replacement ($p=0.579$). Western blot analysis of all isoforms revealed a remarkable decrease in detection of the Ala357-Pro372 isoform (Supplemental Figure S4A). However, the reduced expression of this isoform was not due to a defect in the construct sequence but to a higher protein degradation rate, as shown in Supplemental Figure S4B. Further analysis showed that the Ala357-Leu372 and Ala357-Val372 isoforms do not differ in protein degradation rate. Therefore the differences in the surface expression experiment must be due to a different trafficking pattern of these variants. In this sense co-localization of ZIP4 with calnexin (a protein present in the lumen of the endoplasmic reticulum) indeed showed that those proteins presenting lower surface expression were partially retained in the endoplasmic reticulum (Supplemental Figure S5).

Zinc transport analysis of the different ZIP4 isoforms was performed in two ways. First, we quantified basal zinc content with FluoZin-3 in HeLa cells that overexpressed the various ZIP4 variants during a 24 hour period (Figure 6A), and second, we recorded intracellular zinc uptake upon perfusion with an external solution containing 200 µM Zn^{2+} (Figure 6B). Our results show that basal zinc content in cells overexpressing pathological variants Pro372 and Arg372 did not differ from surrounding non-transfected HeLa cells. On the contrary, all common ZIP4 variants (Ala357, Thr357, Leu372 and Val372) promoted increased intracellular zinc levels. However, and in agreement with their reduced surface expression, Val372 variants (in both Ala357 and Thr357 backgrounds) presented lower basal zinc content compared to Leu372 ($P<0.01$ and $P<0.05$, respectively; one way ANOVA versus the Ala357-Leu372 isoform; Figure 6A). As shown in Figure 6B, cells overexpressing the pathological Leu372Arg and Leu372Pro mutations did not uptake zinc, consistent with their inability to traffic to the plasma membrane. Zinc uptake mediated by the Val372 variants was also consistent with their reduced membrane expression; i.e. the Val372 variants in both Ala357 and Thr357 backgrounds presented significantly lower maximum transport (T_{max}) compared to the Leu372 variant ($P<0.01$ in each case; Figure 6B). However, the time to reach half-maximal transport ($t_{1/2}$) showed no significant difference, indicating that transport kinetics were not markedly different between the four common variants (Figure 6). Overall, these results support the idea that the Val372 variant does not disturb the kinetics of the ZIP4 transporter but leads to lower zinc uptake transport due to reduced surface expression.

DISCUSSION

Leu372Val as the target of an atypical selective sweep in Africa

We began with the observation of extreme population differentiation between Sub-Saharan African and non-African populations involving the Leu372Val polymorphism in the *ZIP4* gene, but no other signals of a classic hard sweep such as long extended haplotype homozygosity (Supplemental Figures S3, S6 and S7). By interrogating and compiling allele frequencies in more than 100 worldwide human populations we further characterized the extreme population differentiation of the Leu372Val polymorphism and confirmed that this result is not an artifact resulting from allele switching (Xue et al. 2009). Furthermore, given the worldwide distribution of the human derived and ancestral alleles (as confirmed by sequencing a Neanderthal and through phylogenetic conservation), this sweep appears to have taken place inside Africa, probably in Sub-Saharan Africa, and not outside the African continent. Notably, the extreme population differentiation for the Leu372Val polymorphism represents the fourth top region within the global genome-wide F_{ST} distribution between CEU-YRI obtained from the 1000 Genomes Project data. The only more extreme F_{ST} CEU-YRI values all involve well-known examples of local geographical adaptation in humans: the *SLC24A5* and *SLC45A2* genes (with an F_{ST} of 0.9826 and 0.9765, respectively), which have been associated with light skin pigmentation in Europeans; and the *DUFFY* gene (with an F_{ST} of 0.9765), which provides resistance to the malaria pathogen *Plasmodium vivax*. Moreover, with the notable exception of *DUFFY* FY*O allele (M T Hamblin and Di Rienzo 2000; Martha T Hamblin, Thompson, and Di Rienzo 2002), most of the extreme F_{ST} values when comparing Africans to non-Africans are usually attributed to local adaptation outside of Africa. It is thus quite remarkable that we have detected such a rare signature of natural selection in the African continent. Moreover, this is congruent with a recent study that finds only limited evidence for classical sweeps in African populations, which is likely due to a combination of limitations in the currently used methodology and specific characteristics of African population history (Granka et al. 2012).

Interestingly, we observe a nearly complete but mild selective sweep for the Val372 variant in Africa, which involves three SNPs with extremely elevated measures of population differentiation, whereas most other commonly used tests for selection show values not even close to genome-wide significance. However, our simulation results indicate that this unusual pattern might be explained by local positive selection in combination with an observed recombination hotspot of moderate strength. At approximately 7 cM/Mb, the recombination rate is only around 7-fold higher than the genomic background, but the hotspot is extended over 3-4 kb. Therefore, it may accumulate a similar number of recombination events over time corresponding to a more typically-sized hotspot of 1 kb and a recombination rate of around 25 cM/Mb. To our knowledge, this is the first example of a selective sweep that is obscured by the effect of a recombination hotspot. It is compatible with earlier theoretical observations that instances of weaker selection in the presence of recombination may not always have an influence on polymorphism statistics (Tennessen, Madeoy, and Akey 2010). Because of the unclear effects of the recombination hotspot, it has not been possible to estimate the age of the sweep using linkage disequilibrium decay related methods (e.g. (Beleza et al. 2013)). Instead, in our simulations we fixed three different selection coefficients and, thus, tested indirectly how a selective sweep could have produced the actual frequencies observed in Yorubans at three different timeframes. In these sense, under the influence of a moderate selection coefficient (which corresponds to the oldest timeframe) we observe that the corresponding distributions of simulated neutrality and differentiation statistics are compatible with the combination of scores observed. Such a pattern seems to indicate an older selective sweep (i.e. between eighty-five and sixty thousand years ago) as the more plausible case.

Other more complex scenarios cannot be entirely ruled out, and could be investigated in more detail. For example: (i) selection acting on standing genetic variation in the sense that the Val372 variant was already segregating when it came under the influence of local selection; (ii) additional directional selection against the Val372 allele in non-African populations; (iii) selection favoring the Leu372 variant on multiple, geographically independent origins mostly in non-African populations; and (iv) 'gene surfing' of any of the two variants on a wave of a population range expansion (Excoffier and Ray 2008). However, we note that it is not necessary to invoke such complex scenarios in preference to the more parsimonious one we propose based on coalescent simulations. Moreover, back-and-forth migrations between Sub-Saharan African, Northern African and Middle Eastern populations after the first Out-of-Africa wave of migration (Henn et al. 2012) could easily explain the observed low-intermediate allele frequencies in Middle Eastern populations without invoking additional selection events.

In the absence of additional linked functional variants in the region, we infer that directional selection has acted on the *ZIP4* gene. This conclusion is supported by: (i) the disease phenotype of acrodermatitis enteropathica, which involves extreme and potentially lethal zinc deficiency and is caused, among others, by diverse mutations at amino acid position 372 in ZIP4 (Schmitt et al. 2009); (ii) the absence of cellular zinc transport in Leu372Arg and Leu372Pro acrodermatitis mutants; (iii) the finding that the Val372 variant leads to reduced zinc transport at the cellular level; and finally (iv) the conservation of this amino acid position across diverse species (Figure 4). Furthermore, we infer the Leu372Val substitution to be the functional site that was the target of selection because it is located in the predicted center of selection (highest F_{ST}), and is the only putative functional polymorphism in the *ZIP4* gene. Of the other two polymorphic variants with somewhat high allele frequency differences between populations, the Thr357Ala substitution (rs2272662) does not show any functional effect and the intronic rs1871535 cannot be associated with any known regulatory function (according to DNAse I hypersensitivity clusters, CpG Islands and transcription factor binding sites information available from the ENCODE data (http://genome.ucsc.edu/ENCODE (Rosenbloom et al. 2012)).

Therefore, both rs1871535 and rs2272662 are likely to be neutral. Other non-synonymous polymorphisms with intermediate allele frequencies in the *ZIP4* gene (Glu10Ala, Ala58Thr, and Ala114Thr) have very low F_{ST} scores and are therefore not considered candidate variants for selection. Additional non-synonymous mutations in *ZIP4* are known to cause acrodermatitis enteropathica but these have extremely low allele frequencies and cannot be considered as possible targets of positive selection.

Possible consequences at the cellular and organ level

Our functional results in transfected HeLa cells indicate that the Val372 form of the ZIP4 receptor has lower relative cell surface expression. Interestingly, we found that this decreased expression translated into reduced zinc transport of the derived Val372 variant at the cellular level. That is, we observed differences in the maximal transport (T_{max}) with no significant differences in the transport kinetics ($T_{1/2}$) between Leu372 and Val372. The functional results observed in transfected HeLa cells are likely to be also applicable to other epithelial cells. The effects of mutations at the 372 codon of ZIP4 on surface expression (in CHO cells) and on zinc transport (in HEK293 cells) have been independently demonstrated for various acrodermatitis variants in mouse cDNA (F. Wang et al. 2004), in addition to the common polymorphisms and the two acrodermatitis variants in human cDNA in this study (four variants in HeLa cells). Since both studies have used different epithelial cell lines, these general results possibly are transferable to cells from the intestine, liver, kidney, and others. However, the critical function of ZIP4 in knockout studies has been shown to primarily affect intestinal uptake of zinc (Geiser et al. 2012b).

In contrast to the Leu372Pro and Leu372Arg acrodermatitis mutations, which served as controls and which showed almost complete absence of zinc transport, both the Leu372 and Val372 variants are capable of carrying out zinc transport in the normal range of concentrations, as expected given their high frequency in the healthy population. The consequences of this difference in zinc transport at the organ and organismal level are currently unclear, although there is a strong indication that this variant may indeed be phenotypically relevant. Indeed, a similar non-synonymous mutation in the porcine homologue of ZIP4 leads to non-pathogenic reduced tissue concentrations of zinc in piglets (Siebert et al. 2012).

Nutritional immunity as a putative selective force

Could the concept of "nutritional immunity" (Kochan 1973; Hood and Skaar 2012) involving zinc explain a putative selective force in Sub-Saharan Africa? According to this hypothesis, the human host restricts access to certain micronutrients, so that pathogens become less virulent. This is a well-known mechanism of immune defense mediated by iron metabolism (Weinberg 1977), and there are indications that zinc metabolism could have a similar function (Kehl-Fie and Skaar 2010; Hood and Skaar 2012). For example, hypoferremia and hypozincemia are both part of the acute phase response to infection and both seem to be influenced by a different zinc transporter from the same family, ZIP14 (Beker Aydemir et al. 2012). We speculate that the selective force behind the extreme F_{ST} pattern of the Leu372Val substitution may be related to pathogens or infectious diseases. It is known that decreased zinc uptake mediated by ZIP4 leads to decreased zinc concentrations in the major organs, as shown in a mouse knockout model (Geiser et al. 2012b). While the phenotypic effect of the Val372 allele in humans is currently unknown, we conjecture that the *in vitro* difference may indeed translate into physiological differences, possibly leading to a slightly decreased uptake of dietary zinc. Fittingly, there is suggestive evidence that African genetic ancestry may involve lower serum levels of zinc (Cole et al. 2010), as African-American children have a fourfold risk of zinc deficiency compared to Hispanic children. This result would suggest that African ancestry may be associated with lower serum zinc levels, although these results may be biased due to differences in lifestyle, socio-economic status etc., and this observation would need to be confirmed by controlled studies. Quantification of such differences would need to account for the expected marginal differences in ZIP4 activity between healthy individuals. Further, any experimental results could be complicated by the possible compensatory effects of protein expression in response to overall cellular and organismal levels of zinc, including circadian regulation. Alternatively, lower zinc concentrations mediated by the Leu372Val substitution in the enterocyte cells could facilitate early diarrheal episodes during a digestive infection in order to reduce the pathogen load on the luminal surface (Hoque, Rajendran, and Binder 2005; Scrimgeour and Lukaski 2008). Similarly, the lower level of expression of the ZIP4 isoform carrying the Val372 variant could also have resulted advantageous if any parasite uses the ZIP4 receptor to enter enterocytes. Furthermore, the selective force may be related to pre-historic differences in dietary zinc due to lifestyle or due to local levels of zinc concentrations in soil and the food chain.

Potential implications – towards a phenotype

No large-scale ethnic comparisons related to serum or tissue zinc concentrations are available. To our knowledge, rs1871534 has not been tested in case-control studies in African populations related to one of the numerous existing infectious diseases like malaria, trypanosmias or Lhassa fever. Therefore it is possible that important findings related to a possible selective force have been missed. In future studies, extension to additional cell lines, and genotype-phenotype association studies in diverse ethnic populations may help to clarify further phenotypic consequences of this non-synonymous polymorphism. Genotype-phenotype association studies should involve African-American or East

African populations in which the Val372 allele is segregating at intermediate frequencies. Candidate phenotypes and traits to interrogate could be serum zinc concentrations, zinc content in hair and nails, serum zinc concentrations after controlled zinc supplementation, and a range of disease traits, especially diseases with an elevated risk in different populations, for example diverse types of cancer in African Americans. As this SNP was not included in the commonly used Affymetrix and Illumina SNP arrays with up to one million variants (although it is included in several of the latest arrays), potential associations of clinical relevance may have been missed. Interestingly, common polymorphisms in other zinc transporters show genome-wide associations with disease traits, such as a non-synonymous variant in the zinc efflux transporter ZnT8 (SLC30A8) and diabetes incidence (Sladek et al. 2007) as well as a regulatory variant in the zinc influx transporter ZIP6 (SLC39A6) and survival in esophagal cancer (Wu et al. 2013).

CONCLUSIONS

The identification of a high-frequency derived allele polymorphism in the *ZIP4* zinc transporter gene (*SLC39A4),* combined with a more complete picture of worldwide allele frequencies and in-depth coalescent simulations, is consistent with a long lasting selective event in Sub-Saharan Africa driven by a moderate selection coefficient. This event did not leave the typical footprint of a selective sweep with long haplotypes or detectable neutral deviations in the allele frequency spectrum of the surrounding region, most likely because of the presence of a moderate recombination hotspot. Through functional experiments we have verified the Leu372Val substitution as the likely causal site. Given that two functionally different alleles of this key component of cellular zinc uptake are distributed so divergently across worldwide populations, our results may point to functional differences in zinc homeostasis among modern human populations with possible broader relevance for health and disease.

MATERIALS AND METHODS

Samples and Genotyping

The G and C alleles at rs1871534 (Leu372Val) have been swapped in various public sources such as HapMap (http://**www.hapmap.org**) or dbSNP (http://www.ncbi.nlm.nih.gov/SNP) that report conflicting allele frequencies in populations with similar geographical origin. This situation led us to repeat the genotyping of this SNP in the Human Genome Diversity Panel (HGDP-CEPH) (Cann et al. 2002). We also genotyped rs2272662 (which causes the Thr357Ala substitution) because, within the *ZIP4* gene, it shows the second highest allele frequency differences between CEU and YRI HapMap populations and allele frequencies were not available at the worldwide level. The rs1871534 and rs2272662 loci were genotyped in the H971 subset (Rosenberg 2006) of the HGDP-CEPH (Cann et al. 2002), representing 51 worldwide populations and in an additional population from Africa: Pygmies from Gabon (N=39)(Berniell-Lee et al. 2009). We also genotyped rs1871534 in North African populations from Western Sahara (Saharawi, N=50), Morocco (Casablanca, N=30; Rabat, N=30; Nador, N=30) and Libya (Libyans, N=50). Genotyping was performed using Taqman assays C__11446716_10 and C__26034235_10 on an Applied Biosystems Light Cycler (7900HR), according to standard protocols. Additional genotypes for rs1871534 were obtained from the Alfred database (http://alfred.med.yale.edu) (Osier et al. 2002; Rajeevan et al. 2003).

Neanderthal Resequencing

The El Sidrón Neanderthal sample SD1253 has been used in many paleogenomic studies due to its high endogenous DNA content and low contamination levels (Krause et al. 2007; Lalueza-Fox et al. 2007; Lalueza-Fox et al. 2009; Briggs et al. 2009; Maricic et al. 2012; Green et al. 2010), attributable in part to having been extracted using an anti-contamination protocol (Fortea et al. 2008). In addition, it has the advantage of having been dated to 49,000 years ago (Torres, Ortiz, and Grün 2010), prior to the arrival of modern humans to Europe. The two orthologous positions for rs1871534 and rs2272662 were amplified using a two-step PCR protocol (Lalueza-Fox et al. 2007) in a multiplexed reaction along with a diagnostic Neanderthal mitochondrial DNA (mtDNA) fragment. After visualizing the PCR products in a low-melting temperature agarose gel, the bands were excised, purified and cloned using the TOPO-TA cloning kit (Invitrogen). Inserts of the correct size were sequenced on an ABI3730 XL capillary sequencer (Applied Biosystems).

Simulations

Simultaneous coalescent simulation of recombination hotspots and selection were carried out using Cosi v1.2 (Schaffner et al. 2005). For underlying neutral demography, we used the best-fit model with slight modifications, similar to a previously used approach (Grossman et al. 2010). Particularly, the migration frequencies were set to zero and the time points of the European and African population bottlenecks were moved back in time to 3,300 generations before present in order to accommodate the long sweep times resulting from the lowest selection coefficient we used (0.5%). The sweep was shifted back 350 generations in order to retain the final population expansions with the advantage of (i) better approximating the fitted model, and of (ii) of generating sufficient singletons when compared to the 1000 Genomes Phase1 data. Subsequent thinning of the simulated data was performed across all populations to account for the underestimation of singletons in 1000 Genomes data. For each simulation, we either used the recombination landscape including hotspots from the YRI population provided by the 1000 Genomes Consortium and based on HapMap 2 trio data (http://1000genomes.org) or alternatively used a constant recombination rate of 8.1702 x 10^{-9} which was calculated as the mean recombination rate in the 100 kb window surrounding *ZIP4*. Simulations had a length of 100 kb, were run in 500 replicates for each scenario and sample sizes were set to 176 chromosomes for Sub-Saharan Africans and 194 chromosomes for Europeans. Regions under positive selection were modeled using a single causal variant that rose to an allele frequency of 0.98 corresponding approximately to that observed now in YRI. We simulated three different selection coefficients (0.5%, 2% and 3%) that lead to different durations of the sweep: 2,938 generations (~60,000 – 85,000 years for generation times of 20 and 29 years, respectively;(Fenner 2005)), 1,469 generations (~30,000 – 43,000 years), or 458 generations (~10,000 – 13,000 years).

Neutrality Tests on Simulated and the 1000 Genomes Data

Neutrality tests on simulated and the 1000 Genomes population data were performed as described by Pybus *et al.* (submitted) and using the 1000 Genomes selection browser (http://pgb.ibe.upf.edu). Briefly, Tajima's D, Fu and Li´s D and Fay and Wu's H were calculated using a sliding window approach with 25 kb windows and approximately 3 kb offset. F_{ST} (Weir and Hill 2002) and XP-

EHH (Pardis C Sabeti et al. 2007) between CEU and YRI were calculated for each polymorphic position.

Cells and Reagents

Human *ZIP4* cDNA encoding the long isoform of the protein and the Ala357 and Leu372 variants was cloned into pcDNA 3.1 (+) expression vector together with an hemagglutinin (HA) tag at the carboxyl terminus as described previously (Tahio 2009). The Leu372Pro and Leu372Arg mutants, as well as the Thr357Ala and Leu372Val polymorphisms, were introduced via site-directed mutagenesis following standard conditions (QuikChange II XL; Stratagene; see Supplemental Table S2 for complete human cDNA and primers used in the mutatgenesis). The six human *ZIP4* isoforms obtained (i.e. Ala357-Leu372, Ala357-Val372, Thr357-Leu372, Thr357-Val372 as well as Ala357-Pro372 and Ala357-Arg372) were confirmed by sequencing with the ABiPrism 3.1 BigDye kit before their use in transfection experiments. HeLa cells were cultured in DMEM plus 10% FBS and, subsequently, each of the various *ZIP4* forms transiently transfected using polyethyleneimine as transfection reagent (PolySciences).

Immunodetection

For the cell surface expression experiments, live cells were incubated with anti HA (1:1000) in DMEM without serum for 1h at 37º before fixation with 4% paraformaldehide. After blocking for 30min (1% BSA, 2% FBS in PBS), cells were incubated with a secondary antibody (1:2000) for 45 min in the blocking solution. For the total cell expression experiments, cells were permeabilized with 1% Triton in PBS for 10min after fixation. Following blocking for 30min (1% BSA, 2% FBS in PBS), cells were incubated in the blocking solution with anti HA (1:1000) for 1h 30min, washed with PBS and incubated with the secondary antibody (1:2000) for 45min. Images were acquired using an inverted Leica SP2 confocal microscope with a 40 x 1.32 Oil Ph3 CS objective. Expression was quantified by measuring chemiluminiscence with a plate reader (24-well plates) using peroxidase-linked anti-mouse antibody (GE Healthcare) as a secondary antibody and SuperSignal West Femto reagent as a substrate (Thermo scientific). Data are presented as the ratio between surface expression and total expression of the transporter. Statistical significance was tested using standard ANOVA.

Zinc Measurements

Cells were transiently transfected with the various *ZIP4* isoforms plus empty ECFP vector for 24-36h. Cytosolic Zn^{2+} signal was determined in CFP-positive cells loaded with FluoZin3 2.5µM (Invitrogen) in a solution containing 140 mM NaCl, 5 mM KCl, 1.2 mM $CaCl_2$, 0.5 mM $MgCl_2$, 5 mM glucose, 10 mM HEPES, 300 mosmol/l, pH 7.4 for 20min. Cytosolic [Zn^{2+}] increases are presented as the difference with respect to the basal signal of emitted fluorescence (510 nm) after adding 200 µM $ZnSO_4$ in a continuous perfusion bath. The kinetics of the various isoforms were calculated using a sigmoidal non-linear regression. In the same set of experiments, basal cellular Zn^{2+} content was estimated as the difference in FluoZin intensity between transfected cells and non-transfected cells before adding Zn^{2+} to the bath. Flourescence intensity was measured using an Olympus IX70 inverted fluorescence microscope, controlled by Aquacosmos software (Hamamatsu).

ACKNOWLEDGMENTS

We thank Jean-Marie Hombert and Lolke van der Veen (Dynamique du Langage, Institut des Sciences De l'Homme, Lyon France) as well as Patrick Mouguiama Daouda (University Omar Bongo, Libreville, Gabon) and the Centre International des Recherches Médicales de Franceville (CIRMF, Gabon) for sharing the Gabon samples. We also acknowledge Karima Fadhlaoui-Zid from the University Tunis El Manar (Tunisia) for providing the Libyan samples as well as the National Institute of Bioinformatics (http://www.inab.org), Nino Spataro for computational support and Marc Torres Ciuró for graphical design help.

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., et al. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, *7*(4), 248–9.

Alam, S., & Kelleher, S. L. (2012). Cellular mechanisms of zinc dysregulation: a perspective on zinc homeostasis as an etiological factor in the development and progression of breast cancer. *Nutrients*, *4*(8), 875–903.

Barreiro, L. B., Laval, G., Quach, H., Patin, E., & Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature genetics*, *40*(3), 340–5.

Beker Aydemir, T., Chang, S.-M., Guthrie, G. J., Maki, A. B., Ryu, M.-S., Karabiyik, A., & Cousins, R. J. (2012). Zinc transporter ZIP14 functions in hepatic zinc, iron and glucose homeostasis during the innate immune response (endotoxemia). (P. V. Nerurkar, Ed.)*PloS one*, *7*(10), e48679.

Beleza, S., Santos, A. M., McEvoy, B., Alves, I., Martinho, C., Cameron, E., Shriver, M. D., et al. (2013). The timing of pigmentation lightening in Europeans. *Molecular biology and evolution*, *30*(1), 24–35.

Berniell-Lee, G., Calafell, F., Bosch, E., Heyer, E., Sica, L., Mouguiama-Daouda, P., Van der Veen, L., et al. (2009). Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. *Molecular biology and evolution*, *26*(7), 1581–9.

Briggs, A. W., Good, J. M., Green, R. E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C., et al. (2009). Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*, *325*(5938), 318–21.

Cann, H. M., De Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., et al. (2002). A human genome diversity cell line panel. *Science*, *296*(5566), 261–2.

Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. E., Livingston, R. J., Rieder, M. J., & Nickerson, D. A. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. *Genome research*, *15*(11), 1553–65.

Cole, C. R., Grant, F. K., Swaby-ellis, E. D., Smith, J. L., Jacques, A., Northrop-clewes, C. A., Caldwell, K. L., et al. (2010). Zinc and iron deficiency and their interrelations in low-income African American and Hispanic children in Atlanta 1 – 4. *American Journal of Clinical Nutrition*, (C).

Dufner-Beattie, J., Wang, F., Kuo, Y.-M., Gitschier, J., Eide, D., & Andrews, G. K. (2003). The acrodermatitis enteropathica gene ZIP4 encodes a tissue-specific, zinc-regulated zinc transporter in mice. *The Journal of biological chemistry*, *278*(35), 33474–81.

Excoffier, L., & Ray, N. (2008). Surfing during population expansions promotes genetic revolutions and structuration. *Trends in ecology & evolution*, *23*(7), 347–51.

Fenner, J. N. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American journal of physical anthropology*, *128*(2), 415–23.

Fortea, J., De la Rasilla, M., García-Tabernero, A., Gigli, E., Rosas, A., & Lalueza-Fox, C. (2008). Excavation protocol of bone remains for Neandertal DNA analysis in El Sidrón Cave (Asturias, Spain). *Journal of human evolution*, *55*(2), 353–7.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–61.

Gardner, M., Williamson, S., Casals, F., Bosch, E., Navarro, A., Calafell, F., Bertranpetit, J., et al. (2007). Extreme individual marker F(ST)values do not imply population-specific selection in humans: the NRG1 example. *Human genetics*, *121*(6), 759–62.

Geiser, J., Venken, K. J. T., De Lisle, R. C., & Andrews, G. K. (2012). A Mouse Model of Acrodermatitis Enteropathica: Loss of Intestine Zinc Transporter ZIP4 (Slc39a4) Disrupts the Stem Cell Niche and Intestine Integrity. *PLoS Genet*, *8*(6), e1002766.

Granka, J. M., Henn, B. M., Gignoux, C. R., Kidd, J. M., Bustamante, C. D., & Feldman, M. W. (2012). Limited evidence for classic selective sweeps in African populations. *Genetics*, *192*(3), 1049–64.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., et al. (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710–22.

Grossman, S. R., Andersen, K. G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D. J., et al. (2013). Identifying recent adaptations in large-scale genomic data. *Cell*, *152*(4), 703–13.

Grossman, S. R., Shlyakhter, I., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science 327*(5967), 883–6.

Haase, H., & Rink, L. (2009). Functional significance of zinc-related signaling pathways in immune cells. *Annual review of nutrition*, *29*, 133–52.

Hamblin, M T, & Di Rienzo, A. (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *American journal of human genetics*, *66*(5), 1669–79.

Hamblin, Martha T, Thompson, E. E., & Di Rienzo, A. (2002). Complex signatures of natural selection at the Duffy blood group locus. *American journal of human genetics*, *70*(2), 369–83.

Henn, B. M., Botigué, L. R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J. K., Fadhlaoui-Zid, K., et al. (2012). Genomic ancestry of North Africans supports back-to-Africa migrations. (M. H. Schierup, Ed.)*PLoS genetics*, *8*(1), e1002397.

Ho, E., & Ames, B. N. (2002). Low intracellular zinc induces oxidative DNA damage, disrupts p53, NFkappa B, and AP1 DNA binding, and affects DNA repair in a rat glioma cell line. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(26), 16770–5.

Hofer, T, Ray, N., Wegmann, D., & Excoffier, L. (2009). Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Annals of human genetics*, *73*(1), 95–108.

Hofer, Tamara, Foll, M., & Excoffier, L. (2012). Evolutionary forces shaping genomic islands of population differentiation in humans. *BMC genomics*, *13*(1), 107.

Hood, M. I., & Skaar, E. P. (2012). Nutritional immunity: transition metals at the pathogen-host interface. *Nature reviews. Microbiology*, *10*(8), 525–37.

Hoque, K. M., Rajendran, V. M., & Binder, H. J. (2005). Zinc inhibits cAMP-stimulated Cl secretion via basolateral K-channel blockade in rat ileum. *American journal of physiology. Gastrointestinal and liver physiology*, *288*(5), G956–63.

Jansen, J., Karges, W., & Rink, L. (2009). Zinc and diabetes--clinical links and molecular mechanisms. *The Journal of nutritional biochemistry*, *20*(6), 399–417.

Kehl-Fie, T. E., & Skaar, E. P. (2010). Nutritional immunity beyond iron: a role for manganese and zinc. *Current opinion in chemical biology*, *14*(2), 218–24.

Kochan, I. (1973). The role of iron in bacterial infections, with special consideration of host-tubercle bacillus interaction. *Current topics in microbiology and immunology*, *60*, 1–30.

Krause, J., Lalueza-Fox, C., Orlando, L., Enard, W., Green, R. E., Burbano, H. A., Hublin, J.-J., et al. (2007). The derived FOXP2 variant of modern humans was shared with Neandertals. *Current biology : CB*, *17*(21), 1908–12.

Küry, S., Dréno, B., Bézieau, S., Giraudet, S., Kharfi, M., Kamoun, R., & Moisan, J.-P. (2002). Identification of SLC39A4, a gene involved in acrodermatitis enteropathica. *Nature genetics*, *31*(3), 239–40.

Lalueza-Fox, C., Gigli, E., De la Rasilla, M., Fortea, J., & Rosas, A. (2009). Bitter taste perception in Neanderthals through the analysis of the TAS2R38 gene. *Biology letters*, *5*(6), 809–11.

Lalueza-Fox, C., Römpler, H., Caramelli, D., Stäubert, C., Catalano, G., Hughes, D., Rohland, N., et al. (2007). A melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals. *Science*, *318*(5855), 1453–5.

Li, C.-R., Yan, S.-M., Shen, D.-B., Li, Q., Shao, J.-P., Xue, C.-Y., & Cao, Y.-H. (2010). One novel homozygous mutation of SLC39A4 gene in a Chinese patient with acrodermatitis enteropathica. *Archives of dermatological research*, *302*(4), 315–7.

Maricic, T., Günther, V., Georgiev, O., Gehre, S., Curlin, M., Schreiweis, C., Naumann, R., et al. (2012). A Recent Evolutionary Change Affects a Regulatory Element in the Human FOXP2 Gene. *Molecular biology and evolution*.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, *338*(6104), 222–6.

Moynahan, E. J. (1974). ACRODERMATITIS ENTEROPATHICA: A LETHAL INHERITED HUMAN ZINC-DEFICIENCY DISORDER. *The Lancet*, *304*(7877), 399–400. d

Neldner, K. H., & Hambidge, K. M. (1975). Zinc therapy of acrodermatitis enteropathica. *The New England journal of medicine*, *292*(17), 879–82.

Ng, P. C. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814.

Osier, M. V, Cheung, K.-H., Kidd, J. R., Pakstis, A. J., Miller, P. L., & Kidd, K. K. (2002). ALFRED: An allele frequency database for anthropology. *American journal of physical anthropology*, *119*(1), 77–83.

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome research*, *19*(5), 826–37.

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

Wang, F., Kim, B.-E., Dufner-Beattie, J., Petris, M. J., Andrews, G., & Eide, D. J. (2004). Acrodermatitis enteropathica mutations affect transport activity, localization and zinc-responsive trafficking of the mouse ZIP4 zinc transporter. *Human molecular genetics*, *13*(5), 563–71.

Wang, K., Zhou, B., Kuo, Y.-M., Zemansky, J., & Gitschier, J. (2002). A novel member of a zinc transporter family is defective in acrodermatitis enteropathica. *American journal of human genetics*, *71*(1), 66–73.

Weinberg, E. D. (1977). Infection and iron metabolism. *The American journal of clinical nutrition*, *30*(9), 1485–90.

Weir, B. S., & Hill, W. G. (2002). Estimating F-statistics. *Annual review of genetics*, *36*, 721–50.

Wu, C., Li, D., Jia, W., Hu, Z., Zhou, Y., Yu, D., Tong, T., et al. (2013). Genome-wide association study identifies common variants in SLC39A6 associated with length of survival in esophageal squamous-cell carcinoma. *Nature genetics*, *advance on*.

Xue, Y., Zhang, X., Huang, N., Daly, A., Gillson, C. J., Macarthur, D. G., Yngvadottir, B., et al. (2009). Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. *Genetics*, *183*(3), 1065–77.

FUNDING

JE was supported through a Postdoc scholarship from the Volkswagenstiftung (Az: I/85 198). This work was partially funded by grants BFU2008-01046, SAF2011-29239, SAF2012-38140, CGL2012-36682 and SAF2010-16725 from the Spanish Ministry of Economy and Competitiveness, Fondo de Investigación Sanitaria (Red HERACLES RD12/0042/0014), and FEDER Funds as well as by grants 2009SGR-1101 and 2009SGR-1369 from Direcció General de Recerca, Generalitat de Catalunya. El Sidrón research was also supported by Consejería de Cultura del Principado de Asturias. MAV is the recipient of an ICREA Academia Award. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Figure Legends

Figure 1. Extreme population differentiation of the Leu372Val polymorphism in *ZIP4* (*SLC39A4*). (A) Distribution of F_{ST} pairwise scores between CEU and YRI in SNPs from the 1000 Genomes data, plotted on a linear and logarithmic scale. The F_{ST} values and corresponding quantiles of the five common SNPs in ZIP4 are indicated. *The F_{ST} quantile of rs1871534 (Leu372Val) was 0.99999977. (B) Distribution of F_{ST} pairwise scores between CHB and YRI in SNPs from the 1000 Genomes data. The F_{ST} values and corresponding quantiles of the five common SNPs in ZIP4 are indicated. **The F_{ST} quantile of rs1871534 (Leu372Val) was 0.99999817. (C) Contour map of worldwide frequencies of the Val372 variant at the rs1871534 SNP. A complete list of populations and allele frequencies is available in Supplemental Table S1.

Figure 2. Human *ZIP4* sequence alignment with chimpanzee and archaic hominids. Archaic hominid sequences are shown from one Denisovan individual and four Neanderthal individuals (three from Vindija and one from El Sidrón which we resequenced for the present study). The orthologous positions of the human rs1871534 and rs227262 SNPs are shown in orange and black, respectively. Note that the reference sequence displayed at the bottom of the figure is the reverse complement of the *SLC39A4* coding sequence. This analysis shows that the *ZIP4* reference sequence carries the ancestral allele shared with archaic hominids and chimpanzee at these two locations.

Figure 3. Genomic context and patterns of selection in a 10 kb region around *ZIP4* (*SLC39A4*). (A) Structure of the human *ZIP4* (*SLC39A4*) gene. (B) F_{ST} scores between YRI and CEU in a 10 kb window centered in the Leu372Val polymorphism (rs1871534). The three indicated SNPs show F_{ST} scores above the 99th percentile (indicated with a black line) of the corresponding

genome-wide F_{ST} distribution. (C) Recombination landscape in YRI. (D) Distributions of diverse neutrality and population differentiation statistics based on coalescent simulations (Schaffner et al. 2005) carried out under neutrality and different selection scenarios. The presence of a recombination hotspot of moderate strength reduced all signals of positive selection in the neutrality tests except for population differentiation. Overall, the observed values (averaged in a 10 kb window around *ZIP4*; green line) are compatible with a moderate selection coefficient (0.5%) simulated under the observed recombination landscape. (E) Coalescent simulations assuming different selection coefficients. Here, the recombination landscape was fixed to the observed landscape in YRI and we tested different selection coefficients (3%, 1%, 0.5% and 0%, the latter corresponding to neutrality). As expected, stronger selection coefficients yielded increasingly stronger deviations from neutrality.

Figure 4. Sequence conservation and clinical relevant variation around the 372 ZIP4 position. (A) Sequence conservation across the vertebrate species tree and the sister protein families SLC39A4 (ZIP4) and SLC39A12 (ZIP12). The highly conserved position Leu372 and the less conserved position Thr357 are indicated. Sequences were downloaded from PhylomeDB (Huerta-Cepas et al. 2011) and aligned in T-Coffee (Notredame, Higgins, and Heringa 2000). (B) Human amino acid variation around the 372 position in acrodermatitis patients and healthy individuals.

Figure 5. Val372 shows reduced membrane surface expression. (A) Immunostaining of the different isoforms of ZIP4 in HeLa cells under permeabilizing (left) conditions for total protein visualization and non-permeabilizing (right) conditions for surface protein visualization. Bright field images are provided to show the presence of cells on the field in all conditions. The acrodermatitis enteropathica variants Leu372Pro and Leu372Arg show absence of membrane expression. (B) Surface expression quantification

normalized by the total amount of transporter of the different ZIP4 isoforms obtained from 12 independent measurements obtained in 4 different transfections. Data are expressed as mean ± SEM. * P<0.05 and *** P<0.001 *vs* Ala357-Leu372, one way ANOVA. The two isoforms expressing Val372 show reduced surface expression compared to the Leu372 isoforms.

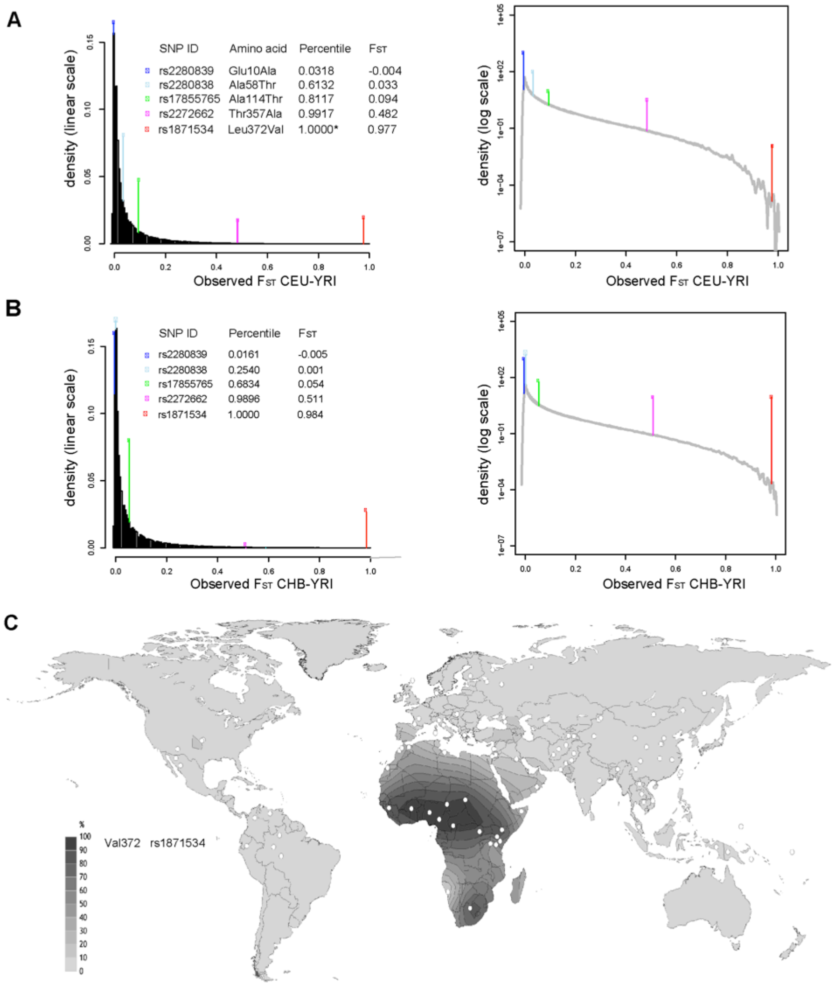
Figure 6. Val372 shows reduced zinc uptake transport. (A) Basal zinc content in HeLa cells transiently transfected with different ZIP4 isoforms plus empty CFP vector. Transfected cells were compared with surrounding non-transfected cells. The two isoforms expressing Val372 show significantly reduced intracellular zinc. (B) Zinc uptake upon perfusion with 200 µM $ZnSO_4$ external solution. Graph bars show the maximum transport (T_{max}) and the time to reach half of T_{max} ($t_{1/2}$) for the different isoforms that reach the plasma membrane, following the color code on the left. Data are presented as mean ± SEM of 3 different transfections and more than 25 cells per condition. Significance was calculated using ANOVA, with the Ala357-Leu372 isoform as reference (*p<0.05, **p<0.01, ***p<0.001). The two isoforms expressing Val372 show reduced T_{max} but no difference in $t_{1/2}$ when compared to the Leu372 isoforms.

Table 1. Common non-synonymous SNPs in the *ZIP4* (*SLC39A4*) gene

	DAF			**F_{ST}**			**Amino Acid**	**Functional Prediction**	
SNP ID[a]	**CEU**	**CHB**	**YRI**	**CEU-YRI**	**CHB-CEU**	**YRI-CHB**	**Replacement**	**PolyPhen**	**SIFT**
rs2280839	0.4812	0.4485	0.4545	-0.0042	-0.0030	-0.005	Glu10Ala	Benign	Deleterious
rs2280838	0.5235	0.4347	0.3803	0.0335	0.0080	0.0006	Ala58Thr	Benign	Tolerated
rs17855765	0.4812	0.4278	0.2601	0.0942	-0.0002	0.0535	Ala114Thr	Benign	Tolerated
rs2272662	0.4941	0.5401	0.0057	**0.4823**	-0.0005	**0.5110**	Thr357Ala	Benign	Tolerated
rs1871534	0.0059	0.0000	0.9787	**0.9765**	0.0002	**0.9837**	Leu372Val	PD	Deleterious

[a] Reported non-synonymous SNPs with minor allele frequencies (MAF) greater than 0.10 in any of the three 1000 Genomes Project populations. Values in bold are above the 99th percentile of the each corresponding F_{ST} genome-wide distribution among the two compared populations. Abbreviations: DAF, derived allele frequency; PD, Probably damaging.

Figure 1.



	SNP ID	Amino acid	Percentile	F_{ST}
■	rs2280839	Glu10Ala	0.0318	-0.004
■	rs2280838	Ala58Thr	0.6132	0.033
■	rs17855765	Ala114Thr	0.8117	0.094
■	rs2272662	Thr357Ala	0.9917	0.482
■	rs1871534	Leu372Val	1.0000*	0.977

	SNP ID	Percentile	F_{ST}
■	rs2280839	0.0161	-0.005
■	rs2280838	0.2540	0.001
■	rs17855765	0.6834	0.054
■	rs2272662	0.9896	0.511
■	rs1871534	1.0000	0.984

Figure 2.

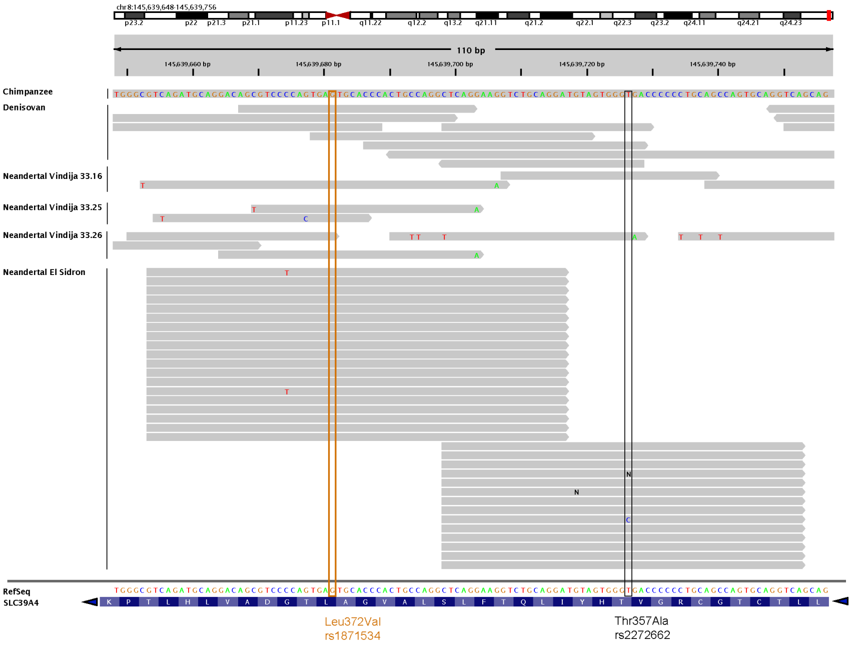

chr8:145,639,648-145,639,756
p23.2
p22
p21.3
p21.1
p11.23
p11.1
q11.22
q12.2
q13.2
q21.11
q21.2
q22.1
q22.3
q23.2
q24.11
q24.21
q24.23
110 bp
145,639,660 bp
145,639,680 bp
145,639,700 bp
145,639,720 bp
145,639,740 bp
Chimpanzee
Denisovan
Neandertal Vindija 33.16
Neandertal Vindija 33.25
Neandertal Vindija 33.26
Neandertal El Sidron
RefSeq
SLC39A4
Leu372Val
rs1871534
Thr357Ala
rs2272662

Figure 3.

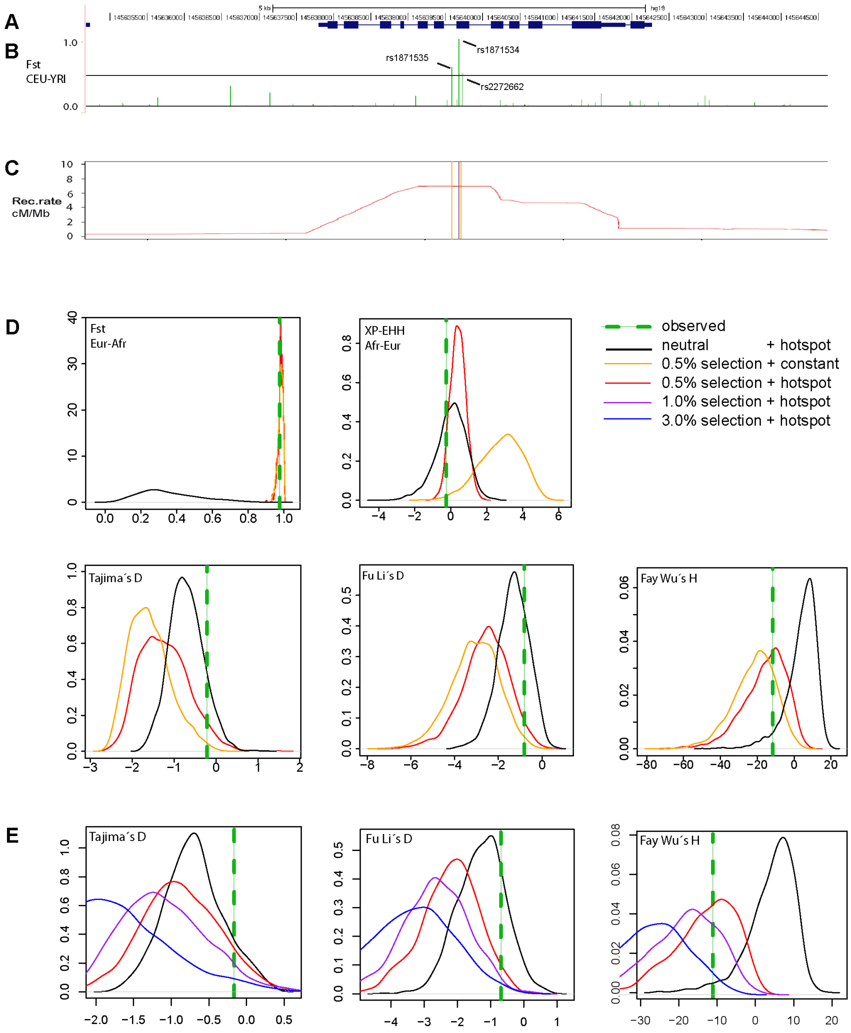
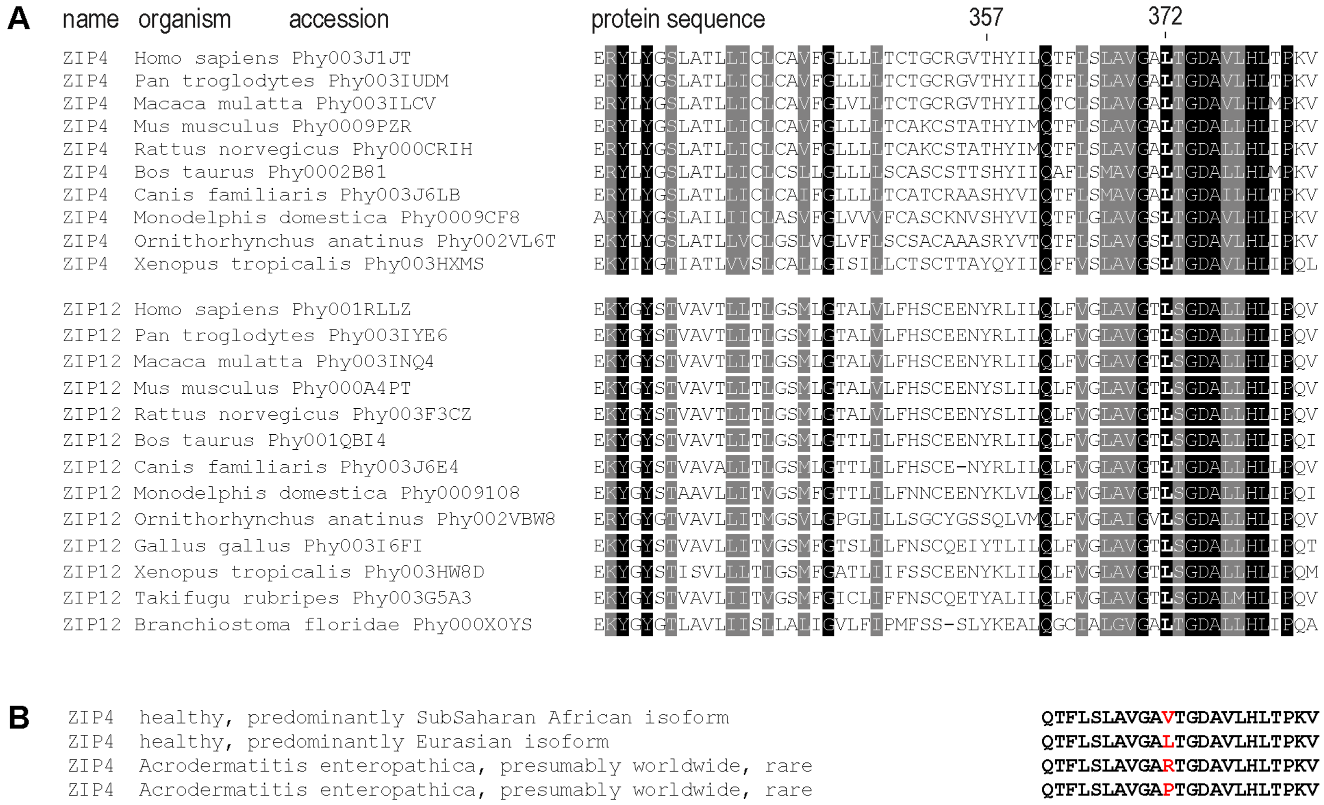

A
5 kb
hg19
145635500
145636000
145636500
145637000
145637500
145638000
145638500
145638000
145639500
145640000
145640500
145641000
145641500
145642000
145642500
145643000
145643500
145644000
145644500
B
Fst
CEU-YRI
1.0
0.0
rs1871534
rs1871535
rs2272662
C
Rec.rate
cM/Mb
10
8
6
4
2
0
D
Fst
Eur-Afr
XP-EHH
Afr-Eur
observed
neutral + hotspot
0.5% selection + constant
0.5% selection + hotspot
1.0% selection + hotspot
3.0% selection + hotspot
Tajima´s D
Fu Li´s D
Fay Wu´s H
E
Tajima´s D
Fu Li´s D
Fay Wu´s H

Figure 4.



A

name	organism	accession	protein sequence
ZIP4	Homo sapiens	Phy003J1JT	ERYLYGSLATLLICLCAVFGLLLLTCTGCRGVTHYILQTFLSLAVGALTGDAVLHLTPKV
ZIP4	Pan troglodytes	Phy003IUDM	ERYLYGSLATLLICLCAVFGLLLLTCTGCRGVTHYILQTFLSLAVGALTGDAVLHLTPKV
ZIP4	Macaca mulatta	Phy003ILCV	ERYLYGSLATLLICLCAVFGLVLLTCTGCRGVTHYILQTCLSLAVGALTGDAVLHLMPKV
ZIP4	Mus musculus	Phy0009PZR	ERYLYGSLATLLICLCAVFGLLLLTCAKCSTATHYIMQTFLSLAVGALTGDALLHLIPKV
ZIP4	Rattus norvegicus	Phy000CRIH	ERYLYGSLATLLICLCAVFGLLLLTCAKCSTATHYIMQTFLSLAVGALTGDALLHLIPKV
ZIP4	Bos taurus	Phy0002B81	ERYLYGSLATLLICLCSLLGLLLLSCASCSTTSHYIIQAFLSMAVGALTGDALLHLMPKV
ZIP4	Canis familiaris	Phy003J6LB	ERYLYGSLATLLICLCAIFGLLLLTCATCRAASHYVIQTFLSMAVGALTGDAILHLTPKV
ZIP4	Monodelphis domestica	Phy0009CF8	ARYLYGSLAILIICLASVFGLVVVFCASCKNVSHYVIQTFLGLAVGSLTGDAVLHLIPKV
ZIP4	Ornithorhynchus anatinus	Phy002VL6T	EKYLYGSLATLLVCLGSLVGLVFLSCSACAAASRYVTQTFLSLAVGSLTGDAVLHLIPKV
ZIP4	Xenopus tropicalis	Phy003HXMS	EKYIYGTIATLVVSLCALLGISILLCTSCTTAYQYIIQFFVSLAVGSLTGDAVLHLIPQL
ZIP12	Homo sapiens	Phy001RLLZ	EKYGYSTVAVTLLTLGSMLGTALVLFHSCEENYRLILQLFVGLAVGTLSGDALLHLIPQV
ZIP12	Pan troglodytes	Phy003IYE6	EKYGYSTVAVTLLTLGSMLGTALVLFHSCEENYRLILQLFVGLAVGTLSGDALLHLIPQV
ZIP12	Macaca mulatta	Phy003INQ4	EKYGYSTVAVTLLTLGSMLGTALVLFHSCEENYRLILQLFVGLAVGTLSGDALLHLIPQV
ZIP12	Mus musculus	Phy000A4PT	EKYGYSTVAVTLLTLGSMLGTALVLFHSCEENYSLILQLFVGLAVGTLSGDALLHLIPQV
ZIP12	Rattus norvegicus	Phy003F3CZ	EKYGYSTVAVTLLTLGSMLGTALVLFHSCEENYSLILQLFVGLAVGTLSGDALLHLIPQV
ZIP12	Bos taurus	Phy001QBI4	EKYGYSTVAVTLLTLGSMLGTTLILFHSCEENYRLILQLFVGLAVGTLSGDALLHLIPQI
ZIP12	Canis familiaris	Phy003J6E4	EKYGYSTVAVALLTLGSMLGTTLILFHSCE-NYRLILQLFVGLAVGTLTGDALLHLLPQV
ZIP12	Monodelphis domestica	Phy0009108	EKYGYSTAAVLLITVGSMFGTTLILFNNCEENYKLVLQLFVGLAVGTLSGDALLHLIPQI
ZIP12	Ornithorhynchus anatinus	Phy002VBW8	ERYGYGTVAVLLITMGSVLGPGLILLSGCYGSSQLVMQLFVGLAIGVLSGDALLHLIPQV
ZIP12	Gallus gallus	Phy003I6FI	EKYGYSTVAVLLITVGSMFGTSLILFNSCQEIYTLILQLFVGLAVGTLSGDALLHLIPQT
ZIP12	Xenopus tropicalis	Phy003HW8D	EKYGYSTISVLLLTIGSMFGATLIIFSSCEENYKLILQLFVGLAVGTLSGDALLHLIPQM
ZIP12	Takifugu rubripes	Phy003G5A3	EKYGYSTVAVLIITVGSMFGICLIFFNSCQETYALILQLFVGLAVGTLSGDALMHLIPQV
ZIP12	Branchiostoma floridae	Phy000X0YS	EKYGYGTLAVLIISLLALIGVLFIPMFSS-SLYKEALQGCIALGVGALTGDALLHLIPQA

B

ZIP4	healthy, predominantly SubSaharan African isoform	**QTFLSLAVGAVTGDAVLHLTPKV**
ZIP4	healthy, predominantly Eurasian isoform	**QTFLSLAVGALTGDAVLHLTPKV**
ZIP4	Acrodermatitis enteropathica, presumably worldwide, rare	**QTFLSLAVGARTGDAVLHLTPKV**
ZIP4	Acrodermatitis enteropathica, presumably worldwide, rare	**QTFLSLAVGAPTGDAVLHLTPKV**

Figure 5.

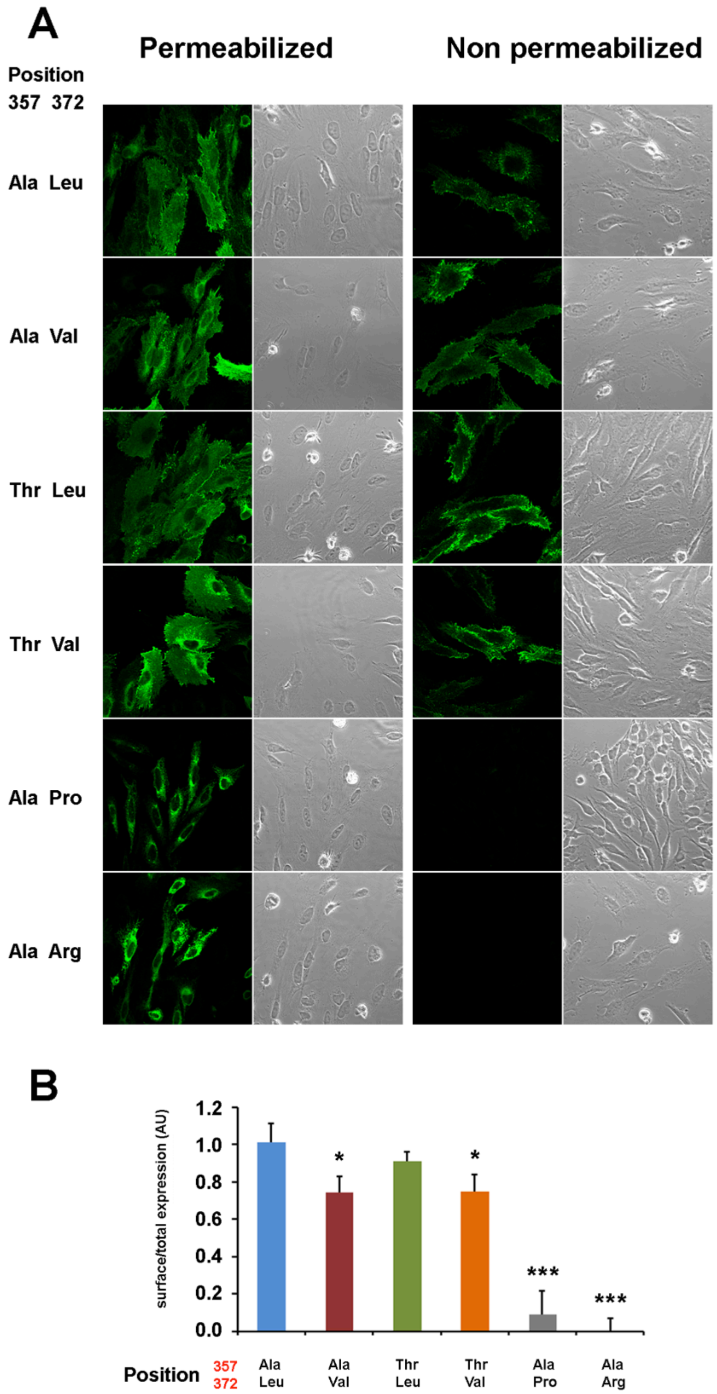
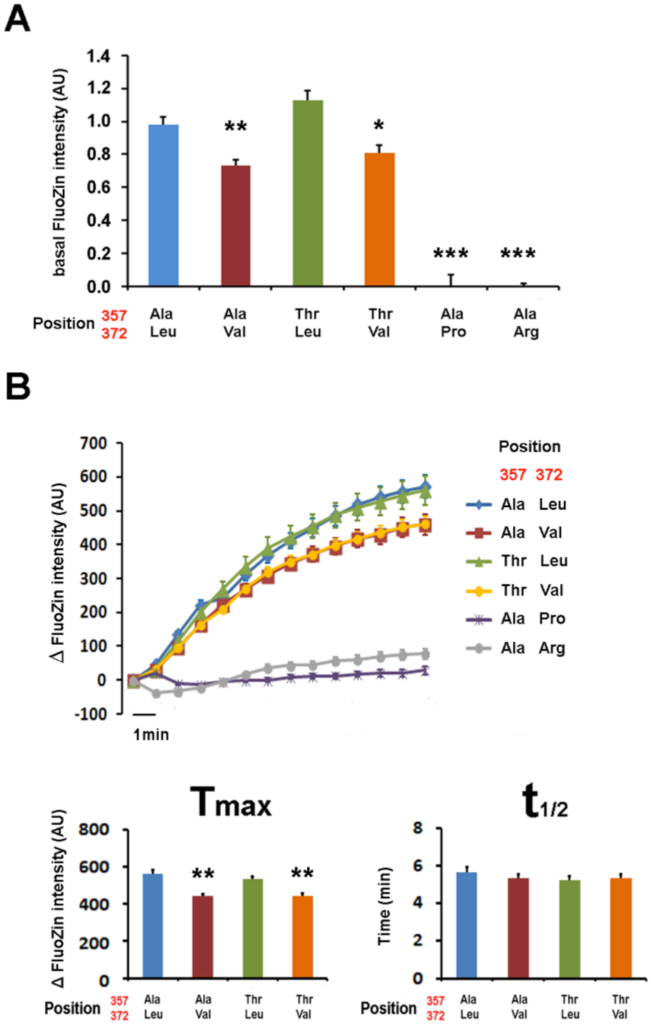

A
Permeabilized
Non permeabilized
Position
357 372
Ala Leu
Ala Val
Thr Leu
Thr Val
Ala Pro
Ala Arg
B
surface/total expression (AU)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
*
*
***
***
Position
357
372
Ala
Leu
Ala
Val
Thr
Leu
Thr
Val
Ala
Pro
Ala
Arg

Figure 6.


A
basal FluoZin intensity (AU)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
**
*
***
***
Position
357
372
Ala Leu
Ala Val
Thr Leu
Thr Val
Ala Pro
Ala Arg
B
Δ FluoZin intensity (AU)
700
600
500
400
300
200
100
0
-100
1min
Position
357 372
Ala Leu
Ala Val
Thr Leu
Thr Val
Ala Pro
Ala Arg
Tmax
Δ FluoZin intensity (AU)
800
600
400
200
0
**
**
Position
357
372
Ala Leu
Ala Val
Thr Leu
Thr Val
t1/2
Time (min)
8
6
4
2
0
Position
357
372
Ala Leu
Ala Val
Thr Leu
Thr Val

Chapter 3

The role of natural selection on coding and non-coding genomic elements: analysis of conserved and accelerated pathways combining divergence and polymorphism data on 20 chimpanzees

Elena Carnero-Montoro[1*], Gabriel Santpere[1*], Natalia Petit[1*], François Serra[2], Christina Hvilsom[3], Daniel L. Halligan[4], Hernan Dopazo[2], Arcadi Navarro[1,5,6, γ], Elena Bosch[1,γ]

(Manuscript in preparation)

[1] IBE, Institute of Evolutionary Biology (Universitat Pompeu Fabra-CSIC), PRBB,
Doctor Aiguader, 88, 08003, Barcelona, Catalonia, Spain
[2] Evolutionary Genomics Lab, Bioinformatics & Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain
[3] Science and Conservation, Copenhagen Zoo, 2000 Frederiksberg, Denmark
[4] Institute of Evolutionary Biology, University of Edinburgh, West Mains Rd, Edinburgh EH9 3JY, Uk
[5] National Institute for Bioinformatics (INB), PRBB, Doctor Aiguader, 88, 08003,
Barcelona, Catalonia, Spain
[6] Institució Catalana de Recerca i Estudis Avançats (ICREA), PRBB, Doctor Aiguader,
88, 08003, Barcelona, Catalonia, Spain

[*] Equal contribution
[γ] Corresponding authors

KEYWORDS: chimpanzee, pathways, natural selection, site frequency spectrum, distribution of fitness effects, fraction of adaptive substitution (α) and adaptive substitution rate (ω_α).

BACKGROUND

Differences accumulated between species can result either from neutral mutations that have been fixed by genetic drift, or from adaptive mutations that have been driven to fixation by positive selection. On the contrary, deleterious mutations are assumed to rarely become fixed. Divergence tests that compare the relative proportion of putatively functional to non-functional (and therefore putatively neutral) substitutions allow to infer the type of selection pressures acting in the putatively functional sites. Recently, these methods have been applied in a pathway-level fashion and have revealed different modes of selection acting over different sets of functionally interconnected genes (Serra et al. 2011). However, these methods cannot clearly distinguish whether an excess of functional fixed mutations is due to the relaxation of purifying selection or to positive selection.

The McDonald-Kreitman test (MKT) compares diversity within a species with divergence between species at two types of sites, namely functional (usually non-synonymous changes) and putatively neutral sites (usually synonymous sites) (McDonald and Kreitman 1991). If the non-synonymous (or putatively functional) variation is neutral, the ratio of functional to neutral variation in divergent sites among species is expected to be equal to the same ratio in polymorphic positions within a species. On the contrary, if the ratio of putatively functional to neutral fixed mutations is significantly higher than the ratio of functional to neutral polymorphisms (i.e. there is an excess of functional divergence) this is taken as evidence for directional positive selection. Importantly, the test also allows quantifying the proportion of functional adaptive fixed differences (α) and the rate at which new adaptive mutations appear (ω_α) (Eyre-walker 2006).

In the last years, a number of studies have yielded important clues about the relative importance of adaptive evolution in shaping molecular evolution in different species by the study of candidate genes, or by genome-wide analysis. Estimates range from low values in yeast (Elyashiv et al. 2010) or humans (Bustamante et al. 2005)(Boyko et al. 2008) to high values in *Drosophila* (Welch 2006) (Sella et al. 2009) or *Escherichia coli* (Charlesworth and Eyre-Walker 2006). Importantly, variability in the α values among species also reflects, at least partially, differences in their effective population sizes, as the proportion of adaptive substitutions tends to be higher in those species with larger population sizes. This observation is consistent with the stronger expected effectiveness of selection relative to drift, as the effective population size of a given species increases. However, the variability of α values sometimes reported for a given species may also reflect the particularities of the different methodologies employed in each study.

Because it uses interdigitated sites, the MKT is robust to variation in mutation rates or to possible different coalescence histories at the different genomic locations analyzed. However, other confounding factors such as the presence of slightly deleterious mutations or particular demographic events that could significantly contribute to the amount of polymorphism may affect the test (Justin C Fay 2011a). Recently, a number of extensions of the MKT have been proposed to correct the possible biases of these confounding factors when trying to estimate the proportion of adaptive evolution. These different modifications mainly use and combine in diverse ways the site frequency spectrum (SFS) of neutral polymorphisms to correct for the effect of demography, and the distribution of fitness effects (DFE) at functional polymorphic sites to correct for the possible presence of slightly deleterious mutations contributing to the polymorphic variation (Eyre-Walker and Keightley 2007) (Keightley and Eyre-Walker 2007) (Boyko et al. 2008)

Interestingly, the contribution of non-coding elements to adaptive evolution has been shown to be higher than that of protein-coding genes, at least in mice and *Drosophila* (Kousathanas et al. 2011)(Mackay et al. 2012b). However, some difficulties arise when applying MKT derivative tests to non-coding elements due to the need of choosing a proper neutral evolving reference to which compare the levels of divergence and polymorphism found in them. Although non-regulatory intronic sequences, synonymous sites and ancestral repeat sequences have been proved to generally evolve neutrally, natural selection could sometimes target them or distort their SFS (Lawrie et al. 2013)(Fernando O, Navarro A, in prep); thus, special consideration needs to be taken in the choice of a neutral reference.

To date, MKT methods have not been applied in a pathway fashion. Such an approach would allow disentangling how different functions are subjected to different evolutionary selective pressures and can improve on the analysis of single genes, since normally genes do not act in isolation. Moreover, none of the pathway analyses of natural selection available have incorporated their putative functional non-coding regulating fraction to ask whether trends in mode of selection are shared among coding and non-coding elements in a given pathway or not.

In this study, we aim to investigate differences in evolutionary patterns between functional pathways and among their corresponding coding and non-coding regions (including promoters, UTRs, introns and trailers) in a chimpanzee population of 20 individuals. For this purpose, we combine both divergence and polymorphism data by applying the DFE-alpha extension of the MKT in two reference pathways that are known to present accelerated and slow rates of protein evolution based on divergence data (the Complement and the Actin pathways, respectively), and in three pathways related to neurological function (Parkinson, Amiloid and Presenilin). The effective population size of *Pan troglodytes troglodytes* (19.7– 39.5*10^{-3}) is one of the largest among the great apes, and is around 3 times higher than that of humans (9.7 – 19.5 *10^{-3}) (Prado-Martinez et al. 2013). By using this chimpanzee population in our study, we hope to have greater power to detect significant differences in the fraction of adaptive substitutions and to distinguish different evolutionary trends among the pathways and their functional elements than if a similar approach was carried in the human lineage.

RESULTS AND DISCUSSION

Selection of biological pathways

We have studied the evolution in chimpanzees of coding DNA sequences (CDS), introns and regulatory regions of diverse functional pathways by combining polymorphism data from 20 individuals of that species with divergence patterns with human. To that effect, we initially selected three different datasets that, according to the literature (Serra et al. 2011) and our own results present particular differentiated evolutionary patterns in the chimpanzee lineage. The study of Serra et al. compared humans chimpanzees, rats and mice. The criteria to decide whether a pathway presented accelerated or decelerated evolution in a certain lineage was two-fold: first, w (dN/dS) values where ranked for the whole genome and it was tested whether certain pathways were exceedingly represented in the tails of the w distributions; second, the number of genes whose w values had increased or decreased since the common ancestor of that lineage were computed and a pathway was considered accelerated (decelerated) if an excess of w increments (reductions) was detected. The pathways we used as references for our study were: (i) the Actin pathway, whose protein coding regions have evolved at a slow rate and they have even decelerated since the common ancestor of humans and chimpanzees. That is, 63% of the genes from the hsa04810 pathway “Regulation of actine cytoskeleton” are in the lower extreme of a genome-wide w distribution and 88% of genes present lower w in the chimpanzee lineage than in the common ancestor with humans. (ii) the Complement pathway, which, on the contrary, shows an enrichment of genes with high dN/dS ratios, with 62% of the genes from the hsa04610 pathway “Complement and coagulation cascades” accumulating in the corresponding high extreme of the w distribution. And, (iii) a set of genes with at least one of their introns showing significant divergence patterns of acceleration in a Likelihood Ratio Test similar to that for (Haygood et al. 2007) (see Materials and Methods).

We further extended our analysis to three neurological pathways of interest which are related to the human neurodegenerative diseases of Parkinson’s (PD) and Alzheimer’s (AD): the Parkinson pathway and the Alzheimer-amyloid secretase and the Alzheimer-presenilin pathways, to which we will refer as the “Amyloid” and the “Presenilin” pathways. Interestingly, although both neurological diseases seem to appear exclusively in humans, AD main neuropathological hallmarks (i.e. Aβ and hyper phosphorylated tau deposition) have also been observed in the aging brain of chimpanzees (Rosen et al. 2008, Gearing et al. 1994). Furthermore, despite the apparent chimpanzee resistance to such neurodegenerative diseases, crucial proteins in AD and PD pathogenesis (i.e. APP, precursor of Aβ**,** tau, and alpha-synuclein) are found highly conserved between human and chimpanzees (Rosen et al. 2008) (Holzer et al. 2004) (Hamilton 2004). However, these proteins are found within well-studied neurological networks whose elements might have suffered different evolutionary trends. Neither the PD nor the AD pathways were found to be especially enriched with a significant proportion of genes displaying particularly low or high w values in a genome-wide human-chimpanzee divergence analysis of natural selection on functional modules. However, both neurological pathways presented a significant proportion of genes for which their corresponding dN/dS ratio had decreased in the chimpanzee lineage since the human-chimpanzee common ancestor (Serra et al. 2011). In particular, 14% of the genes from the hsa05010 “Alzheimer’s disease pathway” presented a significant decrease in w from the ancestor, whereas such percentage was of 11% in the case of the hsa05010 “Parkinson’s disease pathway”. All this makes pathways related to AD and PD interesting suitable test pathways to be included in our combined analysis of divergence and polymorphism data.

Sequencing and descriptive statistics

A complete list of genes included in each of the six aforementioned datasets is given in Supplementary Table S1. Their corresponding coding DNA sequences (CDS), introns and regulatory regions were specifically captured and sequenced in 20 Central Chimpanzee individuals (Supplementary Table S2) to an average depth of 72X (see Materials and Methods and Supplementary Tables S3-S4). A total of 79,650 SNPs (with a Ti/Tv ratio = 2.29) were identified a in a final callable region of 8,599,335 bp (see Methods). Next, we calculated descriptive measures of diversity and human-chimpanzee divergence for all six datasets and five differentiated genomic regions (Tables 1 and 2): coding sequences (CDS), introns, promoters, trailers and untranslated regions (UTR).

The rate of neutral evolution at synonymous sites (dS) was very similar among pathways with Amiloid and Presenilin genes marking the extremes of the range at dS=0.32% and 0.44% (Table 1 and Supplementary Table S5), indicating that on average the genes from each dataset are placed in parts of the genome with broadly similar neutral substitution rates. For coding regions, the average non-synonymous to synonymous divergence rate between humans and chimpanzees is around 0.17-0.25 (Kosiol et al. 2008)

(Hvilsom et al. 2011) reflecting the overall effect of purifying selection acting on most coding parts of the genome. Whereas CDS from the Actin and all three neurological pathways presented similar dN/dS values to that genome-wide average range, those from the Complement and the set of genes with Accelerate Intron evolution presented significantly higher dN/dS values but did not differed between themselves (Supplementary Table S5).

For non-coding regions (i.e. introns, trailers, UTRs and promoters) we considered any polymorphic and divergence site as putatively under selection and measured their rate of substitution per site (dPUS equivalent to dN). We then used the synonymous substitution rates in their corresponding CDS (dS) to obtain equivalent estimations of w (in this case dPUS/dS, equivalent to the dN/dS rates in CDS). We proceeded in the same way for polymorphism data (pPUS/pS). As expected, these dPUS/dS values in all non-coding genomic elements were higher than the corresponding dN/dS values in CDS (Table 2 and Supplementary Table S6), a pattern consistent with the action of purifying selection being stronger on the coding regions of the genome. Notably, except the intronic sequences from the Acc Intron dataset, none of the non-coding regions displayed consistently significant differences in their dPUS/dS values.

Thus, overall divergence descriptives are in agreement with our initial rationale for selecting the Actin and Complement pathways as references, respectively, for constrained and accelerated evolution on CDS; and the Acc Intron dataset as reference for acceleration in non-coding regions. Further confirming that idea, the CDS of the Acc Intron dataset displayed significantly higher dN/dS values than the genome-wide average even though genes in that dataset were selected ignoring CDSs and exclusively on the basis of accelerated divergence in at least one of their intronic sequences.

As for polymorphism on CDS both the Complement pathway and the Acc Intron dataset displayed the higher pN/pS values, even if in most cases differences were not significant. Interestingly, both Amiloid and Parkinson CDS show nominally lower pN values than those of the Actin. Moreover, upon comparison with all other pathways, CDS from the Parkinson pathway also displayed significantly reduced pS. This pattern is consistent with the unfolded site frequency spectrum obtained in CDS, which revealed no significant differences among the datasets except for the Parkinson pathway, where a higher proportion of singletons was found at 0-fold sites (Supplemental Figure S3). The ratio of putatively selected to neutral polymorphism in non-coding regions was clearly higher than that observed in CDS without any clear trend among the different datasets, including the intronic sequences of the Acc Intron dataset.

Alpha (α) and omega-alpha (ω_α) estimation

High dN/dS values and/or accelerated divergence patterns could reflect either the signature of past positive selection or the effect of the relaxation of selective constraints. By contrasting the levels of polymorphism and divergence at putatively neutral and functional sites it may be possible to discriminate between these two different scenarios. Moreover, a number of extensions of the classical McDonald-Kreitman test allow taking into account the site frequency spectrum of the current polymorphism and therefore avoid the possible confounding effects of demography when estimating the rate of adaptation. Here, we have applied one of these extensions, the DFE-alpha method (Keightley and Eyre-Walker 2007), to estimate the proportion (α) and rate (ω_α) of adaptive substitutions in each pathway and genomic element analyzed (Table 3 and Supplementary Tables S8 and S10). Briefly, we considered the overall set of 4-fold degenerate sites from all the genes considered in this study as putatively neutral sites to first infer demographic parameters of the chimpanzee population and then used the inferred demography to assess the distribution of fitness effects (DFE) at functional sites as well as to estimate the α and ω_α parameters for each pathway and genomic element analyzed. For consistency, we also used the overall SFS at all 4-fold degenerate sites as the neutral reference against we checked every pathway or set of elements under analysis.

Evolutionary patterns in reference genes and pathways

We first examine the distribution of fitness effects (DFE). As expected under the action of purifying selection the coding sequences of the Actin pathway shows an excess of new strongly deleterious mutations (see Figure 1, Supplementary Table S7). On the contrary, most mutations in the coding regions of the Complement pathway and those in genes with at least one intron with accelerated substitution rates are estimated to be nearly neutral (Figure 1).

We then focus on rates of adaptive substitution. The coding regions of the Complement and the Actin pathway showed significantly higher and lower adaptive substitution rates (ω_α), respectively (Figure 2A, Table 3). Therefore, in what follows we will consider the Actin and Complement pathways, respectively, as purifying selection and positive selection reference sets, using them to benchmark the coding sequences (CDS) of other pathways. For the Complement pathway, after grouping the CDS in three different percentiles (25, 25-75 and 75) according to their inferred dN/dS values (Figure 2B, Supplementary Table S9), we observed that the higher the dN/dS of the CDS the greater the rate of adaptive evolution was. Moreover, for the three Complement percentiles explored, ω_α values were always significantly higher than that of the Actin set, whereas, in the reciprocal comparison, ω_α values of the Actin pathway were significantly lower in the two most divergent complement percentiles but not in the 25 percentile (Supplementary Table S11). This suggests that the CDS in the whole complement pathway are globally accelerated and that the higher adaptive substitution rates observed are not just the result of extreme outlier genes. For the Actin case, no clear relationship was observed between the dN/dS-based quartiles and the corresponding ω_α values (Supplementary Table S9), and although the three different Actin percentiles showed significantly lower adaptive rates in comparison with the Complement, the reciprocal comparison was not significant (Supplementary Table S11).

Recall that we also selected a reference data set with accelerated substitution rates in introns (Acc. Introns), made of genes with at least one of their introns showing significant divergence patterns of acceleration (see Materials and Methods). As shown in Figure 2A, when considering only the positively selected introns in this set (only the accelerated introns themselves, and not the rest of introns in the corresponding gene) we find strongly significant higher adaptive substitution rates (Supplemental Table S8). The pattern maintains when considering all the introns in the genes that present at least one accelerated intron. Their sequences showed significantly higher adaptive substitution rates (ω_α) in comparison with other pathway-based intron categories (Table 3). Although we are possibly diluting the effect of positively introns when considering all the introns of the dataset, we are still able to clearly capture differences in ω_α values. However, the same dilution may have affected the different non-coding regions in other datasets, and, thus, following the same rationale as for the CDS, we will be using this set of accelerated introns as a reference for positive selection in non-coding sequences. This idea is reinforced when considering the introns in this reference data set for which no significant acceleration was detected (i.e. introns without previous evidence of positive selection from divergence data). These also showed significantly higher adaptive rates (ω_α) in comparison with those from each of the pathways analyzed (Figure 2A, Supplemental Table S8). This result indicates that all introns in the accelerated introns data set follow indeed the same trend towards higher ω_α values and that when using divergence data alone, may be not enough statistical power is reached for all individual significant accelerated introns to emerge.

The combination of divergence and polymorphism data in such reference sets clearly reproduces and refines the evolutionary trends of selection previously inferred using only divergence data. CDS selected for constrained divergence patterns (Actin) do show significantly lower α and ω_α values than the CDS set selected for accelerated divergence patterns (Complement). Moreover, divergence patterns of positive selection are confirmed with polymorphism data in chimpanzees not only in CDS but also in introns. Thus, we confirm that accelerated rates of divergence in these specific preselected elements are due to positive selection and not to relaxation of purifying selection.

Other interesting observations are that whereas we found a high dN/dS ratio in the coding regions of the Complement and the Acc Intron dataset, both α and ω_α values were significantly increased only in the complement CDS. The coding regions of the Acc. Intron dataset did not show significantly different α and ω_α values than those from the Actin pathway. These results suggest that while the increased dN/dS ratio in the coding regions of the Complement pathway (w=0.65, Table 1) have been driven by adaptive evolution, in the Acc. intron dataset such elevated (w=0.49, Table 1) ratio reflects the relaxation of purifying selection acting on the CDS of this dataset.

Evolutionary patterns in three neurological pathways

We next examined the patterns of molecular evolution of the Parkinson's and Alzheimer's pathways (the latter subdivided in the Amiloid and Presenilin pathways) following the same steps than above for the reference pathways. Regarding the DFE, both the Amiloid and the Presenilin pathways display a similar excess of strongly deleterious mutations than the purifying selection reference set (Figure 1). However, whereas in Presenilin CDS we observe a similar proportion of strongly deleterious to nearly neutral mutations, the proportion of nearly neutral mutations is clearly lower in the Amiloid pathway. Parkinson's genes present an intriguing DFE with an excess of very deleterious and deleterious mutations. Accordingly, CDS sites in Parkinson present apparently reduced pN and pS values (Table 1) as well as higher proportion of singletons in the SFS when compared to the Actin pathway (Supplementary Figure S3). Such pattern of constrained diversity is suggestive of the recent action of strong purifying upon the Parkinson pathway in chimpanzees.

As seen in Table 3, when comparing the divergence and polymorphism patterns in CDS from these neurological pathways with those from the reference sets of positive and purifying selection some patterns emerge: the adaptive substitution rates (ω_α) in all neurological pathways are significantly lower than that in the positive selection reference set (as those of the Actin and the Acc intron dataset); none of the Alzheimer's pathways are significantly different from the purifying selection set; but on the contrary, the adaptive substitution rate (ω_α) in the Parkinson's pathway is significantly higher than that of the purifying selection set. These results suggest the action of probably long-term purifying selection constricting the patterns of evolution of the coding sequences of the two Alzheimer's related pathways. Notably, Parkinson's CDS seem to have accumulated a significantly larger proportion of adaptive substitutions than the purifying selection reference set. However, no significant differences are found on the CDS ω_α values of Parkinson and the two Alzheimer's pathways. Overall these results indicate that all three neurological pathways have suffered similar divergence evolutionary constraints, but not as extreme as in the purifying selection reference set. Although CDS from the Complement and the Parkinson pathways display higher adaptative rates than the purifying reference dataset, Parkinson's

CDS do show significant differences in ω_α values with the positive selection reference set. Furthermore, when exploring their corresponding SFS and DFE in the chimpanzee population, Parkinson's CDS also present contrasting recent evolutionary histories to the positive selection reference set.

Interestingly, the proportion (α) and rate (ω_α) of adaptative substitutions follow a similar general trend in all the comparisons performed with only one notable exception (Table 3 and Supplementary Table S10). Whereas we detect significant differences in ω_α values between the Complement and the Parkinson's CDS, they do not differ in their corresponding α values, which, in turn, are significantly higher than those of the remaining datasets. Thus, Parkinson's CDS seem to have accumulated a proportion of adaptive substitutions similar to the reference set of positive selection even though they are evolving at a lower evolutionary rate and are now subjected to strong purifying selection in the chimpanzee population. This observation is consistent with the results by (Serra et al. 2011) who found no particular enrichment of genes with low dN/dS ratios in the chimpanzee lineage but that a proportion of them had significantly decreased their corresponding dN/dS ratios in comparison to the common human-chimpanzee ancestor.

Humans are the only animal known to be susceptible to AD and PD, but A-beta deposition and tau phosphorylation, the main hallmarks of AD, have also been reported in the great apes (N. Kimura et al. 2001)(Rosen et al. 2008)(Gearing et al. 1994). In a recent study analysing aging-associated changes in brains from humans (up to 88 years old) and chimpanzees (up to 51 years old), the authors found significantly higher aging effects in human brains and concluded that these resulted from an extended lifespan in humans (Sherwood et al. 2011). However, there is a substantial lack of data on the neurobiology of aging in great apes, which in part is probably due to the scarcity of available brain samples from those older individuals. Actually, the lifespan of chimpanzees can reach 60 years in captive individuals under continuous medical care (Erwin at al. 2002). Thus, this paucity in severe neurodegenerative processes in chimpanzees can hardly be considered a dogma. Our results indicate that not only main aggregated proteins in AD and PD (APP, tau and alpha-synuclein) are very conserved between humans and chimpanzees (Rosen et al. 2008) (Holzer et al. 2004) (Hamilton 2004) but that the whole neurological pathways explored have been maintained constrained in the chimpanzee lineage. Interestingly, in the human lineage both Parkinson's disease and Alzheimer's disease pathways also display a significant proportion of genes whose dN/dS ratio have decreased since the human-chimpanzee ancestor (i.e. 16% in the Alzheimer's pathway and 14% in the Parkinson's according to (Serra et al. 2011). Thus, overall all these findings will support the idea that the neurodegenerative effects specifically found in humans are probably the result of their extended lifespan rather than a trade-off with specific neurological adaptations.

Evolutionary patterns in coding vs. non-coding regions

After having discussed differences between pathways and established that, overall, the adaptive substitution rate (ω_α) observed in CDS is lower than that of genic non-coding regions (UTRS, introns, promoters and trailers). The questions still remains of whether evolutionary trends detected for a pathway's CDSs are also followed by the non-coding regions in these pathways; and, in correspondence, of whether the acceleration detected in the set of Acc. Introns is corresponded by the rest of elements in these genes.

In Table 3, we group elements by class and examine the differences among datasets for each kind of element. We observe that the differences between pathways detected for a certain class of element do not extend to the others. For instance, whereas the adaptive substitution rate (ω_α) in Complement CDS is significantly higher than that of the remaining CDS sets, the ω_α values from non-coding elements of the Complement pathway do not show any particular trend when comparing with similar categories from other pathways. Interestingly, the opposite is true for the Acc intron dataset: the adaptive substitution rate (ω_α) tends to be significantly higher than that of the rest of studied functional gene groups in other non-coding classes of elements, but not in coding regions (Table 3, Figure 2). These results suggest that natural selection is acting differently in coding and non-coding regions within the same pathways. That is: while a given class of genic elements within a pathway (for instance CDS) may be evolving under similar selective pressures; different classes of elements evolve differently from each other even within the same pathway.

When expanding the comparison of the adaptive substitution rate (ω_α) among CDS, UTRs, introns, promoters and trailers to all pathways analyzed, we do not find any clear pattern of correlation between the different genomic elements within a given pathway either (data not shown). Perhaps the only other striking observation would be that, CDS in the Acc Intron dataset do show the highest proportion of newly nearly neutral mutations and the second lower proportion of very deleterious mutations just after the Complement (Figure 1) as well as elevated pN/pS and dN/dS descriptives (Table 1). However, such patterns probably result from relaxed purifying selection rather than adaptive evolution acting on the CDS of genes with at least one positively selected intron.

It must be noted that we do not perform comparisons between adaptive substitution rates (ω_α) between elements within the same pathways. The reason to avoid doing so is that each class of element presents different functional properties that may introduce biases when estimating α and ω_α considering the element as a whole and using 4-fold degenerate SFS as neutral references. For instance, in contrast to CDS, some non-coding elements present very low levels of constraint, a fact reflected in estimates of DFE, which classify all mutations in introns, promoters and trailers as non deleterious, making estimates of α and ω_α for these elements hardly comparable from those from coding sequences.

Thus, we can conclude that we do observe a general decoupling trend of evolution between coding and non-coding regions of different pathways with differentiated divergence evolutionary patterns in the chimpanzee lineage. Moreover, this general decoupling is extreme in the Parkinson's case. Whereas CDS in Parkinson's present significantly higher ω_α values than those of the Actin pathway as well as a trend towards higher adaptive substitution rate in comparison with the remaining pathway categories except Complement, Parkinson's introns seem to present lower adaptive substitution rates than the remaining groups (with significant differences with the Amiloid, the Complement and the Presenilin pathways, see Table 3).

CONCLUSIONS

Pathway-based inferences of accelerated or conserved protein evolution based solely on divergence data are confirmed when considering polymorphism data. They present higher adaptive substitution rates than those from the remaining pathways. In fact, for the case of accelerated evolution in the Complement pathway, we observe that even the genes that did not present particularly accelerated rates of overall protein divergence, do show high rates of adaptive substitution in their coding regions. This is suggestive of proteins within the same pathway evolving under similar or even parallel adaptive pressures. Furthermore, we find evidence supporting the same pattern in some non-coding regions. The same trend is found for intron sequences that were shown to be accelerated based only on divergence data.. When analyzing genes that have one intron that can be detected as accelerated, the rest of introns in these genes also tend to present higher rates of adaptive substitution.

At least for the pathways under analysis, the adaptive history of coding and non-coding sequences seems to be decoupled. For instance, when evidence for positive selection is inferred from protein divergence, it does not imply that non-coding regions of the same genes will show the signature of the same selective pressures.

Finally, the coding regions of all three neurological pathways display consistent trends of evolutionary constraint in chimpanzees, which probably indicate functional conservation. Interestingly, Parkinson's disease and Alzheimer's disease pathways have a significant proportion of genes whose dN/dS ratio have decreased since the human-chimpanzee ancestor in both the human and the chimpanzee lineage (Serra et al. 2011). Thus, the neurodegenerative effects specifically found in humans are probably the result of their extended lifespan rather than a compromise with neurological adaptations.

MATERIALS AND METHODS

Samples

Blood-derived DNA samples for 20 wild-born non-related chimpanzees (*Pan troglodytes troglodytes*) from Gabon and Equatorial Guinea were obtained from the Research and Conservation Copenhagen Zoo. Detailed information on these individuals is provided in Table S1). DNA concentration was quantified with an Invitrogen QBit fluorometer to ensure a minimum of 6 ng at 50 ng/ul in each sample.

Pathways and regions analyzed

Two functional modules evolving at high and low evolutionary rates in the chimpanzee lineage (SerraF) and a set of genes with introns under accelerated evolution have been compared to three neurological modules related to common neurodegenerative diseases but with no clear trends of selection. A complete list of genes included in our study is given in Supplementary Table S1. In summary, we studied: 58 genes from the module “Complement and coagulation cascade” (KEGG: hsa 04810), 109 genes from the module “Regulation of Actine Cytoskeleton”, 66 genes from the “Alzheimer disease-amyloid secretase pathway (PANTHER P00003), 109 genes from the Alzheimer disease-presenilin pathway (PANTHER P00004) and 87 genes from the Parkinson disease” pathway (PANTHER P0049). We will refer to each of these sets of curated and non-overlapping genes as the “Complement”, “Actin”, “Amiloid”, Presenilin” or “ Parkinson” pathways in the remaining of the text. An additional set of 134 chimpanzee genes with evidence of accelerated in one of their introns have also been included. Signal of positive selection in these introns was obtained by a phylogenetic approach using the HYPHY software and the algorithm implemented in (Haygood et al. 2007). The hypothesis of positive selection in introns was tested using a neutral dataset of ancestral repeats non-overlapping coding sequences located in a window of 100 kb around each genomic intron (see further details in Supplementary Note 1). In total, we sequenced 536 genes and.

From the longest transcript of each particular gene included in these datasets we selected all their exons, a maximum of 1,000 bp flanking each intron boundary, both 5’ and 3’ UTRs, 5 kb of promoter region and 5 kb of trailers. Additionally, we also captured the complete length of those introns with accelerated substitution rates, and we will refer to this set as “Only Ac Introns”.

Capture Design and Sequencing

The complete list of genomic target intervals of interest in the PanTro2 (March 2006) chimpanzee genome (UCSC) was uploaded to the eArray XD software (Agilent Technologies) for custom design of two Agilent SureSelect Target Enrichment kits. Bait Tile parameters were set to use Illumina as a Sequencing Technology, Pair-End Long Read as Sequencing Protocol, and avoidance of masked repeat elements. We retiled those exonic regions not covered by baits, due to avoiding masked repeating regions, with masking OFF option. To ensure the specific retrieval of these genomic regions during enrichment, created baits created with the masking OFF option were then mapped with BLAT and only included in the design when their second match scored below 60 (half the baits length).

All capturing and sequencing procedures were performed at the Genomics Unit of the Center for Genomic Regulation (CRG) Core Facilities. Briefly, 6 μg of genomic DNA from each individual was fragmented to 150-200 bp with the use of the Adaptive Focused Acoustics (Covaris), end-repaired, adenylated, and ligated to specific PE tagged genomic adapters following the standard protocol of the Illumina Paired-End Sample Preparation kit. Different pools of 2-3 indexed libraries were then hybridised in multiplex with the 120 bp biotinylated RNA baits of each custom Agilent SureSelect kits. After enrichment with each individual kit, captured fragments were purified, pooled in two groups (each containing the two sets of captured regions from each of 10 different samples), and sequenced on two lanes of an Illumina HiSeq 2000 System with the use of 75 bp paired-end reads. See detailed information on enrichment and sequencing scheme in Supplementary Table S3.

Mapping, SNP calling and callable genome construction

Burrows-Wheeler Aligner (BWA) was used to align reads to the chimpanzee reference genome (PanTro3), using the default parameters. We used the Genome Analysis Toolkit (GATK) to call indels and to perform re-alignments of reads falling in their surrounding regions. We then carried out recalibration of reads' bases quality scores (BQSR), considering cycle-effect (position in the read) and sequencing chemistry effect (the preceding and current nucleotide). UnifiedGenotype from GATK was used to call genotypes. Afterwards, we filtered SNPs by the variant quality score recalibration (VQSR) method (GATK). We provided a training set consisting of a high-quality SNP subset, i.e. SNPs called with at least 10,000 of phred quality score and not falling in SNP clusters (3 or more SNPs in windows of 10 bp) or in a region at 5 bp or less from an indel. After filtering the whole set of SNPs using VQSR we also remove SNPs falling in clusters and close to indels the same way than in the high-quality subset.

The proportion of genome that is callable and that we finally used in all analyses was calculated by applying the UnifiedGenotype from GATK emitting genotypes for all confident reference and variant sites. We kept only those sites with a confidence genotype call of at least 30 of phred score (the same minimum threshold applied to variable sites). We masked all indel with a padding of 5 bp. All sites with SNPs filtered by the VQSR method and SNP cluster criteria were also removed from the callable fraction of the genome. Finally, from all we finally kept only those positions in which all-20 chimpanzee individuals possess coverage of at least 5X. That left 8,599,335 bp of callable regions including 79,650 SNPs with a Ti/Tv ratio of 2.29. The fraction of SNPs found in coding regions possesses a Ti/Tv ratio of 3.89, consistent with the elevated GC content of exons.

Non-synonymous and synonymous sites

Zero-fold degenerated sites were treated as non-synonymous sites and the four-fold degenerated sites as synonymous. Exonic sequences and genomic annotation from all genes were downloaded from the *Pan troglodytes* genes (CHIMP2.1.4) dataset, at the Ensembl genes 66 database in ENSEMBL (http://www.ensembl.org/biomart/) and lifted over to PanTro3 assembly.

Exclusion of conserved noncoding elements

Conserved noncoding elements (CNEs) within promoters, introns and trailers of all genes studied were identified in the Human genome (hg19) using PhastCons on a 12-way Vertebrate Multiz Alignment. PhastCons parameters were tuned to produce 5% conserved elements in the genome for the vertebrate conservation measurement (expected-length=45, target-coverage=.3, rho=.31). hg19 CNEs coordinates were then mapped to the chimpanzee genome (PanTro3) using the liftOver tool at the Galaxy website (http://main.g2.bx.psu.edu/). CNEs in non-coding sequences can be used as a proxy for functional regulatory motifs in flanking regions surrounding CDS. Unfortunately, their overall length (as well as the total amount of variation found in them) from each specific non-coding region in a given pathway (i.e. promoters, introns and trailers) was not sufficient to evaluate and compare their corresponding evolutionary patterns separately (data not shown), so we simply excluded them from analysis.

Inference of unfolded site frequency spectrum (SFS)

SNPs were oriented with the information of ancestralities in the node separating chimpanzee, human and orangutan species. Ancestral state of all sequenced nucleotides was inferred by parsimony according to the following strategy in the Galaxy website (http://main.g2.bx.psu.edu/): data of divergent sites between panTro3 and hg19 were obtained using the regional variation/fetch substitutions from pairwise alignments tool. For each nucleotide substitution identified, pairwise alignments of panTro3/hg19 and panTro3/ponAbe2 were downloaded using the fetch alignments/fetch pairwise maf blocks tool given our set of genomics intervals. Derived substitutions in panTro3 were considered those where the nucleotide in panTro3 is different to the nucleotide of hg19 and ponAbe2. Identified derived substitutions in panTro3 were compared with data from ancestralities obtained for *Pan troglodytes troglodytes* in (Prado-Martinez et al. 2013), obtaining a 99,1% of coincidences for monomorphic positions and

97% of coincidences with the derived allele when including polymorphic positions in the node separating the three species. To obtain the site frequency spectrum of polymorphism the derived alleles in each category from 1 to n-1 chromosomes were counted. We also counted the number of invariant sites and the derived substitutions. Gaps positions were not taken into account for the SFS.

Descriptives for diversity and divergence

We calculated descriptive measures of diversity and divergence for all pathways analysed by comparing SNPs and substitutions in putatively selected and non-selected sites for all genes and genomic elements in a given pathway. In particular, we counted the proportion of non-synonymous SNPs (pN), all SNPs in non-coding regions as putatively selected SNPs (pPUS), the proportion of synonymous SNPs (pS), the proportion of non-synonymous substitutions (dN), all substitutions in non-coding regions as putatively selected substitutions in non-coding regions (dPUS) as well as the proportion of non-synonymous substitutions (dS). We then obtained the corresponding rates of putatively selected versus neutral SNPs and substitutions as pN/pS and dN/dS in coding regions and as pPUS/pS and dPUS/dS in non-coding regions, respectively.
Putatively selected sites at coding regions were considered to be only 0-fold sites. As non-selected (i.e. neutral) sites for each pathway we considered only the 4-fold sites in each pathway's coding region. Thus, for all non-coding elements we used the same neutral reference, i.e. the one from their respective coding regions.

To evaluate whether pN, pS, dN, dS, pN/pS and dN/dS values were significantly different between pathways we obtained a distribution of all these descriptive values by taking genes randomly 1,000 times. At each round of random selection of genes we required a minimum number of sites analysed equal to the number of sites analysed in the real list of genes for each pathway. Then, for each observed value we calculated the percentile it represents in the distribution of 1,000 randomizations of the compared pathway. We considered that pathway values are significantly different when their observed values fall reciprocally in the tails of the distribution, considering percentiles 2.5% and 97.5% as thresholds.

Distribution of fitness effects of new mutations

Distribution of fitness effects of new mutations (DFE) together with demographic parameters were obtained using a ML-based approach based on polymorphism data as described in (Keightley and Eyre-Walker 2007). Briefly, the method assumes 2 classes of sites (one evolving neutrally and another under mutation-selection-drift balance) and contrasts the folded SFS at these two classes. Under this model sites can be either neutral or damaging. Deleterious effects are sampled from a gamma distribution with *a* (scale) and *b* (shape) parameters. From the gamma distribution 4 different categories of $N_e.s$ ranges are estimated: 0-1, 1-10, 10-100, >100, considered as nearly neutral, mildly deleterious, deleterious and very deleterious, respectively. Demography is modelled in the method considering a 2-ephoc model with one-step change from N_1 initial size equal to 100 to N_2 final size in *t* generation in the past, the ratio of change N_1/N_2 is estimated together with the proportion of unmutated sites *(*f_o*)*. Since neutral evolutionary rates at 4-fold sites are similar among the different datasets, we considered the concatenation of all 4-fold sites (synonymous sites) as the neutral reference for all the putative selected classes analysed.

A simple extension of the method described above allows the calculation of the proportion of fixed adaptive substitutions (α) and the relative rate of adaptive substitution (ω_a) using the parameters calculated in the DFE together with divergence data (Keightley and Eyre-Walker 2007). This is an extension of the classic McDonald-Kreitman-based tests that benefits of accounting for the possible effect of slightly deleterious mutations contributing to polymorphisms and bias due to demographic changes.

Statistical comparisons between pathways and elements

We calculated α and ω_a for our putative selected classes within each class of putative selected elements in all pathways analysed and calculate their confidence intervals by bootstrapping genes using two different strategies. First, we performed random selection of genes up to the total number of genes for each pathway, with replacement. We added all site frequency spectra for that list of genes at all genomic elements, so as the comparison among genomic elements was always done upon the same list of genes. We repeated that 1,000 times and obtained the same number of alpha and omega values. We argued that for some genomic elements with little information, most of it might come from a few genes while others possess little or no-information. We then repeated our bootstrap analysis by bootstrapping independently each genomic element, and requiring that the number of base pairs analysed at each randomly produced gene list should be at least equal than the number of base pairs analysed in the original gene list for each pathway. Both methods gave equivalent results.

To evaluate whether proportions and rates of adaptive changes based on α and ω_a values were significantly different between pathways and elements we applied the following strategy: when comparing different pathways for a particular element studied, for each observed result, we calculated the percentile where it falls in the distribution of bootstrapped values of the compared pathway. We considered the pathway's observed value significantly different from the one it is compared to if it falls in the tails of the distribution considering percentile 2.5% and 97.5% as thresholds.

ACKNOWLEDGEMENTS

The research presented was supported by Ministerio de Ciencia e Innovación, Spain (BFU2008-01046 and BFU2011-29239 to EB) and by Direcció General de Recerca, Generalitat de Catalunya (2009SGR 1101). We also thank the National Institute of Bioinformatics (http://www.inab.org) for the computational support and Marc Torres-Ciuró for graphical design.

REFERENCES

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., et al. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS genetics*, *4*(5), e1000083.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., et al. (2005). Natural selection on protein-coding genes in the human genome. *Nature*, *437*(7062), 1153–7.

Charlesworth, J., & Eyre-Walker, A. (2006). The rate of adaptive evolution in enteric bacteria. *Molecular biology and evolution*, *23*(7), 1348–56.

Elyashiv, E., Bullaughey, K., Sattath, S., Rinott, Y., Przeworski, M., & Sella, G. (2010). Shifts in the intensity of purifying selection: an analysis of genome-wide polymorphism data from two closely related yeast species. *Genome research*, *20*(11), 1558–73.

Eyre-walker, A. (2006). The genomic rate of adaptive evolution. *Evolution*, *21*(10).

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature reviews. Genetics*, *8*(8),

Fay, J. C. (2011). Weighing the evidence for adaptation at the molecular level. *Trends in genetics : TIG*, *27*(9), 343–9.

Gearing, M., Rebeck, G. W., Hyman, B. T., Tigges, J., & Mirra, S. S. (1994). Neuropathology and apolipoprotein E profile of aged chimpanzees: implications for Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(20), 9382–6.

Hamilton, B. A. (2004). alpha-Synuclein A53T substitution associated with Parkinson disease also marks the divergence of Old World and New World primates. *Genomics*, *83*(4), 739–42.

Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.-D., & Wray, G. a. (2007). Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nature genetics*, *39*(9), 1140–4.

Holzer, M., Craxton, M., Jakes, R., Arendt, T., & Goedert, M. (2004). Tau gene (MAPT) sequence variation among primates. *Gene*, *341*, 313–22.

Hvilsom, C., Qian, Y., Bataillon, T., Li, Y., Mailund, T., Sallé, B., & Carlsen, F. (2011). Extensive X-linked adaptive evolution in central chimpanzees, 1–6. doi:10.1073/pnas.1106877109

Keightley, P. D., & Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, *177*(4), 2251–61.

Kimura, N., Nakamura, S., Goto, N., Narushima, E., Hara, I., Shichiri, S., Saitou, K., et al. (2001). Senile plaques in an aged western lowland gorilla. *Experimental animals / Japanese Association for Laboratory Animal Science*, *50*(1), 77–81.

Kosiol, C., Vinar, T., Da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., & Siepel, A. (2008). Patterns of positive selection in six Mammalian genomes. *PLoS genetics*, *4*(8), e1000144.

Kousathanas, A., Oliver, F., Halligan, D. L., & Keightley, P. D. (2011). Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Molecular biology and evolution*, *28*(3), 1183–91.

Lawrie, D. S., Messer, P. W., Hershberg, R., & Petrov, D. A. (2013). Strong Purifying Selection at Synonymous Sites in D. melanogaster. (J. B. Plotkin, Ed.)*PLoS Genetics*, *9*(5), e1003527.

Mackay, T. F. C., Richards, S., Stone, E. a, Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature*, *482*(7384), 173–8.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652–4.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., et al. (2013). Great ape genetic diversity and population history. *Nature*, *499*(7459), 471–5.

Rosen, R. F., Farberg, A. S., Gearing, M., Dooyema, J., Long, P. M., Anderson, D. C., Davis-Turak, J., et al. (2008). Tauopathy with paired helical filaments in an aged chimpanzee. *The Journal of comparative neurology*, *509*(3), 259–70.

Sella, G., Petrov, D. A., Przeworski, M., & Andolfatto, P. (2009). Pervasive natural selection in the Drosophila genome? (M. W. Nachman, Ed.)*PLoS genetics*, *5*(6), e1000495. doi:10.1371/journal.pgen.1000495

Serra, F., Arbiza, L., Dopazo, J., & Dopazo, H. (2011). Natural selection on functional modules, a genome-wide analysis. *PLoS computational biology*, *7*(3), e1001093.

Sherwood, C. C., Gordon, A. D., Allen, J. S., Phillips, K. A., Erwin, J. M., Hof, P. R., & Hopkins, W. D. (2011). Aging of the cerebral cortex differs between humans and chimpanzees. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(32), 13029–34.

Welch, J. J. (2006). Estimating the genomewide rate of adaptive protein evolution in Drosophila. *Genetics*, *173*(2), 821–37.

Figure Legends

Figure 1. **Distribution of fitness effects of new mutations in coding sequences from different pathways.** Bars indicate the proportion of mutations that are nearly neutral (N_es<1), mildly deleterious (1< N_es <10), deleterious (10< N_es <100) and very deleterious (N_es >100), respectively.

Figure 2. **Ratio of adaptive to neutral divergence. A.** Omega (ω_α) values per genomic element and pathway. Significance values for the 95% confidence interval have been obtained by bootstrapping requiring a minimum threshold of size (bp). Values for the 2.5% and 97.5% threshold are indicated. **B**. CDS Omega (ω_α) values comparison between the actin and complement pathways. The comparison is shown overall as well as between the actin and the percentiles 25, 25-75 and 75 of the complement d_N/d_S gene distribution values as calculated in (Serra et al. 2011).

Table 1. CDS polymorphisms and divergence statistical descriptives in 20 chimpanzees.

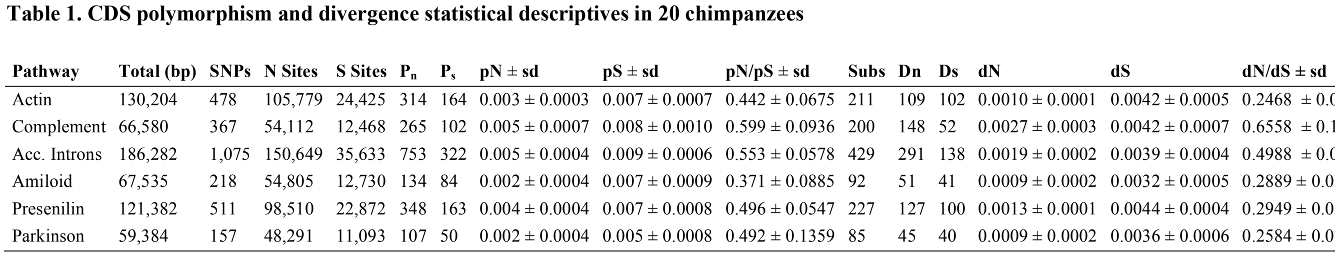


Table 1. CDS polymorphism and divergence statistical descriptives in 20 chimpanzees

Pathway	Total (bp)	SNPs	N Sites	S Sites	P_n	P_s	pN ± sd	pS ± sd	pN/pS ± sd	Subs	Dn	Ds	dN	dS	dN/dS ± sd
Actin	130,204	478	105,779	24,425	314	164	0.003 ± 0.0003	0.007 ± 0.0007	0.442 ± 0.0675	211	109	102	0.0010 ± 0.0001	0.0042 ± 0.0005	0.2468 ± 0.0
Complement	66,580	367	54,112	12,468	265	102	0.005 ± 0.0007	0.008 ± 0.0010	0.599 ± 0.0936	200	148	52	0.0027 ± 0.0003	0.0042 ± 0.0007	0.6558 ± 0.1
Acc. Introns	186,282	1,075	150,649	35,633	753	322	0.005 ± 0.0004	0.009 ± 0.0006	0.553 ± 0.0578	429	291	138	0.0019 ± 0.0002	0.0039 ± 0.0004	0.4988 ± 0.0
Amiloid	67,535	218	54,805	12,730	134	84	0.002 ± 0.0004	0.007 ± 0.0009	0.371 ± 0.0885	92	51	41	0.0009 ± 0.0002	0.0032 ± 0.0005	0.2889 ± 0.0
Presenilin	121,382	511	98,510	22,872	348	163	0.004 ± 0.0004	0.007 ± 0.0008	0.496 ± 0.0547	227	127	100	0.0013 ± 0.0001	0.0044 ± 0.0004	0.2949 ± 0.0
Parkinson	59,384	157	48,291	11,093	107	50	0.002 ± 0.0004	0.005 ± 0.0008	0.492 ± 0.1359	85	45	40	0.0009 ± 0.0002	0.0036 ± 0.0006	0.2584 ± 0.0

Table 2. Non-coding polymorphism and divergence statistical descriptives in 20 chimpanzees

Pathway	Total (bp)	SNPs	pPUS ± sd	pPUS/pS ± sd	Subs	dPUS	dPUS/dS ± sd
Intron							
Actin	1,139,546	10,152	0.0089 ± 0.0004	1.327 ± 0.1310	4,953	0.0043 ± 0.0001	1.041 ± 0.1319
Complement	469,503	4,317	0.0092 ± 0.0005	1.124 ± 0.1139	2,164	0.0046 ± 0.0002	1.105 ± 0.1922
Acc. Introns	1,892,153	20,518	0.0108 ± 0.0002	1.200 ± 0.0815	10,88	0.0058 ± 0.0002	1.485 ± 0.1492
Only Acc Introns	418,919	4,376	0.0104 ± 0.0003	1.156 ± 0.0852	3,053	0.0073 ± 0.0003	1.882 ± 0.1881
Amiloid	678,528	5,825	0.0086 ± 0.0003	1.301± 0.1754	3,037	0.0045 ± 0.0002	1.390 ± 0.2376
Presenilin	1,013,409	9,218	0.0091 ± 0.0005	1.276 ± 0.1125	4,545	0.0045 ± 0.0002	1.026 ± 0.1084
Parkinson	560,49	4,371	0.0078 ± 0.0003	1.730 ± 0.3488	2,298	0.0041 ± 0.0001	1.137± 0.1934
Promoter							
Actin	183,452	1,431	0.0078 ± 0.0004	1.162 ± 0.1235	721	0.0039 ± 0.0002	0.941 ± 0.1347
Complement	101,302	931	0.0092 ± 0.0005	1.123 ± 0.1256	439	0.0043 ± 0.0002	1.039 ± 0.1911
Acc. Introns	229,95	2,411	0.0106 ± 0.0003	1.175 ± 0.0895	1,078	0.0047 ± 0.0002	1.210 ± 0.1342
Amiloid	109,487	943	0.0086 ± 0.0004	1.305 ± 0.1907	459	0.0042 ± 0.0003	1.302 ± 0.2618
Presenilin	181,194	1,459	0.0081 ± 0.0004	1.130± 0.1383	808	0.0045 ± 0.0003	1.020 ± 0.1187
Parkinson	120,311	948	0.0079 ± 0.0003	1.748 ± 0.3794	507	0.0042 ± 0.0002	1.169 ± 0.2301
Trailer							
Actin	206,747	1,729	0.0084 ± 0.0004	1.246 ± 0.1397	845	0.0041 ± 0.0002	0.979 ± 0.1339
Complement	111,924	987	0.0088 ± 0.0005	1.078 ± 0.1225	500	0.0045 ± 0.0003	1.071 ± 0.1963
Acc. Introns	290,773	2,945	0.0101 ± 0.0003	1.121 ± 0.0810	1,467	0.0050 ± 0.0002	1.303 ± 0.1520
Amiloid	124,015	1,052	0.0085 ± 0.0004	1.286 ± 0.1820	504	0.0041 ± 0.0003	1.262 ± 0.2330
Presenilin	251,146	2,157	0.0086 ± 0.0003	1.205± 0.1506	1,051	0.0042 ± 0.0002	0.957 ± 0.1059
Parkinson	165,192	1,283	0.0078 ± 0.0004	1.723 ± 0.3415	696	0.0042 ± 0.0002	1.168 ± 0.2209
UTR							
Actin	105,081	705	0.0067 ± 0.0005	0.999 ± 0.1231	345	0.0033 ± 0.0002	0.786 ± 0.1149
Complement	28,099	220	0.0078 ± 0.0006	0.957 ± 0.1344	114	0.0041 ± 0.0004	0.973 ± 0.1849
Acc. Introns	92,742	934	0.0101 ± 0.0009	1.114 ± 0.1225	382	0.0041 ± 0.0003	1.064 ± 0.1520
Amiloid	71,781	503	0.0070 ± 0.0004	1.062 ± 0.1718	256	0.0036 ± 0.0005	1.107 ± 0.2467
Presenilin	84,996	596	0.0070 ± 0.0004	0.984± 0.11249	291	0.0034 ± 0.0003	0.783 ± 0.1033
Parkinson	60,283	323	0.0054 ± 0.0004	1.189 ± 0.2743	179	0.0030 ± 0.0004	0.823 ± 0.1982

Table 3. Comparison of omega (ω_α) values between pathways for each genomic element analyzed.

	Actin	Complement	Acc. Int	Amiloid	Presenilin	Parkinson
CDS						
ω_α	***-0.06***	***0.89***	***0.07***	***0.12***	***0.06***	***0.29***
Actin	0.482	0.001	0.193	0.098	0.204	0.014
Complement	1.000	0.451	1.000	1.000	1.000	1.000
Acc Introns	0.787	0.004	0.444	0.366	0.538	0.083
Amiloid	0.872	0.006	0.565	0.501	0.658	0.142
Presenilin	0.770	0.004	0.420	0.335	0.508	0.080
Parkinson	0.996	0.015	0.831	0.891	0.951	0.527
Intron						
ω_α	***0.75***	***0.95***	***1.83***	***0.85***	***0.86***	***0.57***
Actin	0.500	0.100	0.000	0.209	0.202	0.776
Complement	0.975	0.506	0.000	0.762	0.762	0.952
Acc Introns	1.000	1.000	0.528	1.000	1.000	1.000
Amiloid	0.824	0.255	0.000	0.492	0.484	0.903
Presenilin	0.860	0.277	0.000	0.524	0.506	0.909
Parkinson	0.033	0.002	0.000	0.021	0.010	0.301
Promoter						
ω_α	***0.44***	***0.74***	***1.01***	***0.64***	***0.84***	***0.65***
Actin	0.394	0.047	0.000	0.161	0.015	0.087
Complement	0.850	0.476	0.046	0.660	0.313	0.600
Acc Introns	0.969	0.940	0.480	0.945	0.809	0.909
Amiloid	0.758	0.270	0.012	0.479	0.172	0.396
Presenilin	0.912	0.705	0.127	0.824	0.497	0.754
Parkinson	0.768	0.282	0.012	0.503	0.184	0.417
UTR						
ω_α	***0.60***	***0.53***	***0.58***	***1.00***	***0.51***	***0.67***
Actin	0.511	0.514	0.526	0.264	0.710	0.345
Complement	0.446	0.431	0.387	0.218	0.636	0.266
Acc Introns	0.487	0.494	0.478	0.254	0.684	0.322
Amiloid	0.745	0.875	0.979	0.565	0.926	0.849
Presenilin	0.420	0.400	0.345	0.203	0.615	0.248
Parkinson	0.556	0.593	0.659	0.310	0.766	0.428
Trailer						
ω_α	***0.56***	***0.85***	***1.29***	***0.54***	***0.63***	***0.80***
Actin	0.520	0.086	0.000	0.553	0.344	0.081
Complement	0.965	0.516	0.003	0.907	0.917	0.483
Acc Introns	1.000	0.976	0.488	0.999	1.000	0.907
Amiloid	0.461	0.064	0.000	0.528	0.287	0.071
Presenilin	0.678	0.158	0.000	0.683	0.494	0.160
Parkinson	0.931	0.422	0.000	0.874	0.854	0.395

Under the diagonal, the percentile in which estimated ω_α values of the pathways (in rows) fall in the bootstrapped distribution of ω_α values of the corresponding compared pathway (in columns). The reciprocal comparison is shown above the diagonal. Upper (dark grey) and lower (light grey) significance thresholds are set to the 0.975 and 0.025 percentiles of the bootstrapped distribution. Black cells contain the percentile of the estimated ω_α value of a given pathway within its own bootstrapped distribution of ω_α. Cells in cursive bold contain the observed ω_α values.

Figure 1.

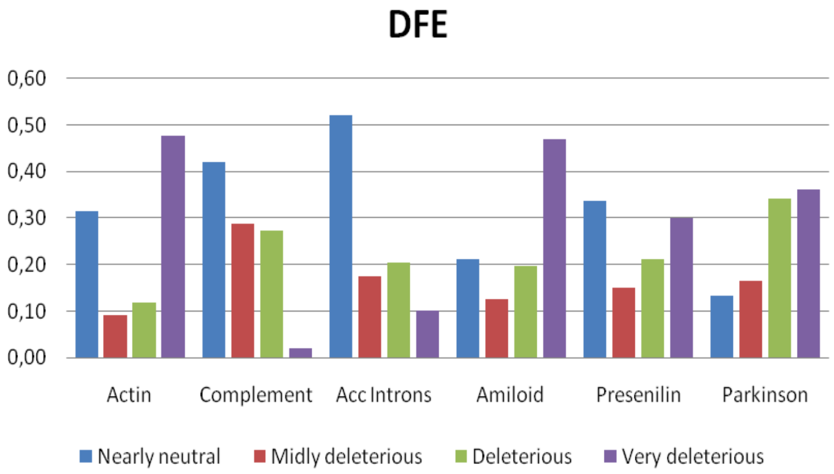


Figure 2.
A.

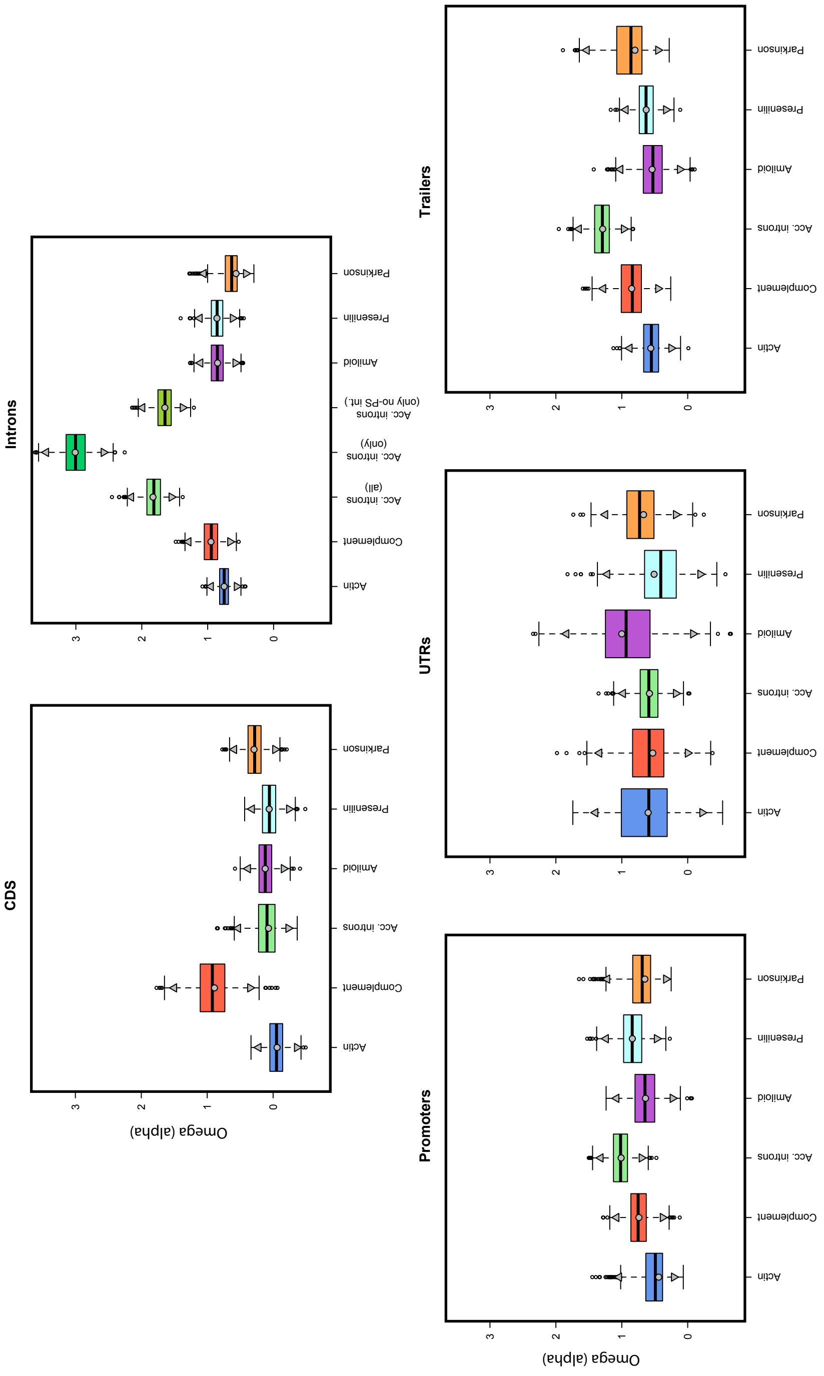
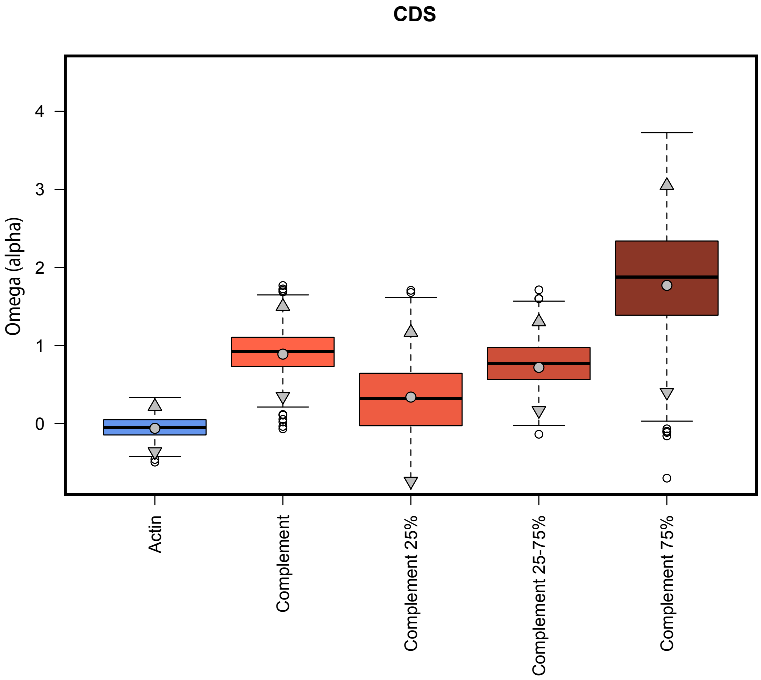

CDS
Introns
Promoters
UTRs
Trailers
Omega (alpha)
0
1
2
3
Actin
Complement
Acc. introns
Amiloid
Presenilin
Parkinson
Acc. introns (all)
Acc. introns (only)
Acc. introns (only no-PS int.)

Figure 2

B.


CDS
Omega (alpha)
4
3
2
1
0
Actin
Complement
Complement 25%
Complement 25-75%
Complement 75%

Discussion

1. The adaptive role of the *CD5* gene in East Asian populations

In the study presented in chapter 1 of the results section of this thesis we describe a new example of human recent adaptation. We provide convincing evolutionary and functional evidences for a process of positive selection acting on the immune-receptor *CD5* gene in East Asian populations.

Previously, different genome-wide scans of positive selection had analyzed genotype data and found a number of different molecular signatures for positive selection in the region where *CD5* is located (Moreno-Estrada et al. 2009)(Williamson et al. 2007). The region is located in chromosome 11 and spans more than 400 kb where many other functionally relevant genes are located (figure 1). Such high gene density makes it difficult to recognize which variant could have risen up in frequency and driven the selective signature. Moreno-Estrada et al. (2009) made a step forward to elucidate which gene could have been the actual target of selection. By analyzing all the potential functional variation present in the whole 400 kb region they predicted that an Ala471Val substitution (rs2222971) in the CD5 receptor was the most plausible target for such selective sweep. Their suggestion was made based on the high population differentiation observed on such non-synonymous SNP, the presence of the derived allele on the selected haplotype and, furthermore, on the *in silico* prediction of the substitution as probably damaging, which suggests that the change has likely an impact in the protein function. However, the relevance of the Moreno-Estrada et al. (2009) findings regarding rs2222971 as putative target of selection were limited because they based their adaptive inferences on genotype data and on *in silico* predictions.

The aforementioned study together with the parallel recognition of *CD5* as a pathogen recognition receptor (PRR) of specific fungal particles (Vera et al. 2009) were the initial points that motivated us to further investigate *CD5* as a candidate gene for recent human adaptation.

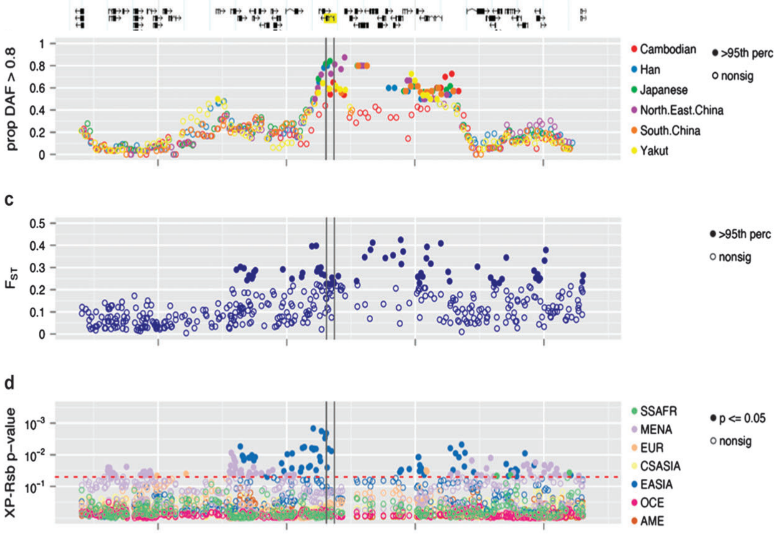


Figure 1. Significant signatures of selection in the *CD5* region on East Asians. Statistical tests, based on allele frequency spectrum (DAF), population differentiation (F_{ST}), and LD-patterns (XP-Rsb) were performed on genotype data (Modified from Moreno-Estrada et al, 2009)

As it has been extensively described in section 6 of the introduction, the never-ending challenge for survival imposed by pathogens has been one of the major selective pressures in our evolutionary history.

One of the most important elements in host-pathogen interactions are the PRRs (Pattern Recognition Receptors) that are in charge of initially recognizing pathogen molecules, and consequently, triggering the subsequent immune responses. Fascinatingly, recent population genetics studies on such elements have revealed several cases of positive selection and have given important insights into the process of adaptation of both interacting species (see Introduction, section 6). Nonetheless, the adaptive role of the scavenger-receptor cysteine-rich superfamily (SRCR-SF) members was not recognized until this work.

The CD5 receptor belongs to the SRCR-SF of soluble-secreted and membrane-bound receptors. Members of this family present a broad range of functions, as for instance pathogen recognition, immune-modulation, and epithelial homeostasis. In spite of their high functional heterogeneity, they all share one or several very ancient and highly conserved SRCR domains that are around 100 amino acids long and are characterized by their high and well-defined cysteine content. SRCR key residues show a very high degree of conservation, as the cysteine positions, while other residues have evolved more freely allowing the development of versatile functions. Although most SCRC-SF members are found in mammals, they have also been found in other vertebrates, invertebrates, algae, and even in sponges. Such wide distribution suggests an early appearance of the SCRC-SF members in the animal phyla, and indicates that they likely belonged to the first elements of the ancient innate immune system. Several membrane-bound receptors showing SRCRs are indeed PRRs that recognize PAMPs and thus, are directly involved in pathogen recognition. This has been shown to be the case, for example, of SR-AI-II-II, SCARA5, S5D-SRCRB, Spa-α , CD136, CD6, and recently CD5 (Vera et al. 2009)(Martínez et al., 2011). Those elements constitute *per se* excellent candidates to interrogate their possible role in pathogen-driven adaptation.

CD5 is expressed in thymocytes, T cells and in the B1a subset cells. It possesses an extracellular region containing 3 SRCR domains, a transmembrane domain plus a cytoplasmatic region. The large cytoplasmatic region of CD5 devoid intrinsic enzymatic activity but is involved in intracellular signaling by undergoing structural modification and allowing the binding of accessory molecules. It contains several extremely highly conserved docking sites, as for example the ITIM-like domain, that are essential for its function. In brief, the cytoplasmatic domain is the scaffold of the protein that allows the signal received by *CD5* to be further translated into different responses.

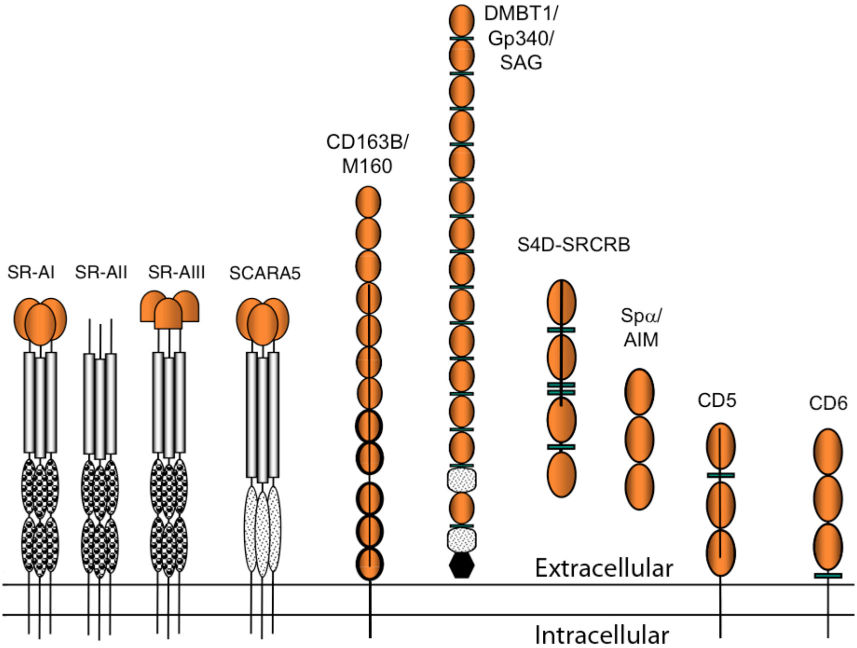


Figure 2. Members from the scavenger receptor cysteine-rich (SRCR) involved in pathogen recognition. Domains in orange are SRCR domains shared among all of them

The putative adaptive variant that we identified as the target of positive selection is located in the cytoplasmatic part of CD5 receptor, very close to the ITIM-like domain. Its position, in such a conserved and functional important region, points out the likely functional relevance of the change. Indeed, it is predicted to be probably damaging by *in silico* analysis. However, and very importantly, experimental *in vitro* functional differences between the two alleles of this non-synonymous polymorphism, the specific cellular phenotypical adaptative response of the Val471 allele and the putative selective force that selected such allele, remained unknown.

In our study, we were able to confirm the previous signatures of selection on *CD5* gene in East Asians, by resequencing its complete sequence in 60 individual from 3 HapMap populations including the African, European and Asian ancestries Resequencing allowed us to obtain not only the complete spectrum of variation but to perform more robust neutrality tests based on sequence data, and not limited to genotype data. Importantly, we gave statistical significance to our results based on the distribution of expected values produced by extensive coalescent simulations. Simulations allow overcoming one of the toughest caveats of genome-wide studies based on outlier approaches (see Introduction, section 3). We simulated the expected values under an *ad-hoc* coalescent model as described in (Schaffner et al. 2005) which takes into account the specific demography of the studied populations, together with the particular genomic characteristics of the region analyzed, two factors that can profoundly impact the neutral expectation of diversity . Furthermore, we were able to confirm that the Alanine to Valine change at the position 471 of the CD5 receptor is, in fact, the most plausible target of positive selection, as that no other putative functional variant on CD5 showed such high allele population differentiation while being linked to the selected haplotype.

At this point, an important episode in our research was the establishment of a close collaboration with the Immunology Service from the Hospital Clínic in Barcelona. Such collaboration allowed the integration of functional experimental analysis in our population genetics framework to overcome the lack of previous biological knowledge on the impact of this change in the immunological responses mediated by CD5 receptor. The performance of functional analysis on transfected cells and peripheral mononuclear blood cells (PMBC) helped us to understand how the derived Val471 variant could have facilitated adaptation to specific environmental pressures in the East Asian population.

Based on our results in transfected cells, an heterologous system, and in peripheral blood mononuclear cells (PBMC) from different individuals, we concluded that the hyper-signaling immune responsiveness of East Asians as carriers of the selected Val471 could have conferred them resistance to past fungal pathogen exposure. The higher production of IL-8 supports this hypothesis. IL-8 release is known to be a key player in host defense against some invasive fungal infection such as *Candida albicans* (Mostefaoui et al. 2004). This observation, together with the recent finding that the extra-membrane domain of CD5 is implicated in specifically sensing fungal particles, add further evidence to support the hypothesis that the derived Val variant could have been favored in East Asians as it confers a more efficient immune response against such pathogens.

To unravel which specific selective force triggered a particular adaptive event is often very helpful to identify when the selection occurred. The worldwide frequency distribution of the Val471 variant suggest that the selected derived variant was already segregating before it increased frequency in East Asians. Since Native American populations also seem to present extremely high frequencies of the derived allele we inferred that the selection event occurred recently sometime between the colonization of East Asia and that of the Americas (see figure 3).

Unfortunately, to the best of our knowledge, no clear record of fungal epidemics is available in such timeframe for East Asian populations that could be the selective pressure triggering this adaptation, so other possible phenotypic features conferred by this variant cannot be discarded.

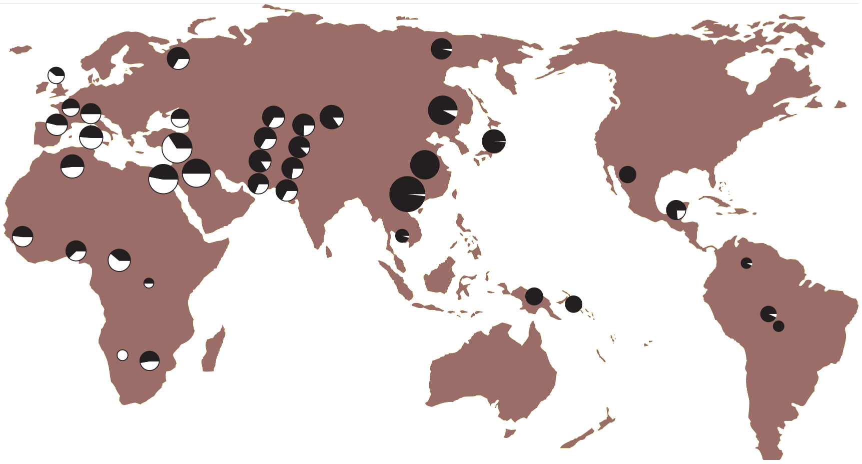


Figure 3. Pie chart of worldwide allele frequency distribution at the Ala471Val polymorphism. White represents the ancestral state while black represents the derived one. The adaptive derived Val is fixed in most East Asian and in Native Americans.

An important feature of CD5 to take into account is that it is not only involved in the innate adaptive immunity, but has also an important role in regulating the adaptive immune responses. In summary, CD5 negative regulates the antigen-mediated lymphocyte receptor activation through physical association to both T and B cell receptors (TCR and BCR, respectively) as it has been shown in mouse deficient models (reviewed in Lozano et al., 2000). Intriguingly, broader implications of CD5 in lymphocyte survival and tolerance have been found out very recently (reviewed in Soldevila et al 2011). Up-regulation of *CD5* in T cells has been associated to inhibition of apoptosis, and to IL-10 production in B cells. Both processes enhance lymphocyte survival chances. On the other hand, *CD5* up-regulation has also a role in auto-reactive T and B cells avoidance by increasing activation antigen thresholds of lymphocytes and by regulating the IL-10 production of B regulatory cells (see figure 4).

The pleiotropic nature of the CD5 receptor makes it more difficult to understand its implication in adaptation as the same change can impact multiple immunological functions and, likely, influence pathogenesis of immune-related diseases related to survival and tolerance, as for example cancer and autoimmunity. To date, and to the best of our knowledge, two association studies have revealed an association between the A471V alleles and immune-related diseases. In particular, the Ala471 variant was associated with an increased risk for B-cell chronic Lymphocytic Leukemia(B-CLL) prognosis (Sellick et al., 2008). B-CLL is a lymphoproliferative neoplastic disorder that currently represents a major threat for Western countries. Interestingly, very recently an extensive GWA analysis of rheumatoid arthritis has also identified the Ala471 variant to be associated with this autoimmune disease (Eyre et al. 2012).

Regulation of CD5 expression is known to be a key process to both prevent cancer and autoimmune disease, but in completely different directions. Down-regulation of CD5 expression has shown to be linked to anti-tumoral responses, as it promotes cellular apoptosis and leads to a more reactive lymphocyte response that can participate more efficiently in killing malignant cells. On the contrary, CD5 expression has been shown to protect from some autoimmune diseases. CD5 expression enhances IL-10 release, which suppresses immune responses and assures the maintenance of proper self-tolerance levels (Dalloul 2009). Although the importance of CD5 in fine-tuning the immune response is clear, the direction in which the A471V can influence the pathogenesis of such responses is far from being completely understood.

In this aspect, we have further collaborated with our immunologist colleagues to assess the implication of the CD5 A471V polymorphism in the lymphocyte proliferative responses of both T and B cells as well as in the pathogenesis of autoimmune disease systemic lupus erythematosus (SLE) (data not published).

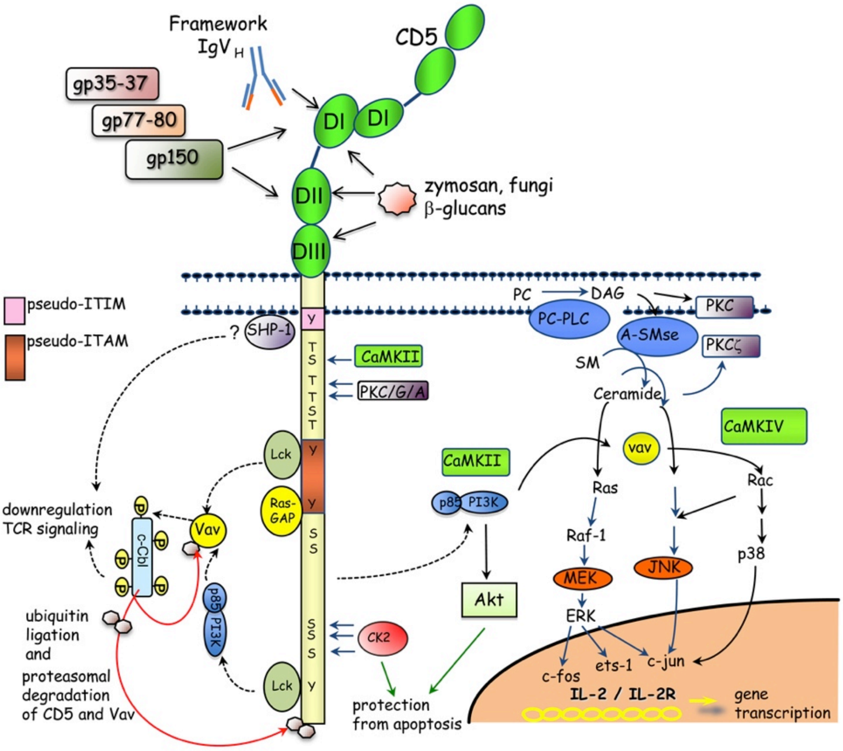


Figure 4. Lymphocyte responses triggered by CD5 activation. Multiple signaling pathways are associated to the cytoplasmatic motifs of CD5 correceptor. Stimulation of CD5 extramembrane domain by natural or exogenous ligands provoke the signaling activation of certain pathways and ultimate lead to process as TCR dowregulation, cytokine release and/or inhibition of apoptosis.

The functional data presented in the aforementioned study shows indications that after cross-talking TCR or BCR (but not the CD5 receptor) of peripheral blood from both T and B cells, homozygotes for the derived Val allele present lower lymphocyte activation responses than Ala homozygotes, as measured by lymphocyte proliferation and cytokine release (IFN-γ, and IL-10). Such observation is in concordance with our previous finding of cells carrying the selected derived Val allele showing hyper-signaling responses and the negative regulatory function of CD5. The higher signaling response of Val carriers would further activate the negative regulation capabilities of CD5, which will be translated in an increase of TCR/BCR threshold, and will ultimate lower the lymphocyte proliferative responses. On the other hand, carriers of the ancestral Ala variant showing deficient signaling responses, will maintain low the Ag-binding TCR/BCR thresholds and will show hyper lymphocyte proliferative responses that could eventually influence autoimmune disorders.

SLE is a systemic chronic complex autoimmune disease that is characterized by hyper-reactive T and B cells, auto-Ab production and immune-complex deposition. Genetic predisposition to SLE is the result of the contribution of the combined effect of multiple variants from a large number of genes, each allele contributing only mildly. Associated loci identified in the last years thanks to GWAs only account for around 15% of the heritability of the disease. Some of the identified alleles contribute specifically to different clinical manifestations of SLE, leading to early and severe forms of SLE. The aberrant function of T and B cells seen in SLE patients could be the results of deficient control of Ag-receptor-mediated T and B cell signaling mediated by the action of regulatory molecules involved in homeostasis of immune response and/or maintenance of self-tolerance, as for example that could be the case of CD5 receptor, as we have demonstrated. Indeed, *CD5* expression is up-regulated in many autoimmune disorders (Youinou and Renaudineau 2011).

In line with such observations and our functional results, we found a significant association with the ancestral defective-signaling hyper-proliferative Ala variant and a severe clinical form of SLE denominated SLE Nephritis. Of note, the positive selected adaptive variant shows, in this case, pleiotropic effects influencing a better diagnosis in an autoimmune diseases, and not to poorer diagnosis, as it has been reported for most of others adaptive alleles (Barreiro and Quintana-Murci 2010)(Raj et al. 2013) that are examples of the so called antagonist pleiotropy. The antagonist pleiotropy hypothesis claims that alleles that are beneficial for one trait in a certain moment of a lifetime could be detrimental for another trait, in a different stage (Williams 2001). Recent evolutionary thinking is beginning to address the differences between adaptive alleles contributing differently to disease pathogenesis in terms of their evolutionary time of apparition, although not much is known in this aspect (Di rienzo, personal communication)

Our preliminary new results reveal functional evidences of the importance of CD5 receptor as an immune-modulator molecule in lymphocyte activation and support its involvement in SLE pathogenesis, which was previously unknown. Taking into account that multiple disease-related genes are shared between different autoimmune disorders (Richard-Miceli and Criswell 2012), and both implication of *CD5* in SLE and RA, it would be worthy to further investigate whether *CD5* could be a new shared genetic factor in autoimmunity.

However, whether such new functional properties and disease association can tell us something more about its role in past adaptation or not, needs to be further discussed. In this regard, the recent knowledge of CD5 expression as a molecule marker of B10 regulatory cells is interesting. B10-reg cells are known to mediate the host response against pathogens by suppressing immune responses, and consequently, preventing harmful hyper-proliferative responses in host tissues. Such regulation is thought to be modulated by IL-10 release. It has been suggested that as a host-pathogen coevolution outcome, pathogens could have developed mechanisms to induce B-cell proliferation and, therefore, increase IL-10 production. That would suppress immune responses to infection and would allow the pathogen to longer persist and increase its chances to infect the host tissues, as it has been shown for the protozoan *Leishmania major* and the nematode *Brudia pahangi* (DiLillo et *al*, 2010). On the other hand, the adaptive state of CD5, that is the isoform carrying Val and present in most East Asian populations, shows reduced B-cell proliferation and, thus, less IL-10 production compared to the ancestral isoform. Such immune response could have conferred pathogen resistance to East Asians. By allowing longer and more successful immune responses after infection, carriers of the Val471 variant would increase their chances to survive to a geographically restricted pathogen.

In summary, we have provided convincing evidence of the important role of CD5 in modulating the immune responses, both in the maintenance of the immune homeostasis and after infection by pathogen fungi. We have also proved how the inferred putative adaptive variant directly influences such processes. Moreover, an implication in the pathogenesis of an autoimmune disease also seems to be revealed. Hence, we have demonstrated that evolutionary approaches can help to identify new functional alleles influencing present-day disease pathogenesis. According to our results, East Asian and Native American populations are predicted to present a less severe clinical form of autoimmune diseases due to the lesser proliferative responses led by the adaptive Val471 variant. However, it is unlikely that the allele was selected for this reason, as these diseases have emerged only in our modern history and do not seem to compromise sufficiently our fitness. It is more likely that the autoimmune trait is a secondary pleiotropic product of a past process of adaptation. Due to the role of CD5 in host defense and being the interaction with pathogen one of the stronger selective pressures in the human evolutionary history, it seems reasonable to think that the adaptive variant somehow conferred resistance to past pathogen encounters. Although due to the broad functional repertory of CD5, other scenarios cannot be fully discarded, we have provided different explanations about how the selected variant could have conferred pathogen resistance to East Asians.

2. The adaptive role of the ZIP4 gene in Sub-Saharan Africans

In the work presented in chapter 2 of the results section, we have described our second case of human adaptation. We provided compelling evidence for positive selection at the human zinc transporter ZIP4, which is coded by the *SLC39A4* gene*,* in this case in Sub-Saharan populations.

Prior to our work, a genome-wide scan studying genetic differentiation on genotype data from the HapMap project revealed that the rs1871534 polymorphism (leading to a Leu to Val change at codon position 372) is one of the non-synonymous variants with the largest differentiation among human populations (Barreiro et al. 2008). Thanks to extensive genotyping, compilation of public data for this polymorphisms and resequencing of a Neanderthal sample, we have been able to completely detail the worldwide geographical allelic frequency distribution of the polymorphism (see figure 1) and the unclear ancestral state. First, we observed that, surprisingly, the derived state is the most frequent allele in Sub-Saharan African populations, where it is often fixed. Second, we observed that outside Africa, the highest frequencies of the derived allele were found in Europe and West Asia.

Because of such extreme differentiated allele distribution between African and non-African populations, we suspected that positive selection could have driven such differentiation pattern.
High F_{ST} values are often produced by positive selection; however, they are usually accompanied by other molecular signatures of positive selection (Barreiro et al. 2008) (see Introduction, section 2). Enigmatically, we did not observe any other signatures of

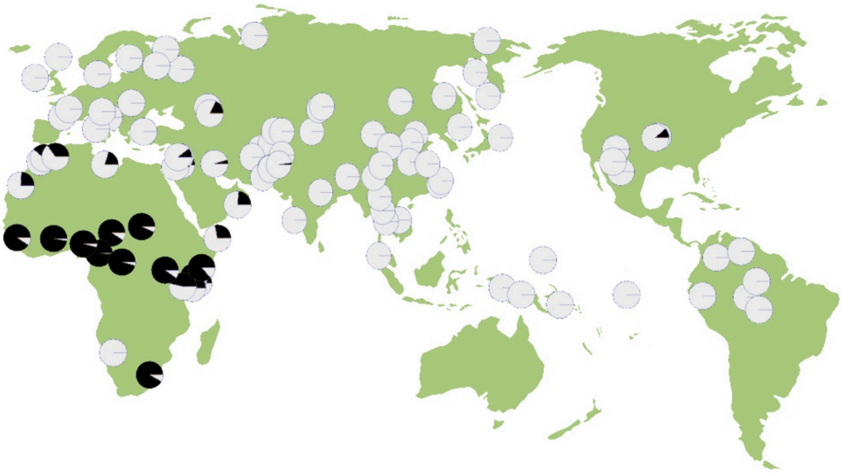


Figure 1. Worldwide allele frequency distribution for the Leu372Val polymorphism. White represents ancestral state, while black represents derived state.

positive selection, such as allele frequency spectrum deviation from neutral expectations or extensions of haplotype blocks, as it is expected when adaptation has occurred The identification of a recombination hotspot located in the same position where the *SLC39A4* gene is found, made us wonder whether such an extreme pattern of differentiation could be explained by drift only, or whether possible additional molecular signatures of positive selection in this genomic region could have been erased by recombination.

To answer this question, we took advantage of coalescent simulations one more time. In this case we simulated the scenarios of neutrality and of directional selection acting on a new mutation which appeared in Africa shortly after the split from non-African populations . We used the parameters described by Schaffner et al. (2005) to take into account the demography of the studied populations and simulated both scenarios in presence or in absence of the recombination hotspot. Under the directional selection scenario, three different selective coefficients were considered.

Our results revealed that the observed data could not be explained solely by neutrality or positive selection. On the contrary, a scenario of a long-lasting intermediate selective coefficient acting together with a mild recombination hotspot fits best the observed variability. It also explains the presence of a still non-fixed derived allele segregating at high frequencies in African populations.

Thanks to our simulations, we were able to confirm that an incomplete selective sweep has likely provoked this extreme allele frequency differentiation pattern, and that other molecular signatures of selection were probably erased by recombination. This is, as far as we are concerned, the first example described of a selective sweep that is blurred due to a particularity of the genomic architecture: in this case, the presence of a recombination hotspot.

However, this demography did not allow for additional populations such as the ancient and early divergent branches groups of Pygmy and/or San population(Veeramah et al. 2012), which likely represent additional useful information on this selective event. In order to sketch out more realistic but also more complex evolutionary scenarios,

Curiously, they show opposite ancestral allele frequency trends, the allele being fixed in the former, and almost absent in the latter. This fact, together with the intermediate allele frequencies seen in both North Africans and people from Middle Eastern populations, could question the date of the appearance of the derived allele, or the geographic areas where natural selection shaped the frequency of this SNP (see figure 1). Different scenarios could explain this allele distribution:

i. We could claim that the ancestral state of the polymorphism is present in San and non-Africans populations and that the derived allele was selected, after the Out-of-Africa migrations, in Bantu-speakers, which subsequently have migrated from their homeland in West Central Africa to most of the southern half of the continent. In this scenario the ancient Pygmy population would be expected to retain high frequencies of the ancestral allele, which is not the case. However, due to the recent admixture of these populations with other Bantu-speakers ones, we cannot reject this scenario.

ii. The derived allele may have appeared before the divergence of the San and Pygmy populations. The absence of this allele in a sample of 6 individuals could be due to chance, since it is such a small sample. The population that migrated out of Africa went through a population bottleneck, which may explain the low frequency of the derived allele in non-African populations. In addition or alternatively, natural selection could also be acting against the derived allele outside Africa

iii. Alternatively, the absence of the derived allele in the San sample could be explained by recent population specific selection against the derived allele in the San.

At this point, none of the three scenarios cannot be firmly rejected. However, the most parsimonious scenario seems to be one in which only one mutation appears in Sub-Saharan Africa -either after the migrations out of Africa or before- and it is then subsequently swept.

Ideally, dating the allele appearance would help elucidating the evolutionary story of adaptation. However, current methods rely on the length of haplotypic blocks (Beleza et al. 2013). Unfortunately, this task is not possible for this allele since the presence of a recombination hotspot has broken the expected haplotype-block structure.

Other, more complex situations have also been discussed in our manuscript, as for example, the possibility that the allele was already segregating in the population prior to the selection to happen, or that the observed allelic differentiation could be the result of gene surfing of any of the variants in a wave of population expansions. However, our simulation results, together with our functional results argue for positive selection targeting ZIP4.

As described in section 6.2 of the introduction, the maintenance of a balanced metal homeostasis is a key process to ensure the proper function of many cellular processes which are metal-dependent. Moreover, since micronutrients are essential elements that are obtained exogenously, many mechanisms have appeared during the evolution to guarantee the acquisition of adjusted metal levels under different critical physiological conditions (as for example, deficiency or excess). Some evidence for this is the large number of existing micronutrient transporters which are differentially expressed in cellular types, developmental stages and even, in cellular compartments, and also the fact that they are tightly regulated by multiple mechanisms, both at transcriptional and post-translational levels, including responsiveness to their own presence.

Importantly, the study of ZIP4 represents the first example of human adaptation involving a micronutrient transporter. In our study, we researched the possible functional effect of the Leu372Val change at ZIP4 transporter, which besides providing evidence for a possible adaptive role of this non-synonymous SNP may help clarifying which adaptive phenotype could have been favored in African populations. The wide distribution of both alleles suggested that any possible phenotypic change does not likely have a large impact on protein functions, but more likely a mild effect that may be difficult to detect in our heterologous system. In order to scale and distinguish known pathological effects from more physiological changes, we decided to include two mutations at the same position leading to Acrodermatitis

Enteropathica, (AE), a disease phenotype characterized by zinc deficiency.

Because zinc is known to be deficient in many regions of Sub-Saharan Africa, and because zinc deficiency leads to many physiological and developmental impairments (Cole et al. 2010), we speculated that the change, if detectable, could have been involved in optimizing intracellular zinc levels by increasing zinc uptake. However, our experiments in zinc transporter revealed a completely opposite trend. To our surprise, the uptake experiments performed showed that carrying the derived Val present in most African populations lead to significantly less ZIP4 cell surface expression, which translates into significantly reduced zinc transport when compared to the ancestral Leu, present in non-African populations. On the other hand, the AE mutations lead to failure of ZIP4 trafficking to the cell membrane, and thus, show almost absent zinc uptake, as was expected. We were able to demonstrate that neither the Val nor the Ala polymorphic variants behave as the pathological mutations.

Since reduction of zinc uptake cannot explain adaptation to zinc deficiency, as we had speculated, we had to invoke other evolutionary scenarios. Interestingly, the allelic distribution of the derived allele follows a cline distribution with a hot nucleus in West African populations. This pattern has been previously observed in several environmental factors, but also in other adaptive alleles, as for example the Duffy locus known to protect African populations from malaria (see figure 2). Infectious diseases in Africa are a major threat, both because the economic situation of most of the countries does not allow relying on a proper health system that can effectively prevent and treat the infectious diseases –meaning they easily spread and become epidemic- ; but also because the prevalence of parasites in these warm tropical regions is very elevated.

Interestingly, the allele distribution seen for the Leu372Val polymorphism matches quite well the distribution of some of the major pathogens in Africa, such as the protozoans *Plasmodium sps* and *Trypanosomiasis sps* or the helminth *Ascaris sps* (see figure 2). Although correlation does not directly implies causation, such observations lead us to wonder whether the presence of the Val allele in Africans could be somehow related to the host-defense adaptation to pathogens, as observed, for example, for the Duffy allele in Africans.

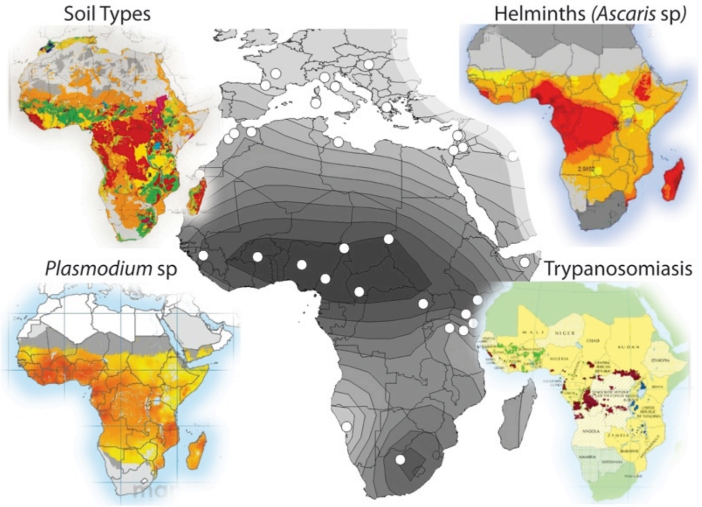


Figure 2. Cline distribution of Val372 variant correlates with particular pathogen and soils types distribution in Africa. Adapted from (Bethony et al. 2006) (Simarro et al. 2010) (Hay et al. 2009).

According to the nutritional immunity hypothesis, the host often restricts cellular metal availability, which is needed for the pathogen to grow and complete its life cycle, and thus pathogens become less virulent and do not compromise the host survival. This process has been very well described for iron, and it was shown that hypoferremia is an acute phase response to infection (Hood and Skaar 2012); likewise, it has also been recently shown that recruiting zinc and manganese is also a host-defense strategy (Kehl-Fie and Skaar 2010). Likewise, the less zinc is available the more protected would be the host, a scenario that completely matches our experimental data and the correlation observed between allele frequencies and pathogen prevalence.

It could be, then, that carrying a Val allele would protect African individuals from very high frequent contact with different pathogens. Alternatively, the lower level of expression of the ZIP4 receptor observed for the Val allele could also have resulted advantageous if the presence of the receptor in the enterocyte membrane is used by particular infectious parasites to invade the host.

To understand the phenotypic effect of the Leu372 Val substitution at the individual level further evidence should be gathered. First, our experiments have been carried out at a cellular level, by over-expressing ZIP4 in transiently transfected cells in one type of epithelial cell line and by using supra-physiological levels of zinc to quantify its transport. However, very likely our results can be transferable to other cell lines, and also to the organismal level where the substitution could broadly affect physiologically its carrier, as it has been shown for other non-synonymous mutations in piglets in a homolog protein to ZIP4 (Esteve-Codina et al. 2013), and as it has been revealed by the physiological consequences of ZIP4 knockout studies (Geiser et al. 2012). Investigation of the impact of this polymorphism in real healthy tissues from individuals of different origins (Africans vs non-Africans), by means of total protein expression and zinc concentration quantification plus, a case-control study of tissues from different diseases in which zinc is involved could help us to understand how the polymorphism could further influence the phenotype and the pathogenesis particular diseases. Indeed, zinc levels are known vary and influence many disorders as cancer or diabetes (Rink and Haase 2007), which makes us think that our polymorphism could have a role as a susceptibility factor in such diseases, as it directly influences the ZIP4 expression.

Lack of evidence regarding a clear relation between this non-synonymous SNP to any complex phenotypic trait makes it difficult to further speculate about alternative adaptive hypotheses. Unfortunately the rs1871534 SNP is not included in most common genotyping chips that have been used in recent genome-wide analysis of infectious diseases or other complex phenotypes, so its possible implication in disease has not been tested in large case-control association studies. Inexistence also of genome-wide studies of zinc concentration involving different ethnic comparisons also limits further conclusions about the global contribution of this, and other variants, to the global tissue metal content.

Apart from revealing a new case of human adaptation, our study is unique since it provides an example of a selected gene that could not have been detected by most current genome-wide methods for detecting positive selection in humans. Moreover, as it has been discussed above, it is the first time that a human polymorphism in a zinc transporter has been proven to lead to a functional change that impacts the function of the protein. Further implications for both points can be easily foreseen.

First, it is reasonable to think that there are many other examples of hidden molecular signatures of possible selection in the human genome. The recognition of such hidden or incomplete selective sweeps that other genome-wide scans of positive selection have failed to detect, could help to explain, for example, the almost complete absence of classical selective sweep events in Africans (see Introduction, section 2.2).

Second, base on the relevance impact of this polymorphism in human populations, we hope that new research lines in evolutionary biology and also in medical genomics will intend to clarify the contribution of micronutrient metabolism transporters and related elements to adaptation, and also their implication in disease, something that to date, has been very little covered.

3. Evolutionary trends in coding and non-coding elements from pathways with differentiated patterns of divergence.

In chapter 3 of the results section, we have presented a pathway-based study of the action of natural selection on coding and non-coding genic elements (i.e. promoters, UTRs, introns and trailers). In our study we aimed to contrast the rates of adaptive evolution among the different genomic elements and pathways analyzed following the McDonald-Kreitman test (MKT) comparative framework of combining polymorphism patterns within a species (in chimpanzees in our case) with divergence data (between humans and chimpanzees). To that end, we performed an extended custom enrichment with the Agilent technology and resequenced the regions of interest in 20 chimpanzee individuals using an Illumina instrument.

Due to the use of the so-called Next Generation Sequencing Technology, we have been very strict about including only sites that show high quality sequencing information for all the individuals. In the MKT derivative approach that we applied, the SFS of a neutral reference is used to infer both the demography for the population whose polymorphism is analyzed as well as the distribution of fitness effects on the functional sites when estimating the proportion and rates of adaptive evolution.

Furthermore, as in our case, the adaptive rates for all elements and pathways were estimated using the same neutral reference, which consisted in the concatenation of all 4-fold synonymous sites. Although it is not possible to completely exclude sites under the influence of selection in any neutral reference, our approach should avoid any variability of α and ω_α values due to the use of different neutral references. Moreover, although the rates of adaptive evolution can be somehow biased by different factors (see

Introduction, section 2.3), the use of a common neutral reference allows us to make proper comparisons between pathways in each class of element analyzed and to reveal true differences between them. On top of that, our framework allowed us to analyze the evolutionary trends in coding and non-coding sequences without requiring different theoretical approaches.

Thanks to our study, we have been able to confirm that the signatures of acceleration previously inferred from divergence data both at the coding regions of the Complement pathway and at the intronic sequences of our set of Accelerated introns are, indeed, a result from adaptive evolution. Moreover, and very intriguingly, we have demonstrated that high fractions of adaptive substitutions are not only distributed around the exons or introns with higher *d*N/*d*S (ω) values, but in the whole collection of elements analyzed (i.e. we find high ω_α values in the coding regions of all genes of the Complement pathway and in all the introns of the Acc Intron dataset). This suggests that the method applied is very efficient in detecting adaptive evolution, even in regions that had not previously been identified as rapidly evolving when using only divergence-based methods.

In the case of the Complement pathway, such observations suggest that the whole set of coding regions of the pathway is likely evolving under similar selective pressures. This is congruent with the biological function of the pathway, and also with the parallel action of natural selection on the particular set of functional related elements configurating the pathway (see Introduction, section 4.2). Elements from the complement system are involved in the early response to infections, and thus, are part of the innate immunity. As I have extensively illustrated in the introduction, there is plenty evidence of pervasive positive selection among elements of the immune system (section 6.2). Most studies on this topic have identified the role of individual genes related to immunity in population and species adaptation to pathogens (as we have done in our two previous chapters in the results section). Similarly to a recent (Daub et al. 2013), we investigate adaptation to pathogens in the chimpanzee lineage using a pathway framework, and provide new evidence for the action of polygenic adaptive evolution within the pathway. In this case, such polygenic adaptation is not due to the accumulation of incremental changes in allele frequencies on multiple genes affecting a particular phenotypical trait, but to a significant accumulation of adaptive substitutions in the coding regions of the whole pathway, which will be evolving in a coordinated mode.

As for the set of genes showing introns with accelerated evolution, we had no initial evidence that they were functionally related, so the scenario of parallel evolution between different functionally interconnected genes cannot be invoked. However, our results in this set of genes nicely reveal that if a particular non-coding element in a gene such an intron is involved in adaptation, it is likely that its other accompanying elements of the same type (i.e. remaining introns) are under the same selective force. In addition, we can also see how other non-coding elements, such as promoters and trailers of the same genes (for which we had no previous evidence of high evolutionary rates), also seem to display high rates of adaptive evolution. By contrast, this trend is not seen on their corresponding coding sequences. This observation suggests that there is, at least in this set of genes, a general decoupling of evolutionary trends among their coding and non-coding sequences. It would be worthy to further investigate this tendency genome-widely, and to elucidate whether specific sets of genes, or specific traits, are more likely to accumulate adaptive substitutions and show higher rates of adaptive evolution in their regulatory regions than in their coding sequences, or *vice versa*.

This issue has been partially addressed by Haygood et al. (2010), which work was based on surveys of positive selection in humans. In this case, the authors showed that some functional categories (such as those related to neuronal activity) showed enrichment in signatures of positive selection in their non-coding elements, while others (such as immunity and defense) showed enrichment in their coding elements. Interestingly, we detect the same tendency in the coding regions of the Complement pathway, although we have no evidence of it in the putative regulatory regions of the two neurological related pathways that we included in our study focused on the chimpanzee lineage.

What might account for the strong contrast in the distribution of signals of positive selection between different functional categories is an open question. I do speculate, as others have previously done, that genes that are expressed in many different tissues and whose function is essential do not usually undergo changes in their coding regions, as they usually are recessive and this would disrupt their function. In these types of genes, mutations in the non-coding regulatory sequences could be more easily accepted and selected as they could favor an advantageous change in gene expression, in a given tissue or cell-type. On the other hand, genes that are somehow redundant in function as it is the case of many immune-related genes- could easily accumulate mutations in their coding regions which could favor new adaptive traits without highly compromising a correct immunological function. To answer these open questions, I can foresee a follow-up study consisting of a genome-wide pathway-analysis based on MKT. Applying this method to the high coverage of sequencing data for the individuals of the 1000 Genomes Project and the recently published Great Apes (Prado-Martinez et al. 2013) would surely provide insights about the contribution of different pathways and genic elements to the adaptation of both species.

Finally, all genic elements from the neurological pathways investigated seem to display constrained patterns of evolution in the chimpanzee population although clearly not as extreme as in our purifying selection reference set in CDS (i.e. the Actin pathway). Interestingly, whereas Parkinson's CDS present a tendency towards a high adaptive rate in comparison to most of the pathways analyzed but the Complement, introns in the Parkinson pathway display significantly lower adaptive rates with all the remaining pathways analyzed but the Actin. This pattern further extends the previous observation of a decoupling trend of evolution between the coding and non-coding regions of the chimpanzee genome.

Concluding Remarks

The first two studies of this thesis provide functional and evolutionary characterization of two different adaptive variants in humans, which illustrate well the variety of approaches that might be undertaken depending on the *a priori* evidence for the action of positive selection provided by genome-wide studies.

For example, the case of the *CD5* gene represents a nice example of a classical signature of positive selection. In this study, we had clear initial evidence for the presence of a selective sweep in East Asian populations along a large genomic region of around ~400 kb. Part of the importance of the study resides in the identification of the adaptive variant within a high gene-density region by exploring the complete genetic variability of the region in detail. The adaptive variant in this case is almost fixed in the East Asians, where it has swept other derived neutral linked variants to high frequencies, creating a typical signature of hitchhiking. On the contrary, in the study of *ZIP4* gene, we only had three highly differentiated SNPs one of which was one of the most highly differentiated non-synonymous variants in the genome; however, no additional evidence of selection was found. In this case, an important step in the study was to prove that the observed variability pattern is compatible with the action of positive selection together with the presence of a hotspot in the same region where the gene is located.

An important point shared by both studies is the functional characterization of the molecular relevance of these non-synonymous substitutions and, in particular, the understanding of the functional impact of the allele that selection has favored in relation to a putatively adaptive phenotype. Without doubt, the opportunity to perform functional studies in collaboration with other research groups has strengthened both evolutionary studies.

Furthermore, the adaptive variants identified in both cases are functional variants which could also have important epidemiological implications (i.e. they may be influencing differences in health and disease among different human populations). In the case of *CD5*, new preliminary results indicate that the selected variant plays a role in determining susceptibility to certain autoimmune diseases, which has also been described for other genes involved in adaptations driven by pathogens, as explained by the hygiene hypothesis. Further research on the role of *CD5* and regarding the selected variant will help to better understand such pathogenic role as well as to the development of new therapeutic strategies. As for the *ZIP4* case, potential epidemiological consequences can be easily foreseen and have been presented but need to be further confirmed.

Finally, in the last chapter of the thesis we go beyond the study of candidate genes to explore how natural selection acts in functionally related genes (i.e. pathways). Our work demonstrates that the joint use of polymorphism and divergence data can yield important insights to study adaptive evolution and provides evidence supporting the existence of differentiated rates of adaptive evolution among different pathways, and among their corresponding coding and putative regulatory elements.

We hope that our meaningful results will motivate new studies to (i) reveal new examples of human adaptation, especially in cases where the signature of the selective sweep has been erased; (ii) further explore the implication of the *CD5* and *ZIP4* genes in disease phenotypes; and (iii) further understand the contribution of different pathways and genomic elements to adaptation.