Gene regulatory factors in the evolutionary history of humans

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

Dissertation

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet
Informatik

vorgelegt von
MSc. Alvaro Perdomo-Sabogal
geboren am 08. Januar 1979 in Armenia, Quindio (Kolumbien)

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Peter F. Stadler, Institut für Informatik, Leipzig
2. Prof. Dr. Andrew Torda, Zentrum für Bioinformatik, Hamburg

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 24. August 2016 mit dem Gesamtprädikat “magna cum laude"

Abstract

Changes in *cis*- and *trans*-regulatory elements are among the prime sources of genetic and phenotypical variation at species level. The introduction of *cis*- and *trans* regulatory variation, as evolutionary processes, has played important roles in driving evolution, diversity and phenotypical differentiation in humans. Therefore, exploring and identifying variation that occurs on *cis*- and *trans*- regulatory elements becomes imperative to better understanding of human evolution and its genetic diversity.

In this research, around 3360 gene regulatory factors in the human genome were catalogued. This catalog includes genes that code for proteins that perform gene regulatory activities such DNA-depending transcription, RNA polymerase II transcription cofactor and co-repressor activity, chromatin binding, and remodeling, among other 218 gene ontology terms. Using the classification of DNA-binding GRFs (Wingender et al. 2015), we were able to group 1521 GRF genes (~46%) into 41 different GRF classes. This GRF catalog allowed us to initially explore and discuss how some GRF genes have evolved in humans, archaic humans (Neandertal and Denisovan) and non-human primates species. It is also discussed which are the likely phenotypical and medical effects that evolutionary changes in GRF genes may have introduced into the human genome are; for instance, speech and language capabilities, recombination hotspots, and metabolic pathways and diseases.

In addition, by exploring genome-wide scan data for detecting selection, we built a list of GRF candidate genes that may have undergone positive selection in three human populations: Utah Residents with Northern and Western Ancestry (CEU), Han Chinese in Beijing (CHB), and Yoruba in Ibadan (YRI). We think this set gathers genes that may have contributed in shaping the phenotypical diversity currently observed in these three human populations, for example by introducing regulatory diversity at population-specific level. Out of the 41 DNA-binding GRF classes, six

groups evidenced enrichment for genes located on regions that may have been target of positive selection: C2H2 zinc finger, KRAB-ZNF zinc finger, Homeo domain, Tryptophan cluster, Fork head/winged helix and, and High-mobility HMG domain. We additionally identified three KRAB-ZNF gene clusters, in the chromosomes one, three, and 16, of the Asian population that exhibit regions with extended haplotype homozygosity EHH (larger than 100 kb). The presence of this EHH suggests that these three regions have undergone positive selection in CHB population. Out of the 22 GRF genes located within these three KRAB-ZNF clusters, seven C2H2-ZNF GRF genes (*ZNF695, ZNF646, ZNF668, ZNF167, ZNF35, ZNF502,* and *ZNF501*) carry nonsynonymous SNPs that code nonsynonymous SNPs that change the amino acid sequence in their protein domains (linkers and cystine-2 histidine-2 amino-acid sequence motifs). Six GRF genes located on the EHH region on the chromosome 16 of CHB have been associated with obesity (*KAT8*, *ZNF646*, *ZNF668*, *FBXL19)* and blood coagulation (*STX1B* and *VKORC1*) in humans. In addition, we also detected genetic changes at GRF sequence level that may have resulted in subtle regulatory changes in metabolic pathways associated with glucose and insulin metabolism at population-specific level.

Finally, acknowledging that a representative fraction of the phenotypic diversity we observed between humans and its closely related species are likely explained by changes in *cis*-regulatory elements (CREs), putative binding sites of the transcription factor GABPa were identified and investigated. GABPa is GRF protein member of the E-twenty six DNA-binding proteins class. GABPs control gene expression of many genes that play key roles at cellular level, for instance, in cell migration and differentiation, cell cycle control and fate, hormonal regulation and apoptosis. Using ChIP-Seq data generated from a human cell line (HEK293T), we found 11,619 putative GABPa CREs were found, of which 224 are putative human-specific. To experimentally validate the transcriptional activity of these human-specific GABPa CREs, reporter gene essays and knock-down experiments were performed. Our results supported the functionality of these human-specific GABPa CREs and suggest that at least 1,215 genes are primary targets of GABPa. Finally, further analyses of the data gathered depict scenarios that bring together transcriptional regulation by GABPa with the evolution of particular human

speciation and traits for instance, cognitive abilities, breast morphology, and lipids and glucose metabolic pathways and the regulation of human-specific genes.

By studying genetic changes in *cis*- and *trans*- regulatory elements in humans in two different evolutionary time frames, species evolution and population genetics, we were able to show how genome regulatory innovations and genetic variation may have contributed to the evolution of human- and population specific traits. Here, we conclude that human-specific changes in regulatory elements are likely introducing subtle regulatory variation in key pathways at physiological level, for instance, in glucose/insulin, lipids metabolism and cognitive abilities. Some of these changes may have resulted in adaptive responses that left signatures of positive selection at human population specific level.

Acknowledgment

I would like to thank those who significantly helped me to keep myself together, motivated and combative when the difficult times arrived. This mainly refers to my sister Ericka, family, and best friends. Special memorable thanks go to the one who already left, my dad, who before leaving put incredible effort in making of this, a feasible trip into this scientific career.

I would also like to thank to Katja, for her commitment with the supervising process, for her valuable time and massive patience during the academic discussions.

Many thanks to Lydia and Daniel, who apart from being devotedly committed with the Friday's friendly catching up sessions, also had the time to provide help and feedback when the times were cloudy and the codes were messy.

Many thanks to Peter for generating a space with the facilities, the people and the environment to do proper scientific research.

Many thanks to Jens and Petra who were always available for sorting out logistic issues.

Many thanks to the folks from the Bioinformatics institute who always provided me with a friendly and enjoyable environment.

Contents

Abbreviations

Alu:	mobile elements in the genome
bp:	base pairs
CEU:	Utah Residents (CEPH) with Northern and Western Ancestry
CHB:	Han Chinese in Bejing, China
ChIP-Seq:	Chromatin immunoprecipitation followed by sequencing
CMS:	Composite of multiple signals test
CREs:	*cis*-regulatory elements
DE:	Differentially expressed
DBDs:	DNA-binding domains
DNA:	Deoxyribonucleic acid
EHH:	Extended Haplotype Homozigosity.
Eq:	Equation
eQTL:	Expression quantitative trait loci
ETS:	E-twenty six transcription factors
EREs:	Endogenous repetitive elements
ERVs:	Endogenous retroviruses
FDR:	False discovery rate
FPKM:	Fragments Per Kilobase of transcript per Million mapped reads
GRFs:	Gene Regulatory Factors
HYK:	Hasegawa, Kishino and Yano substitution model
iBAQ:	Intensity-based absolute quantification
IDs:	Stable identifiers
Kb:	Kilo base pairs
Kya:	Kilo years ago.
KRAB-ZNF:	Krüppel associated box zinc finger gene

LCT:	Lactose gene
LD:	linkage disequilibrium
LINEs:	long interspersed nuclear elements
lncRNA:	Long non-coding RNA
MAST:	Motif Alignment and Search Tool
MAF:	Minor Allele Frequency
MEME:	Multiple EM for Motif Elicitation
NGS:	Next generation sequencing
PWM:	Position-specific weight matrix
SINE:	Short Interspersed Nuclear Elements
SVA:	SINE-VNTR-Alu, a composite hominid-specific retrotransposon family
TFs:	Transcription Factors
TSS:	Transcription Start Site
UCSC:	University of California Santa Cruz Genome Browser
VNTR:	variable number tandem repeat
YRI:	Yoruba population in Ibadan, Nigeria

Introduction

The past fifty years have seen the development and application of numerous experimental techniques and statistical methods to identify genomic elements that might be playing important roles in shaping genetic and phenotypical variation, and diversity within species, species diversification and speciation. From a broad perspective, these strategies have been used to explore evolution at different scales and in an extensive range of organisms. Evolutionary events can be explored into two distinct hierarchical time frames: population genetics and species evolution (Sesink Clee and Gonder 2012) These different time scales encompass either the diversification within a given species, where the offspring has the very similar genetic background as the ancestor, or the large-scale patterns in which the origin of new species from previously existing or extinct ancestral types has taken place (species evolution) (Eldredge 1989) (Figure 1). Population genetics, as evolution on a small scale, mainly integrates mechanisms that causes changes in the allele frequencies in the gene pool in a period of time that covers from few to several generations for a population (Reznick and Ricklefs 2009). The main mechanisms driving population genetics processes mainly act at population's level, where mutation, migration, drift and selection (positive, negative, balancing) accumulate changes over time. The effects and intensity of these processes and their relation with the environment may lead to population differentiation or to the birth of new species trough the continuous accumulation of small and horizontal changes over time (Dobzhansky 1941) (Figure 1). In contrast, species evolution, understood as the evolutionary change at the level of species, takes place at higher levels, resulting in large and complex changes, thus giving origin to new groups, for instance, families and genera (Valentine and Jablonski 2003; McGowen et al. 2014). Studying species evolution involves the understanding of processes like developmental constraints, variation in the diversification rates (species selection), speciation, and extinctions,

among others (Figure 1). Nonetheless, the processes resulting in strong genetic differentiation and speciation are both hard to be observed.

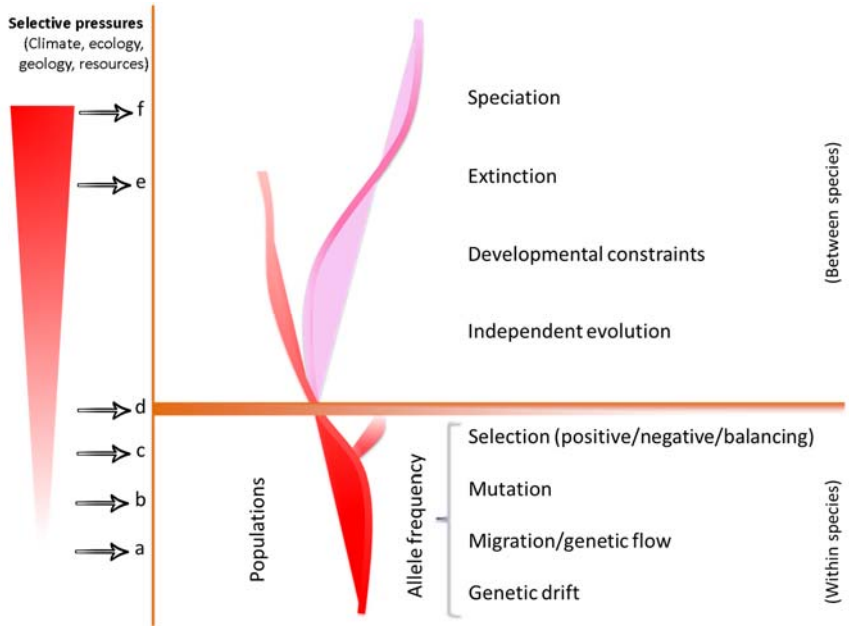


Figure 1. Hierarchical relations between population genetics and species evolution in context of environmental influence. The lower level depicts organisms arranged within populations, their interactions and the association with the environment. Left axis (red) indicated the time scale. (a-f) describe some of the dynamics governing the evolutionary changes. (a) Mutation and local adaptation. (b) Insufficient resources causing migration. (c) Reduction in gene flow causing population differentiation. (d) Geographical barrier (isolation by distance). (e) Environmental challenge to which the species cannot rapidly adapt to (ecological success). (f). Speciation/Extinction

1.1. Emergence of modern humans

From a population genetics perspective, deciphering the molecular mechanisms that underlie the evolutionary history of anatomically modern humans (AMHs), their current genetic and phenotypic diversity, is among one of the most compelling challenges in contemporary genomic and transcriptomic research. Following the advent of genetic data, three major competing hypothetical models of modern human origins are still widely used to explain the emergence of AMHs. The *multiregional model* (MRE), as well known as polycentric hypothesis, the widely accepted model of a recent African origin ("*Out of Africa*" hypothesis; RAO), both mainly based on archeological observations, and the *Assimilation* model. The *Assimilation* model integrates arguments of MRE and RAO, and adds new

perspectives from recent ancient DNA studies (Figure 2) (Bräuer et al. 1997; Gibbons 2011). The MRE suggests that the AMH's traits are the result of an interlinked and extensive social and biological network first established by ancestral species such as *Homo erectus* around 1.8 million years (Wolpoff et al. 1988). Such network was characterized by the constant social and cultural interaction, and the gene flow between the evolving Eurasian AMH populations. It is assumed that this facilitated the species-wide evolutionary change and promoted the local diversity we currently observe in AMH populations, and simultaneously prevented speciation (Figure 2) (Wolpoff et al. 1988; Wolpoff et al. 2000).

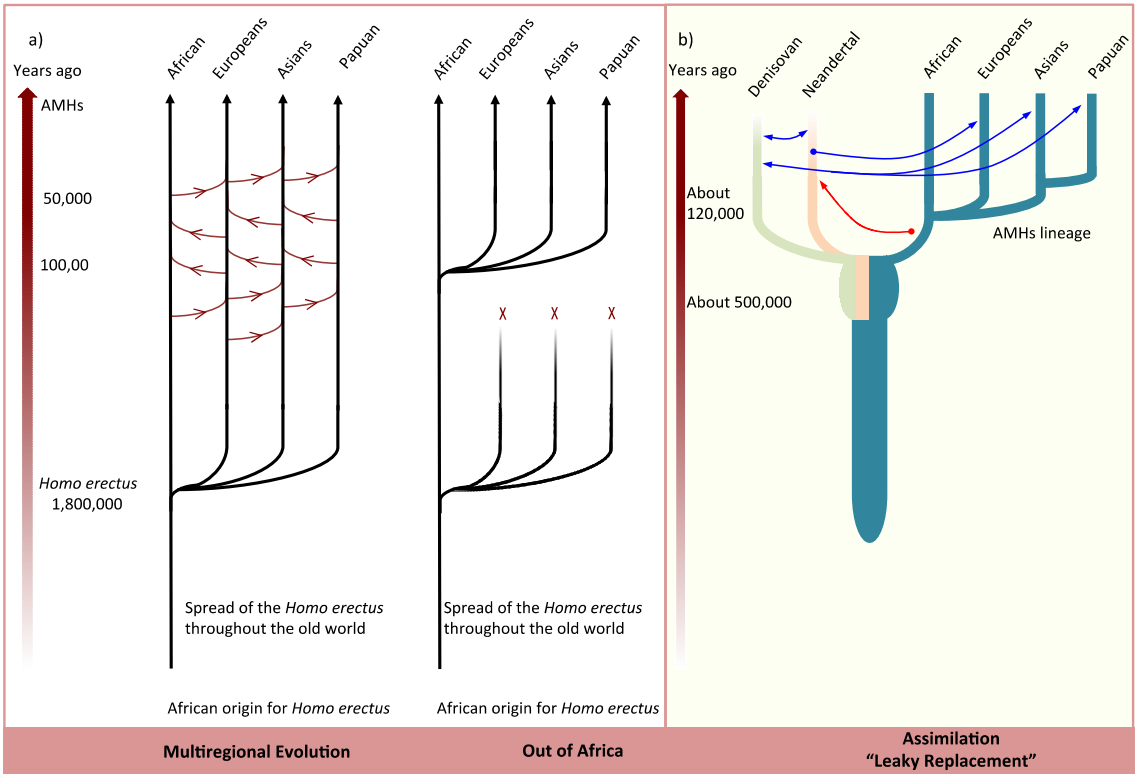


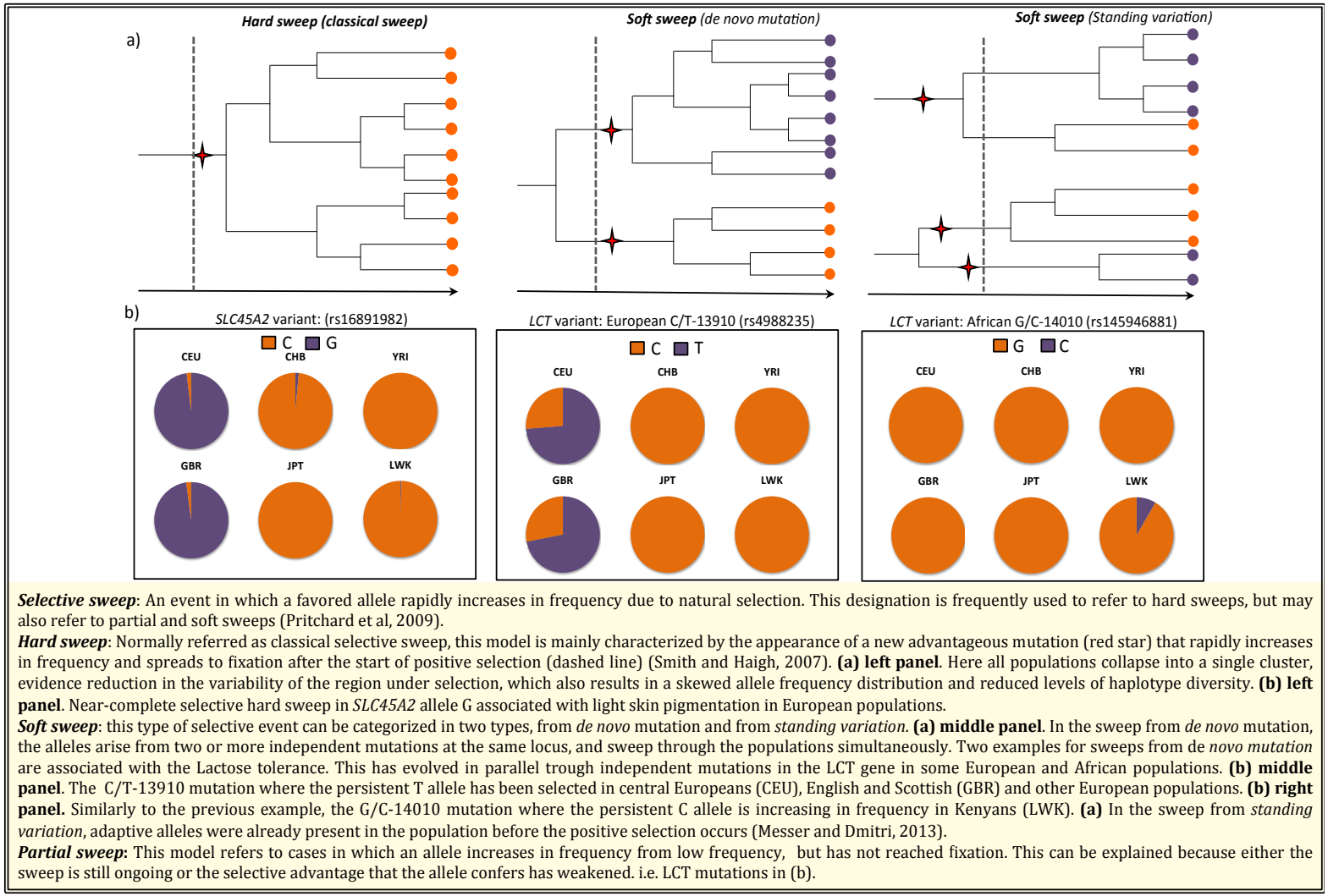
Figure 2. Diagram representing each of the three competing models. (a) Multiregional and Out of Africa models.. Red arrows indicate gene flow between populations in the multiregional model Modified from Gibbons (Gibbons 2011). (b) The illustration of the assimilation model. In this scenario. Blue arrows indicate gene flow between AMHs, Neandertal and Denisovan in, the gene flow was depicted as inferred by Kuhlwilm et al. (2016). Five interbreeding events are shown (blue arrows). Inferred gene flow from a population closely related to AMHs towards the Altai Neandertal (red arrow)(Kuhlwilm et al. 2016).

Conversely, the RAO model posits that AMH's features descent from a population with a geographic origin in Africa that spread throughout the continent and

migrated to Eurasia (Stringer and Andrews 1988) substituting other hominin species such as Neandertal and Denisovan (Gibbons 2011; Tryon and Bailey 2013). However, new genetic evidence suggests that interbreeding was common between some AMH populations and archaic hominin species such as Neandertal and Denisovan (Green et al. 2010; Reich et al. 2010; Kuhlwilm et al. 2016). Such evidence strongly supports the work of Bräuer et al. (1997), the leaky replacement, which was initially and for long ignored (Gibbons 2011; Tryon and Bailey 2013). This model introduces the replacement of archaic humans while hybridizing with some AMH populations followed by the decline in the populations of the latter two archaic humans (Figure 2).

Different scenarios of human dispersals and routes out of Africa have also been recently depicted (Oppenheimer 2009; Mellars et al. 2013; Veeramah and Hammer 2014; Reyes-Centeno 2016). Independently of the differences these scenarios may represent in the understanding of AMH's evolution, it is clear that the processes that followed AMH's migrations out of the African continent involved adaptation, colonization and interbreeding with other hominin species (Green et al. 2010; Reich et al. 2010; Alves et al. 2012; Kuhlwilm et al. 2016). While human adaptation to new conditions was taking place, it is assumed that an immense repertoire of genetic and phenotypic variation was simultaneously shaped. A representative amount of this variation has had a direct impact on the ability to face and survive different challenges such as changes in climate and diet, resistance to prevalent pathogens, and diseases (Vasseur and Quintana-Murci 2013). Accordingly, traits that confer particular advantages or disadvantages on the individual fitness are subjected to natural selection (Figure 3) (Gillespie 1991).

Several classical examples on how particular genes have bestowed adaptive traits in AMH populations, for instance, the Lactose persistence and melanin synthesis in Europeans, among others, are well documented.



Selective sweep: An event in which a favored allele rapidly increases in frequency due to natural selection. This designation is frequently used to refer to hard sweeps, but may also refer to partial and soft sweeps (Pritchard et al, 2009).
Hard sweep: Normally referred as classical selective sweep, this model is mainly characterized by the appearance of a new advantageous mutation (red star) that rapidly increases in frequency and spreads to fixation after the start of positive selection (dashed line) (Smith and Haigh, 2007). **(a) left panel**. Here all populations collapse into a single cluster, evidence reduction in the variability of the region under selection, which also results in a skewed allele frequency distribution and reduced levels of haplotype diversity. **(b) left panel**. Near-complete selective hard sweep in *SLC45A2* allele G associated with light skin pigmentation in European populations.
Soft sweep: this type of selective event can be categorized in two types, from *de novo* mutation and from *standing variation*. **(a) middle panel**. In the sweep from *de novo* mutation, the alleles arise from two or more independent mutations at the same locus, and sweep through the populations simultaneously. Two examples for sweeps from de *novo mutation* are associated with the Lactose tolerance. This has evolved in parallel trough independent mutations in the LCT gene in some European and African populations. **(b) middle panel**. The C/T-13910 mutation where the persistent T allele has been selected in central Europeans (CEU), English and Scottish (GBR) and other European populations. **(b) right panel.** Similarly to the previous example, the G/C-14010 mutation where the persistent C allele is increasing in frequency in Kenyans (LWK). **(a)** In the sweep from *standing variation*, adaptive alleles were already present in the population before the positive selection occurs (Messer and Dmitri, 2013).
***Partial sweep*:** This model refers to cases in which an allele increases in frequency from low frequency, but has not reached fixation. This can be explained because either the sweep is still ongoing or the selective advantage that the allele confers has weakened. i.e. LCT mutations in (b).

Figure 3. Types of selective sweeps observed in population genomics data.

In the first example, the Lactose gene (*LCT*) is associated with the capacity to assimilate and digest lactose, which dissipates during the childhood in humans. However, for some European populations the lactase metabolic activity persists in adults. As previously mentioned, AMH's genetic population differences may help us to identify traits that have conferred particular selective advantages, for instance, to changes in diet, sun exposure, malaria resistance, among others. In populations from the North of Europe, two allelic variants have been associated with lactase persistence, in a uniquely common haplotype that is present in around 77% of the individuals. The frequency and extension of this haplotype is considered a hallmark of natural selection on the human genome (Figure 3) (Bersaglieri et al. 2004). A similar evolutionary event has been described for the Solute Carrier Class 45 Member 2 (*SLC45A2*) gene, which is associated with mediating melanin synthesis in several species, including humans. This gene carries particular single nucleotide polymorphisms (SNPs) that have been associated with light skin color in Europeans (Soejima and Koda 2007), and melanin index variations in Indian populations (Jonnalagadda et al. 2016) (Figure 3).

1.2. Evolutionary changes after human lineage split from non-human-primates

From a species evolutionary perspective, it has been widely acknowledged that many phenotypical differences found between AMHs and non-human primates are mainly driven by changes in the mechanisms controlling gene expression rather than by the emergence of novel structural genes (Britten and Davidson 1971; King and Wilson 1975). Although, recent evidence suggests that clade or species-specific genes might be driving the evolution of new physiological functions at reproductive, brain, and immunological level, for instance, in humans and chimpanzees (Sudmant et al. 2010; Zhang et al. 2011; Geschwind and Konopka 2012). Thus, it may be the combinatorial effect between the evolution of gene regulatory elements, sequence specific changes in protein domains, and the appearance of new genes, among other molecular mechanisms, what have contributed in the evolution of species-specific traits.

Studies of diversity by using genomics and transcriptomics have importantly contributed to the understanding of the ways in which some genomic innovations and sequence changes in genes might contribute to generate alternative phenotypes between species. Changes at genomic level shaping the mechanisms in which proteins interact with DNA or with other proteins to control gene expression are considered essential in the evolution of species heterogeneity and phenotypic diversity (Wray 2007; Wittkopp and Kalay 2012). Gene regulation is mainly mediated by gene regulatory factors (GRF), a group of proteins that directly or indirectly interact with DNA to regulate the expression of other genes. Two widely explored genes that are examples of how genetic variability on GRF genes might have profoundly changed the evolutionary history of humans are Forkhead box protein 2 (*FOXP2*) (Lai et al. 2001; Enard et al. 2002; Konopka et al. 2009; Reimers-Kipping et al. 2011; Maricic et al. 2013) and PR domain-containing protein 9 (*PRDM9*). The FOXP2 protein carries two human-specific amino acids (Enard et al. 2002). It has been suggested that these two amino acids are connected with molecular processes moderating the evolution of speech and language, both distinctive human traits (Lai et al. 2001). Experimental research has shown that the set of genes that are regulated by FOXP2 differs between human and chimpanzee (Konopka et al. 2009), which suggest the likely effect that these two amino acids caused on altering regulation of gene expression and the evolutionary improved speech capabilities in humans. Recent evidence suggests that *FOXP2* variation may be important in language disabilities, but that the contribution of common variants to the language capability of an individual are unlikely to cause an appreciably effect (Mueller et al. 2016). In addition, it was also found that *FOXP2* exhibited signatures of positive selection in humans (Enard et al. 2002; Zhang et al. 2002) and that FOXP2 target genes show strong evidence of positive selection as well (Fisher et al. 2013). Another example of a GRF with relevance in the evolutionary history of humans is *PRDM9*, a zinc finger GRF protein with histone methyltransferase activity and highly variable tandem-repeat zinc finger domains. *PRDM9* plays a key role in determining sequence-specific recombination hot spots in the humans genome (Thomas et al. 2009) and other non-human primates. *PRDM9* carries human-specific changes in amino acids that determine its binding site specificity in humans when compared with chimpanzee. This suggests

that *PRDM9* binds and regulate different genomic regions in humans and chimpanzees. In addition, recent evidence suggest that *PRDM9* sequences from Neandertals and Denisovan were closely related to the ones in present day humans (Schwartz et al. 2014). Zinc finger domains of *PRDM9* display ubiquitous signatures of positive selection on the amino acids responsible for interaction with DNA (Schwartz et al. 2014).

Despite the understanding of the molecular bases that are driving the evolution of human phenotypical differences between populations and with other species is still inceptive, it is clear that genetic variation affecting transcriptional machinery may have a profound effect on fine-tuning the expression of genes involved in particular traits we observe in humans and non-human primate species. Therefore, studying and identifying genetic changes that might be involved in altering gene expression by either introducing variation in non-coding DNA regions involved in regulation of the transcription of nearby genes, or in positions that code for DNA or protein-protein interaction domains, becomes essential for understanding the human regulome, and in a broader perspective, human evolution.

Considering that elucidating how regulatory mechanisms have evolved in humans from a genomic perspective still represents an open challenge with a vast repertoire of possibilities to be explored and identified, we defined and implemented different strategies aiming to characterize and meaningfully contribute to the understanding of the evolution of human GRFs. These strategies involved establishing an inventory of all putative GRF genes of the human genome; the detection of GRFs that exhibit signatures of positive selection in three different human populations; and finally, the identification of human-specific transcription factor binding sites for a particular GRF protein that plays essential roles at cellular, hormonal, neurological and mitochondrial level.

The first chapter of this thesis introduces the strategies for building a comprehensive and up to date catalog of GRFs of the human genome. This catalog gathers information from the most representative inventories so far created for humans in the last decade.

We also grouped all GRFs that have DNA-binding properties into 41 different classes according to Wingender et al, (2015). This chapter also includes a literature review of the biological evidence that connects particular GRFs with the evolution of humans from three different evolutionary windows. It also discusses the likely consequences that evolutionary changes may have introduced in humans from a phenotypic and medical perspective. Consequently, we first explored the evolutionary changes that occurred during the evolution of humans and non-human primates, secondly, the genetic variation after AMHs split from archaic humans, and finally the genetic differences in GRFs within modern humans. Part of the results of this work was published as review paper in the journal Current Opinion in Genetics and Development. In addition, we also used the GRF catalog in a collaborative research project that was recently published in the Journal Frontiers in Genetics. Full references:

> ***Perdomo-Sabogal A**, Kanton S, Walter MBC, Nowick K. 2014. The role of gene regulatory factors in the evolutionary history of humans. Curr. Opin. Genet. Dev. 29:60.*
>
> *Berto S, **Perdomo-Sabogal A**, Gerighausen D, Qin J, Nowick K. 2016. A consensus network of gene regulatory factors in the human frontal lobe. Front. Genet. [Internet] 7. Available from: http://www.frontiersin.org/bioinformatics_and_computational_biology/10.3389/fgene.2016.00031/abstract*

In the second chapter of this thesis we present the results obtained from exploring genome wide data for detecting GRF genes that are candidates for positive selection occurring in three human populations. Using the most recent catalog for GRFs (Chapter one), data obtained from genome-wide scans for detecting positive selection and the information from the 1000 genomes project, we extensively explored which GRF genes and classes are located in genomic regions that are candidates for positive selection in three particular AMH populations. As results, we present a set of GRF candidate genes for positive selection, and introduce six of the larger classes of GRFs that show enrichments for genes exhibiting signatures of positive selection. We also describe three regions harboring multiple *Krüppel* associated box domain zinc finger genes

(KRAB–ZNF) and that present extensive signatures of selection that are larger than 100 kilo base pairs (kb). We finally present some examples of how single nucleotide variants occurring in GRF genes may introduce regulatory diversity in humans, thus possibly leading to the evolution of particular traits such as obesity, blood coagulation, reproduction, and insulin/glucose metabolic pathways in humans. The results of this research are included in a manuscript under preparation.

The third chapter of this thesis presents the results obtained from integrating five different approaches to identify human-specific binding sites for the GRF GA-binding protein alpha subunit (GABPa). From an experimental perspective, three strategies were implemented: chromatin immunoprecipitation followed by massively parallel DNA next generation sequencing (ChIP-Seq), RNA interference (siRNA) and reporter gene assays. From a computational perspective, two main approaches were used: comparative genomics by using multiple sequence alignments and gene ontology analysis. As results, around 6000 GABPa binds sites were identified. Among these, it was possible to pinpoint regulatory regions that are human-specific. In addition, we also detected that GABPa regulates the expression of several KRAB-ZNF genes, and human-specific genes. We also linked *GABPa* to the regulation of human and primate-specific genes that have been associated with human cognitive disorders and neuromuscular functions, mitochondria biosynthesis, and embryo development. The results of this work were recently published in the journal Molecular Biology and Evolution. Full reference:

> ***Perdomo-Sabogal A****, Nowick K, Piccini I, Sudbrak R, Lehrach H, Yaspo M-L, Warnatz H-J, Querfurth R. 2016. Human lineage-specific transcriptional regulation through GA binding protein transcription factor alpha (GABPa). Mol Biol Evol.33:1231–1244*

Special considerations: The chapter three include experimental data that was collaboratively generated with research partners.

Chapter 1

Gene Regulatory Factors, key genes in the evolutionary history of modern humans

2.1. Introduction

At transcriptional level, gene regulation is mediated via GRFs that directly or indirectly interact with DNA and other co-factors to up or down regulate the transcription of other genes (Ryan et al. 1999; Briers et al. 2009; Iyengar and Farnham 2011; Wingender et al. 2013; Perdomo-Sabogal et al. 2014). The GRF proteins are typically characterized by the presence of one or more DNA-binding domains or domains that moderate the interaction with DNA-binding transcription factors (Hughes 2011; Iyengar and Farnham 2011). Genetic changes altering the way proteins regulate the expression of other genes or the configuration of enhancers and promoters, are thought to mainly contribute in the evolution and heterogeneity of phenotypes (Wray 2007; Wittkopp and Kalay 2012). Despite a representative number of studies have shown that sequence changes in cis-regulatory elements (CREs) are more frequent and significantly contribute in driving phenotypical variation at species level (Wray 2007; Wittkopp and Kalay 2012), a growing group of new evidence suggests that changes in the DNA sequence of regions coding for binding domains of GRFs, may also play a major role on shaping phenotypic diversity and evolution at population level (Nowick et al. 2011; Jacobs et al. 2014; Cheatle Jarvela and Hinman 2015; Zhang et al. 2015; Barrera et al. 2016).

In humans, GRFs have been cataloged in several different ways, either using their biological function, tissue-specific expression, or their role as DNA-binding or co-factors (Brivanlou and Darnell 2002; Vaquerizas et al. 2009; Weirauch and Hughes 2011; Wingender et al. 2013). Several other ways have also been considered, either based in the way they interact for regulating genes expression, for instance, by their

ability to interact with co-factors and induce the reorganization of the nucleosome structure and chromatin remodeling (Figure 4) (Messina et al. 2004; Voss and Hager 2014). As result, several scattered inventories of the human GRF genes have been produced in the last decade (Messina et al. 2004; Vaquerizas et al. 2009; Ravasi et al. 2010; Nowick et al. 2011; Tripathi et al. 2013; Wingender et al. 2013; Edgar Wingender et al. 2015). Despite the different efforts to identify and build the most complete set of human GRFs, mainly implemented for DNA-binding transcription factors (TFs) in the above-mentioned works, these independent studies have resulted in fragmented or incomplete information. In some cases, especially in those pioneering works, it is difficult to currently track back the information for such sets of genes.

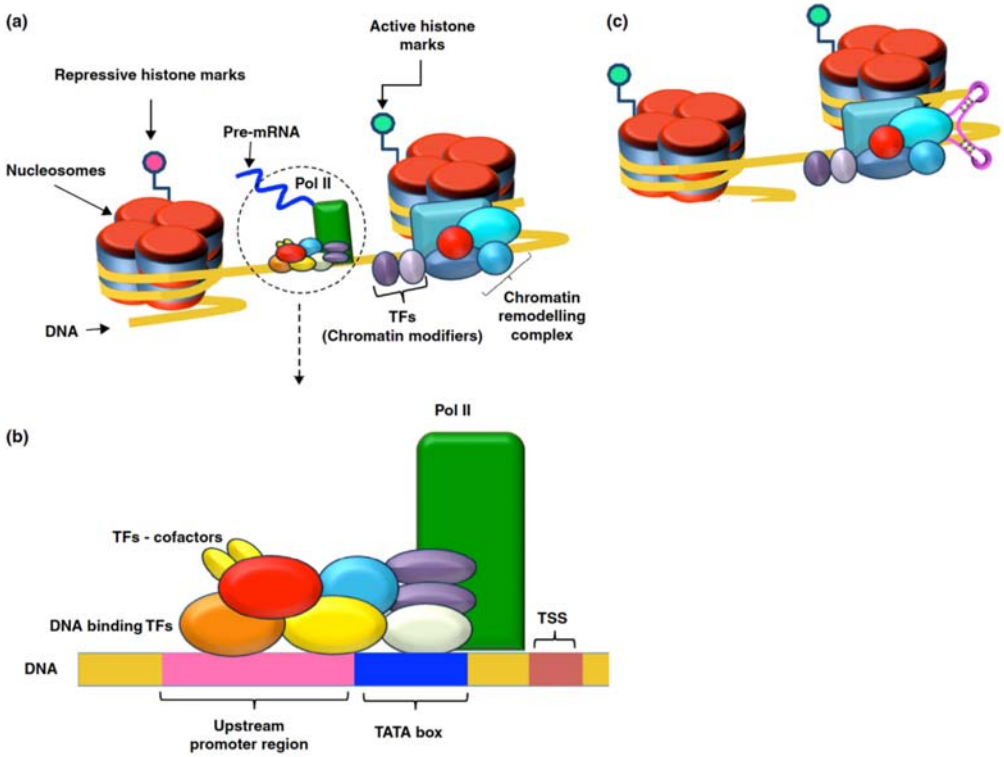


Figure 4. Illustrative representation of chromatin structure and chromatin-mediated gene regulation. (a) Dynamical interaction between GRFs to modify the chromatin architecture and enable the access of the gene regulatory machinery to the DNA, thus regulating the initiation of transcription. (b) GRFs bind the promoter regions, recruit other co-factors and fine-tune the pre-initiation complex to regulate gene expression. The pre-initiation complex helps to position the RNA polymerase II. (c) GRFs also interact with other molecules, for instance, ncRNAs, to down regulate gene expression. Modified from (Perdomo-Sabogal et al. 2014).

Taking into account these functional roles and the scattered information from the most representative seminal works in the area of the human GRFs (Messina et al. 2004; Vaquerizas et al. 2009; Ravasi et al. 2010; Nowick et al. 2011; Corsinotti et al. 2013; Tripathi et al. 2013; Wingender et al. 2013; Edgar Wingender et al. 2015), we built a comprehensive catalog of human GRF genes here. As resource for studying the evolution of GRFs in humans, this catalog enabled us to gather information about the likely role of GRFs in the evolutionary history of humans, archaic humans and non-human primates (Perdomo-Sabogal et al. 2014). In addition, it was a useful resource for identifying GRF genes that might be playing important regulatory roles in the human frontal lobe, a work we recently published in Berto et al, (2016). Finally, it also allowed us to extensively look for candidate genomic regions exhibiting signatures of positive selection in genomic regions where GRF genes are located for three modern human populations. (CEU, CHB, YRI) (Chapter II).

2.2. Results

2.2.1. An updated comprehensive catalog of GRFs for studying regulatory evolution in human

The catalog presented here gathers information obtained from seven different studies performed for GRF in human, and that are widely acknowledged as seminal works in the area. Based on the strategies we implemented, our catalog of GRF includes the majority of genes reported in these studies (Figure 5). Our catalog includes 3037 (90%) of the GRFs that were recently reported in "TFcheckpoint", a TFs database for human (Tripathi et al., 2013). Twenty-six out of the remnant 211 genes listed in TFcheckpoint matched the criterion "regulation of transcription"; however, none of them was supported by the literature. Consequently, we decided to exclude them from our final GRF gene list. Genomic coordinates and gene identifiers of all GRFs in this list correspond with the information available for the human reference genome version GRCh37/hg19. In total, we cataloged 3362 GRFs (Supplementary Table S1, supplementary data file). Based on the strategies implemented here for building this GRF catalog, it additionally includes 287 GRF genes .

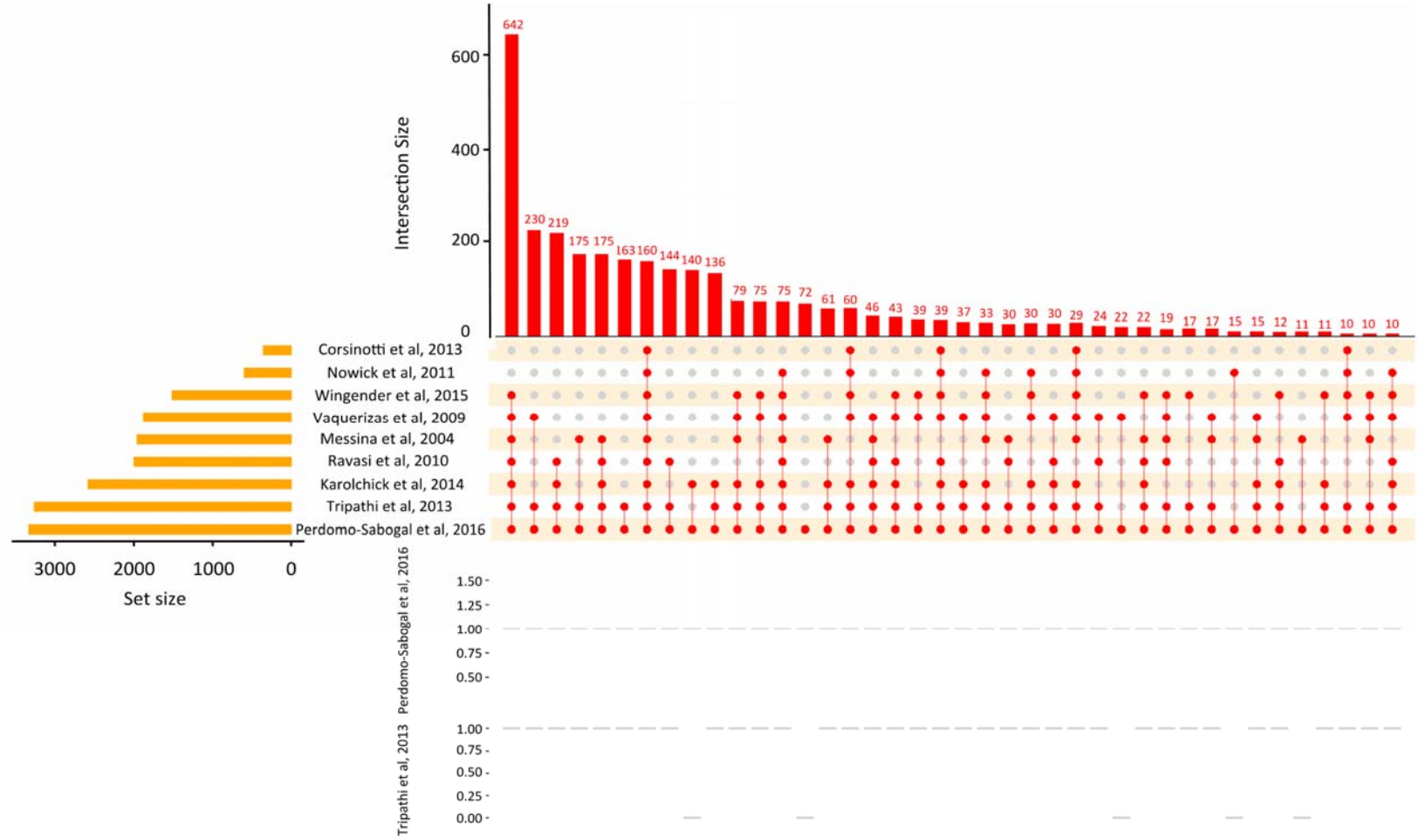


Figure 5. Catalog of GRFs genes present in the human genome. Overlap between our GRF list and the different seminal works that have been conducted for humans during the last two decades

A detailed and curated classification of DNA-binding GRFs into their functional classes exists, which enabled us to functionally classify 1521 GRF genes (~46%) into 40 out of 42 transcription factor classes reported (Figure 6). Two GRF classes, C6 zinc cluster and E2-related, do not have any gene reported in our list. Out of those genes with available information about their class, zinc finger genes are by far, the most abundant type of GRF genes (807 members). C2H2 is the most abundant classes of GRFs with 695 members, of which 415 are KRAB-ZNF, followed by Homebox Domain and basic Helix-Loop-Helix (bHLH) genes (229 members and 109).

For our own research interests and downstream analyses, we decided to split the TF class C2H2 into two separated classes. The non KRAB-ZNF domain genes were kept in the C2H2 class, and we generated a new category containing all C2H2 gene that have a KRAB domain. The rationale behind this decision was based on several reasons, but mainly because there are some indications that suggest that KRAB-ZNF genes might be undergoing different evolutionary processes than other C2H2 genes. For instance, KRAB-ZNF genes represents more than 60% of the C2H2 genes (Figure 6), it includes fast-evolving genes, and copy number variations of these genes have resulted in genomic innovations in humans (Nowick et al. 2011; Nowick et al. 2013). In addition, some members of this group have also been recently connected with important evolutionary mechanisms to down-regulate retrotransposable elements in humans (Jacobs et al. 2014; Lukic et al. 2014; Najafabadi et al. 2015) and human-specific changes might be of interest for understanding recent human evolution.

2.2.2. Evolutionary changes in GRFs after humans split from chimpanzees

Genomic variation is a primary source of phenotypic diversity within and between species. At species level, the acquirement of new genomic elements can have a strong impact on the way particular traits evolve. New genes play essential roles in organismal evolution, as it is observable by the extraordinary diversity they represent in numbers and types among species (Chen et al. 2013). Among the 3362 genes included in our catalog, we identified 15 GRF genes that are uniquely present in humans (*BOLA2,*

BOLA2B, TCEB3C, BAGE2, ZNF138, SCXB, ZNF705B, FOXD4, ZNF658B, FOXD4L4, FOXD4L5, ZNF322P1, SSX1, SSX4B, DMRTC1) (Perdomo-Sabogal et al. 2014), meaning they have no ortholog in other species (Zhang et al. 2010). Despite of being in the genome of one or just few species, new genes can provide new molecular and cellular functions with indispensable roles at developmental, reproductive, neurological and behavioral level (Chen et al. 2013). Results from mRNA expression profiles for different human tissues (Wilhelm et al. 2014) indicated that seven out of the 15 human-specific GRFs are highly expressed in cerebral cortex, testis, ovary, lymphatic and endocrinal system (Figure 7). Therefore, it will be interesting to discern how these 15 human-specific GRF have been integrated into the pathways regulating functional programs in humans or if they are involved in the development of human-specific morphological characteristics.

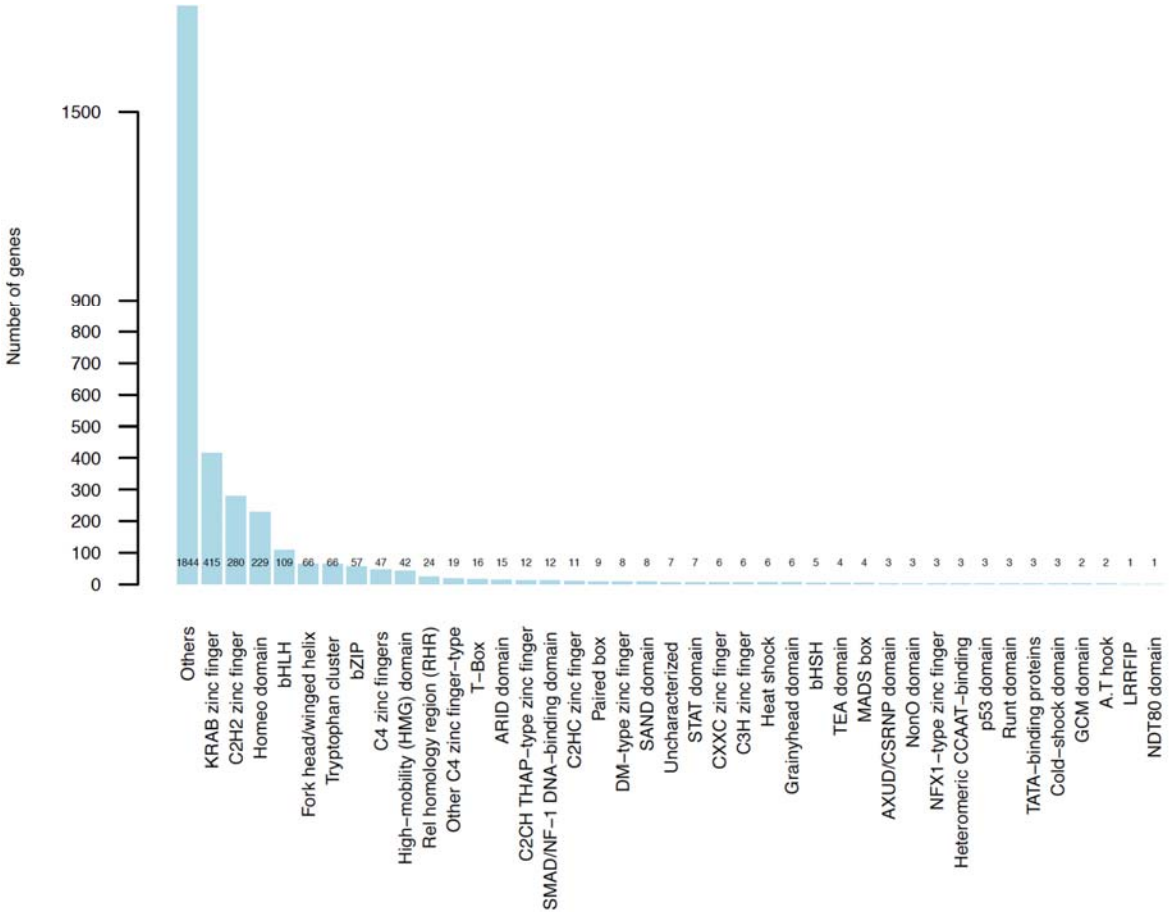


Figure 6. Classification of DNA-binding GRF genes within our catalog according to Wingender et al, 2015.

Genetic disease association studies are frequently used for understanding how human-specific genes or genes showing copy number variants in close related primate species could result in phenotypical traits, physiological constraints and medical consequences. We used the gene-disease association database "DisGeNET" (Piñero et al. 2015) to explore if some of these human-specific genes have been associated with altered phenotypes in humans. Six of the 15 human-specific genes have been connected with diseases such as aggressive cell sarcomas and nervous system pathologies (*SSX1* and *SSX4B*) (Crew et al. 1994; Yawata et al. 2011; Piñero et al. 2015), melanomas (*BAGE2*) (Xie et al. 2002), and substance-related disorders such as tobacco use and nicotine dependence (*ZNF138)* (Rose et al. 2010). In addition, putative roles of the TCEB3C indicate it may act as tumor suppressor protein in small intestinal neuroendocrine tumors (Edfeldt et al. 2014).

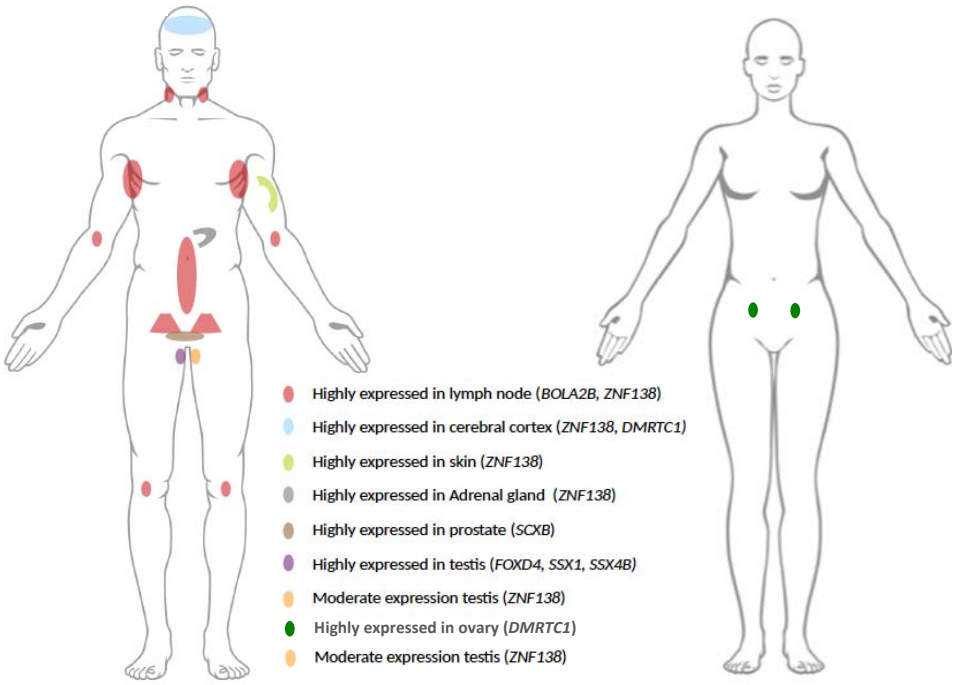


Figure 7. Human-specific genes are over-expressed in lymph node, cerebral cortex, testis, skin, adrenal gland, and prostate. Expression levels of seven out of 15 human-specific GRF genes are based on the median RNA expression-profiles(Wilhelm et al. 2014). Except tissues that are sex-specific, all genes represented here were found to be highly expressed in both males and females (Wilhelm et al. 2014). Figure adapted from ProteomicsDB. (Wilhelm et al. 2014).

By comparing the modern human genome versus the genomes of archaic hominin species, it was also possible to pinpoint some GRF that exhibit higher copy number variants in Neandertal and Denisovan. Some of these GRFs have strong phenotypical effects if mutated or over expressed in AMHs (Table 1). For instance, Histone Cluster 1 gene (*HIST1H2BN*) and Tripartite Motif Containing 26 (*TRIM26*) have been associated with circulatory and cardio vascular diseases such as Behcet syndrome (Piñero et al. 2015), a common syndrome in Middle East and Asian AMHs populations (Durrani and Papaliodis 2008). Another example is Double Homeobox 4 (*DUX4*), a GRF found overexpressed at mRNA and protein levels (Dixit et al. 2007) in facioscapulohumeral muscular dystrophy 1 myoblasts. This altered phenotype results in progressive skeletal muscle weakness.

Table 1. Gene regulatory factors with copy number variants in archaic humans. Genetic association studies revealed some of the likely medical consequences these genetic variations could have had in Neandertal and Denisovan. (¥) Genes with higher copy number variants in Denisovan. (*) Genes showing higher copy number variation in Neandertal (Perdomo-Sabogal et al. 2014)

Chromosome	Strand	Gene Start (bp)	Gene End (bp)	HGNC symbol	Ensembl Gene ID	Genetic Association (Disease class)	Genetic Association (complex disease and disorders)
6	-	26043455	26043885	*HIST1H2BB*	ENSG00000196226		
6	+	27806323	27823487	*HIST1H2BN*	ENSG00000233822	Cardiovascular, Immune	Cardiovascular diseases, lupus erythematosus systemic
7	-	102113565	102119354	*POLR2J**	ENSG00000005075	Cancer	Urinary bladder neoplasms
9	-	116237	118417	*FOXD4*	ENSG00000170122		
9	-	70426623	70429731	*FOXD4L4**	ENSG00000184659		
9	-	70175707	70178815	*FOXD4L5**	ENSG00000204779		
18	-	102908	112287	*DUX4*	ENSG00000258389	Developmental	Muscular dystrophy
18	-	44554573	44556449	*TCEB3C*	ENSG00000183791		
6	-	30152231	30181271	*TRIM26*¥	ENSG00000234127	Immune	Behcet syndrome, lupus erythematosus, multiple sclerosis, bipolar disorder

2.2.3. Evolution of GRFs in AMH population

Genomic comparisons between human individuals from different human populations provide a great opportunity to detect evolutionary events associated with particular human traits. Since AMHs migrated out of Africa the first time, AMHs have undergone

substantial morphological changes. As consequence, AMHs display particular population specific features such as lactose persistence, skin color, hair thickness, height, among others.

Recent genome-wide scans for signatures of positive selection have identified a large set of candidate genomic regions (Sabeti et al. 2007; Pickrell et al. 2009; Chen et al. 2010; Metspalu et al. 2011; Grossman et al. 2013). Overall, the overlap between the genomic regions under selection reported in these studies is not high. Nonetheless, by using the genomic coordinates of these regions and the ones from our GRF catalog, it was possible to identify several GRF genes that are likely candidates for positive selection in AMHs. In ten cases, the same GRF gene was reported in more than two studies (Table 2). It is also interesting that some of these GRF are located in regions showing population-specific signatures of positive selection (Supplementary Table 2), which may also indicate population-specific regulatory mechanisms. For instance, in the WW domain containing oxidoreductase (WWOX) gene, a TF gene that has been found in a region that shows signatures of a recent selective sweep (Table 2), in Utah residents with northern and western European ancestry, carry a SNP that has been associated with changes in lipid metabolism and cardiovascular disease risk in humans. Despite further research is required, it is possible that WWOX might be involved in changes of lipid metabolism at population specific level.

Table 2. GRFs located on genomic regions that have been identified as candidates for positive selection in AMHs (Perdomo-Sabogal et al. 2014).

Gene symbol	Number of genes	Source
WWOX	1	(Sabeti et al. 2007; Pickrell et al. 2009; Grossman et al. 2013)
MYEF2, FBN1	2	(Sabeti et al. 2007; Pickrell et al. 2009)
ZMYM6	1	(Sabeti et al. 2007; Grossman et al. 2013)
PPARA, KCNH5	2	(Pickrell et al. 2009; Metspalu et al. 2011)
HIF1A, SNAPC1, DPF1	3	(Pickrell et al. 2009; Grossman et al. 2013)
KCNH7	1	(Metspalu et al. 2011; Grossman et al. 2013)
CTNND2, BMI1, AFF2, BBX, NFE2L2	5	(Sabeti et al. 2007)
RGS9, ERBB4, ATF6, PHF19, DUSP12, RFX3, CIITA, NCOA7, APC, TRIM14, SETBP1, POLR2K, FOXE1, HSF2, YTHDC1, HEY2	16	(Pickrell et al. 2009)
CLOCK, MSTN, LIN28B, ISX	4	(Metspalu et al. 2011)
ANKRD45, RRN3, SFPQ, SIN3A, SLC30A9, CCDC71, RNF135, NCOA1, PCGF1, HIRA, MCM6, ASXL2, FOXP1, RHOA, TERF2IP, TAX1BP3, HIPK1, KCNIP4, RFX5, ADNP2, ZBTB41, PAPOLA, POGZ, FMNL2, ACTR5, PAWR, LHX8, USF1, EBF1, LBX2, CHD2, ARIH2, PHTF1	33	(Grossman et al. 2013)

2.3. Discussion

Regulation of gene expression involves a wide and complex repertoire of mechanisms that are, to a large extend, fine-tuned by GRF proteins. As *trans*-acting molecules, GRFs directly or indirectly interact with DNA to up or down regulate gene expression. As group, GRFs have evoked the interest since they might be playing essential roles in determining species-specific phenotypes. It is plausible that the appearance of new genetic changes occurring on GRF genes are a prime source for the molecular diversification and divergence in hominins and other non-human primates, and the effects of natural selection on AMHs.

By using several pre-existing inventories we built an updated list of GRF genes for human. In contrast to previous works, this catalog includes genes involved in different regulatory mechanisms such as DNA-binding, co-factors, histone and chromatin modifiers, among others. The overlapping strategy followed by batch coordinates conversion and manual inspection of different databases, brought us to estimate in about 3362 GRF genes. Compared with the first inventory of TF genes realized for humans by Messina (2004), we additionally cataloged 45 zinc finger genes and 30 Homeobox genes. The number of bHLH genes reported in our catalog is smaller than the one reported in Messina's study. Nonetheless, our catalog includes the number of bHLH genes reported in the latest classification of human DNA-binding TFs (109 members) (Wingender et al. 2015). Compared with the most recent inventory of human GRFs "TFcheckpoint" (Tripathi et al. 2013), our catalog additionally includes 287 genes. This catalog becomes the most complete inventory of human GRF. As source of information, this catalog is of great utility for exploring the roles of GRF in human evolution. For instance, for identifying GRF genes that might be involved in important regulatory roles at tissue-specific level, gathering information about human-specific GRFs, comparative genomics with other hominin and non-human primate species, among other uses.

2.3.1. Newly evolved GRF in humans

Genomic innovations such as the appearance of new genes, changes in copy number variants or new regulatory elements such as transcription factor binding sites can strongly affect the way phenotypes evolve. Out of the 3362, 15 GRFs correspond to human-specific genomic innovations, while other 17 are Hominidae-specific. Apart from the number of human and primate-specific genes that have been identified so far (Zhang et al. 2010), over-expression of younger genes in testis seems to be a constant (Chen et al. 2013). Four human-specific GRF genes were found highly expressed in human reproductive organs: testis (*FOXD4*, *SSX1, SSX4B*) and ovary (*DMRTC1*) (Wilhelm et al. 2014).

The appearance of human-specific genes as result of copy number expansions in humans have also been connected with the modulation of brain functions (Fortna et al. 2004; Sudmant et al. 2010; Geschwind and Konopka 2012). Analyses of expression-profiles of the human brain have shown enrichment of human-specific genes in the prefrontal cortex, a human trait that displays distinctive structural characteristics and cognitive capabilities when comparing human with other primates (Zhang et al. 2011). Two newly evolved GRFs in human are overexpressed in brain (*ZNF138* and *DMRTC1*) (Wilhelm et al. 2014). This suggests, although not conclusively, that some newly evolved GRF genes might be playing human-specific roles in tissue-specific regulatory networks at reproductive and neurological level in humans.

2.3.2. Human-specific GRFs and disease

Disease association studies evidenced that six of the human-specific GRF are connected with altered phenotypes in humans. For instance, chromosomal translocations that result in gene fusion between the Synaptotagmin gene (*SYT*) and members of the Synovial Sarcoma X Breakpoint, genes *SSX1* and *SSX4B,* have been linked with aggressive cell sarcomas and nervous system pathologies. The gene B Melanoma Antigen (*BAGE2*) is another example of human-specific genes involved in the

occurrence of tumor of melanin-forming cells. This gene has been found exclusively expressed in 22% of melanomas (Xie et al. 2002).

GRF that exhibited copy number variants in Neandertal and Denisovan can have profound effects in human phenotypes if mutated or over expressed. For instance, some of them are associated with circulatory and cardio vascular diseases (*HIST1H2BN* and *TRIM26*) such as Behcet syndrome (Piñero et al. 2015), a pathology characterized by the inflammation of the blood vessels. Another example is *DUX4*, a GRF highly expressed in patients with muscular dystrophy that results in the continuous skeletal muscle weakening (Dixit et al. 2007). This suggests that changes in GRFs may have also resulted in detrimental phenotypes in archaic hominins. It also highlights some consequences these evolutionary changes could have had on individuals carrying copy number variants for these genes in hominin species (Perdomo-Sabogal et al. 2014).

Identifying human-specific genes, in general, allows the exploration of how the addition of new genes may lead to phenotypic diversity and evolution. Newly evolved GRFs contributing with human-specific functions may have a broader impact on pathways associated with important speciation traits in humans, for instance, with evolved reproductive, brain and immunological roles. Therefore, the identification of newly evolved genes that perform GRF functions in humans opens the possibility to experimentally explore their roles or effects in human-specific regulatory pathways. For instance, the implementation of ChIP-Seq, gene silencing, gene editing, among other experimental approaches on this set of genes, would be of great utility for understanding their roles at different biological levels in humans.

2.3.3. A first glance of evolutionary changes in GRF within AMHs populations

The occurrence of a new mutation can introduce a new phenotypic attribute that leaves population specific signatures in the genome after fixation (Wollstein and Stephan 2015), for instance, selective sweeps. Main characteristic of a selective sweep is the considerable reduction in the genetic variability in the neighborhood of a particular

new mutation as consequence of strong natural selection. Selection acting on particular gene variant also has an effect on the neighboring alleles and on the recombination rates of a given region, thus resulting in an overrepresentation of a particular haplotype (Messer and Petrov 2013).

It is relevant to highlight that the aforementioned scans for positive selection in human population genomic data have implemented methods mainly designed for detecting hard selective sweeps. In most cases, the authors have focused on those regions exhibiting extreme values (strong conservative measures). Therefore, it is likely that other types of selective events, for instance, soft sweeps have been excluded from the final reported regions. Thus, three main general conclusions can be depicted as result of these studies. First, the different scans for positively selected genes have resulted in a non-overlapping list of candidates. Second, a combinatorial strategy by implementing different statistical approaches could help to reduce the number of false positives by comparing different values for particular candidate genes. It could also help to reduce the effects that the demographic patterns such as bottlenecks, expansions, genetic flow, among others, can generate on diversity, thus disentangling when a signature of selection should be considered more reliable. Third, overall, regulatory elements seem to substantially contribute to both adaptive substitutions and deleterious polymorphisms that might have had key implications for human evolution (Enard et al., 2014).

Due to the strong conservative strategies that have been implemented for analyzing the results of such scans for selection, we suggest they might be ignoring an entire class of influential adaptive signatures that still remain uncovered in the genomes of AMHs populations. Consequently, we think the implementation of these methods for detection of selection, followed by less conservative strategies for detecting positively selected regions, and the targeting of specific groups of genes, for instance, GRFs, is important to detect some still not described signatures of selection.

2.4. Conclusion

Despite in the last two decades a majority of studies have suggested that changes in CREs elements is the prime source for species differentiation and phenotypical diversity, recent evidence suggest that appearance of new GRF genes and changes in the proteins of these genes are additional sources for driving phenotypical diversity. We presented here comprehensive catalog of GRF for the human genome, with some examples of how such catalog of genes may contribute to studying and understanding the roles of GRF proteins in the evolutionary history of human-specific traits, but also we gathered information of utility for designing new experiments to deeper explore and validate the potential role of these genes in defining human-specific phenotypes.

2.5. Materials and Methods

2.5.1. Building the GRF gene catalog

We brought together seven different GRF gene lists previously published for human. These gene sets included TFs transcriptome and sequence evolution studies and TFs inventories (Messina et al. 2004; Vaquerizas et al. 2009; Ravasi et al. 2010; Nowick et al. 2011; Corsinotti et al. 2013; Tripathi et al. 2013) and functional classifications (Wingender et al. 2013; Wingender et al. 2015) (Figure 8). The whole process involved overlapping genes based on their stable identifiers (IDs), batch coordinate conversions (liftOver) (Karolchik et al. 2014), manually inspection using different databases and gene ontology for associating gene product attributes. Using customized Perl scripts, and information from ensemble (Flicek et al. 2014), UCSC genome browser (Karolchik et al. 2014), UniProtKB (The UniProt Consortium 2015), and HUGO (Eyre et al. 2006) databases, among others, we overlapped the gene lists that provided the frequently recommended stable identifiers (IDs). For instance, we used identifiers such as official gene symbols, gene and transcripts ensemble, UCSC, protein, and RefSeq IDs. These identifiers were used as keys.

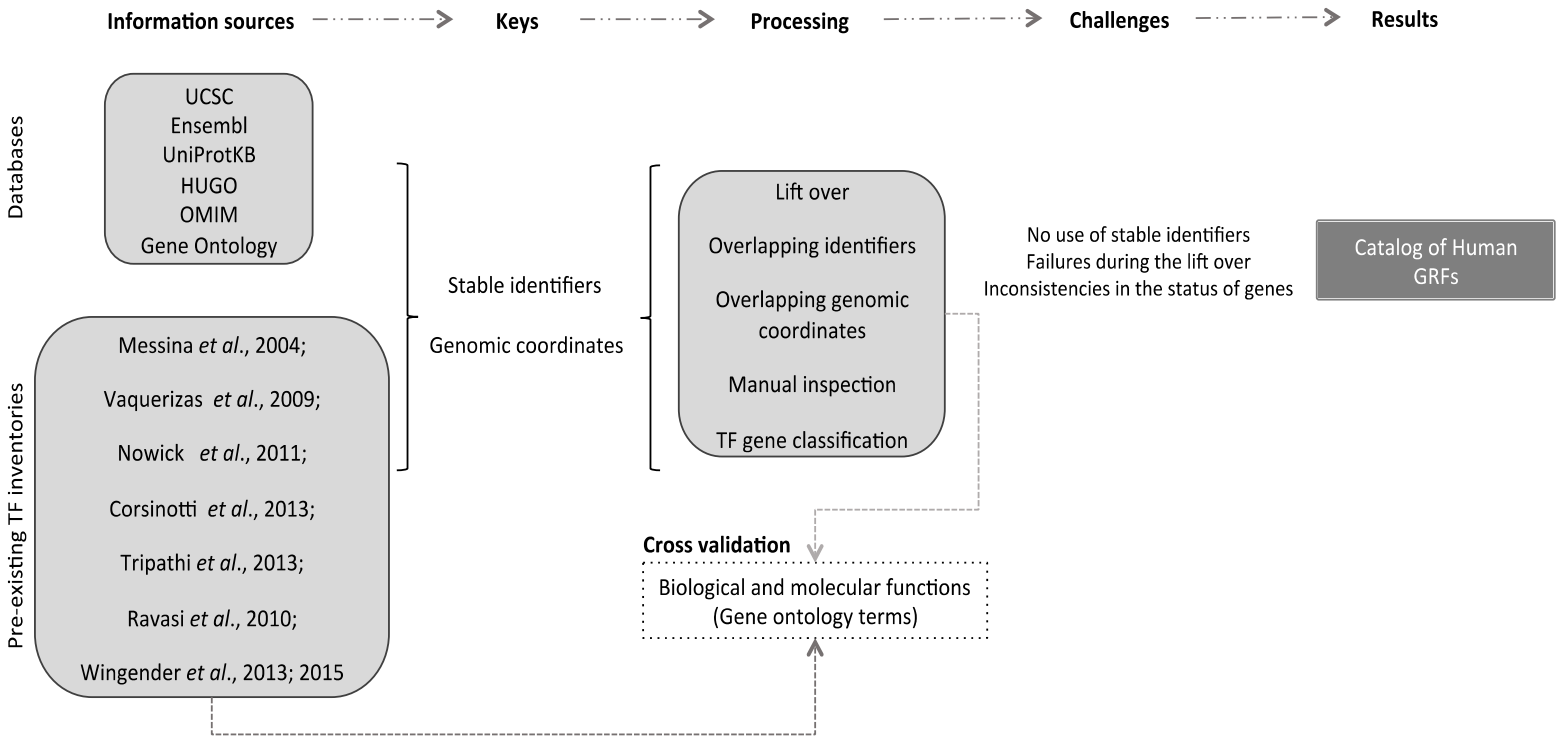


Figure 8. This workflow represents the step-by-step procedure for building the GRF catalog presented here. The selected scientific publications sourced here were chosen based on their efforts and contributions to identify and categorize human GRFs.

In those cases where it was not possible to overlap particular identifiers, mainly explained by changes in the names due to updates between versions of the reference genome, we lifted over the genomic coordinates between assemblies. Subsequently, and using customized perl scripts and bedtools (Quinlan and Hall 2010), we performed a new overlap between sets of coordinates. In those cases where none of the two previous strategies was successful, we used the gene symbols and genomic coordinates to look for additional information in the aforementioned databases. Despite the efforts, it was not possible to track back all the information reported in some of these studies. This is mainly explained by the use of non-official stable IDs and failures when performing the liftOver because of retired genes from previous genome versions in the version GRCh37/hg19. In addition, there were also discrepancies between the current and previous status of genes, for instance, loci that were previously considered as genes, but now cataloged as pseudogenes, or ncRNAs.

Since some of the GRFs reported in the aforementioned studies were present just in one of these inventories, we introduced an additional criterion to consider a gene within our catalog. Despite it is known that the molecular mechanisms that regulate transcription are not yet fully understood for many proteins coded by GRF genes, we required that singletons have to be associated with a list of gene ontology (GO) terms we particularly customized. To do so, we built a list of GO terms to identify genes that are likely to perform GRF activities. We selected molecular and biological GO terms such as regulation of transcription, DNA-depending transcription, RNA polymerase II transcription cofactor and co-repressor activity, chromatin binding, remodeling, among other 218 terms, (Supplementary Table S3, supplementary data file). Then we sourced the GO information for the whole set of genes for the human reference GRCh37/hg19, in total 27,993 genes and around 80,900 transcripts, from the UCSC genome browser. This list also included information about genomic coordinates, several types of stable identifiers (i.e. Ensembl IDs, official gene symbols, UCSC IDs, RefSeq, among others). These two GO terms lists allowed us to first, overlap identifiers and genomic coordinates between the singletons, the GO terms for the whole set of human genes and the customized set of GO terms we generated, and second, cross-validate all GRF genes versus the GO customized list.

For our own investigative purposes, we identified the smallest and the biggest coordinates of all transcripts of a gene to define its whole genomic region. In those cases where some GRFs presented multiple overlapping transcript coordinates, we used the smaller gene start and the biggest gene end, thus covering the whole genomic region for each particular gene (Figure 9a). In cases where the same gene had one to several non-overlapping transcripts (Figure 9a), we decided to keep them separately. In such cases we added an additional tag at the end of the official gene symbol. This tag consists of a dash (-) followed by a capital letter from (A) to the number of non-overlapping transcripts. We also kept separately those GRF genes that have multiple copies located on the same chromosome, but do not overlap at all (Figure 9b). For instance, we slightly modified the names for copies of the gene that encodes for the BolA-like protein 2 from BOLA2B to BOLA2B-A and BOLA2B-B. Among other particularities of the catalog we built here, there were some cases where GRFs had the same stable gene symbol but the copies were located in different chromosomes, for instance, the gene for protein phosphatase 2 regulatory subunit B (PPP2R3B). This gene is assigned to two genomic locations, one on the chromosome Y and the other on the chromosome X (Figure 9b). In such cases we slightly modified the names for copies of the gene from PPP2R3B to PPP2R3B -A and PPP2R3B-B. In cases where alternative gene names were used to address conjoined genes, as well as known as fusion genes that produce readthrough transcripts, we kept them separately. For instance, the genes ZNF670 and ZNF670-ZNF695 (readtrough) where at least part of one of the coding regions from each gene is present (Figure 9c), were kept separately.

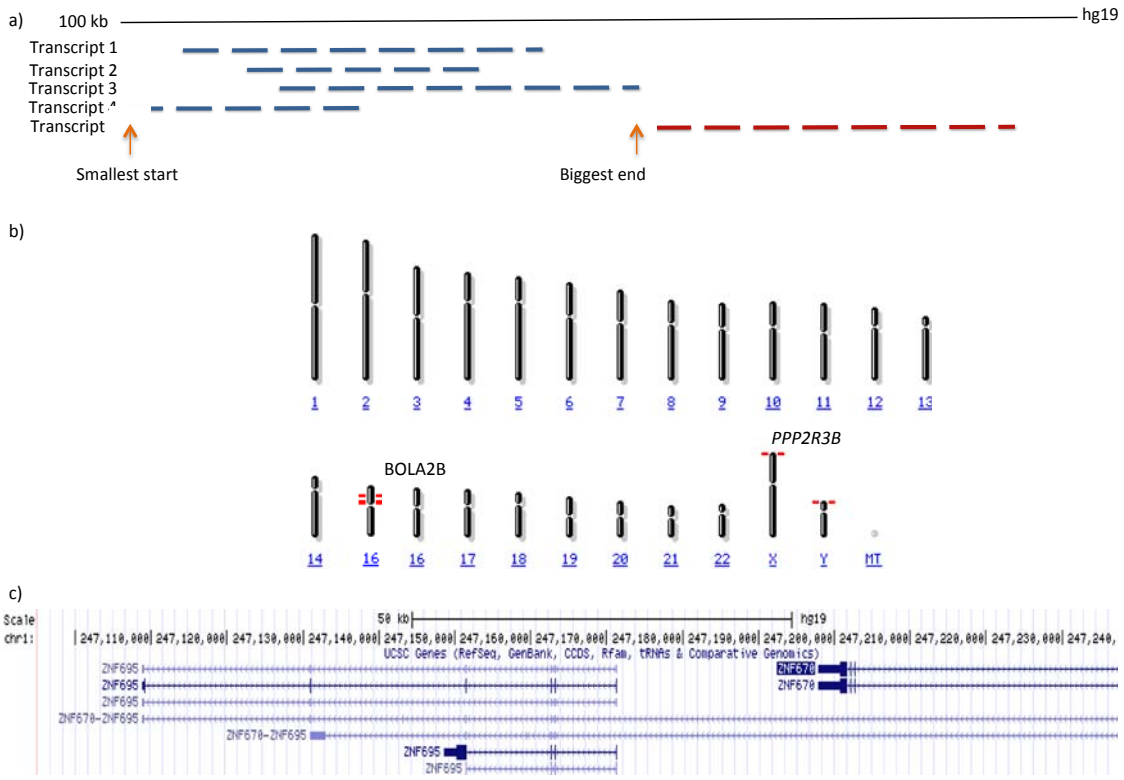


Figure 9. Particularities found for some genomic regions while building the GRF catalog for human. (a) Genes having transcript variants that have been assigned to non-overlapping genomic neighboring regions. (In red) example of non-overlapping transcript variant for the same gene. (b) Genes having copies located on the same chromosome but in different locations (i.e. BOLA2B gene) or genes having two different genomic locations due to copies located on different chromosomes (i.e. *PPP2R3B* gene). (c) Example of conjoined or readtrough GRF genes *ZNF670-ZNF695*.

In addition, and based on the results obtained by Huntley et al, (2006), we decided to keep several putative GRFs within our catalog, even when they are currently considered as pseudogenes, lincRNAs, or in the worst case scenario, they have a retired status. The rationale behind keeping these genes in our catalog lays over the manual curation Huntley and collaborators carried out (Huntley et al. 2006). Additionally, these genes show all the characteristics of protein coding genes, for instance, open reading frame with no stop codons in the coding sequence (Nowick et al. 2011). For such cases, 22 KRAB-ZNF in total, the official gene symbols were tagged as follows: (+N) indicates the gene was reported as new in Huntley et al., (2006) but now is either considered a pseudogene, or has retired status; (+NP) are not considered protein coding genes (i.e. ncRNAs; pseudogenes).

Chapter 2

Positive selection on GRF genes as source for regulatory diversity in human populations

3.1. Introduction

3.1.1. The genetics of AMHs' adaptation

During the last 95 ka, AMHs have widespread out of the sub-Saharan Africa to successfully adapt and colonize different world latitudes (Fu et al. 2013). Multiple factors have influenced the genome diversity in AMHs and may be responsible, to some extent, of the phenotypical variation. This process has involved adaptation to a diverse collection of habitats such as hot and cold, dry and humid, forests, desserts, and different altitudes, among others (Pritchard et al. 2010). It has also involved the implementation of new diets, and the development of immunological resistance to diseases and pathogens characteristic of such environments. These differences in climates, resources, pathogens exposed humans to strong changes that acted as selective forces (Vasseur and Quintana-Murci 2013).

To what extend have these factors influenced genomic adaptation in quantitative traits, and how can such traits rapidly respond to changing selective pressures by shifting the allelic frequencies of a vast repertoire of polymorphic sites? If they have left signals in the genomes of AMHs' populations, can we identify the genes or groups of genes that have been mostly influenced? Our understanding of human genome variation has considerably improved during last decades. The continuing development of large scaled datasets by, for instance, SNP arrays and whole genome sequences of a large number of individuals from multiple AMH populations (The International HapMap Consortium and International HapMap Consortium 2005; 1000 Genomes Project Consortium 2012) have prompted the interest in finding such genomic regions, especially those likely to carry signatures of natural selection.

Given that selection acts at phenotype level, alleles showing evidence of selection are likely to be of functional relevance.

Genomic regions, under the assumption that benign genetic variants that increase fitness would have been conserved (Vasseur and Quintana-Murci 2013), are useful for interpreting how history and the legacy of natural selection have impacted the human genome. They are also important for understanding the evolutionary dynamics that have governed the AMHs adaptation (Biswas and Akey 2006). In addition, the assessment of the intensity and type of selection acting on different human genomic regions facilitates pinpointing of candidate loci that are likely to be related with rare or severe diseases, and genes that are most likely to be involved in susceptibility or resistance to diseases (Vasseur and Quintana-Murci 2013). For instance, diseases exhibit geographical dissimilarities at population level such as metabolic and autoimmune medical conditions (Karlsson et al. 2014).

Changes in the gene regulatory mechanisms and subtle variation in gene expression levels seem to prevail in the way phenotypical traits diversify, in particular, in a short evolutionary time scale (Bornberg-Bauer et al. 2010; Jones et al. 2012). Therefore, it is expected that a substantial fraction of the genetic and phenotypic differences we currently observe within AMH populations, are likely to be a consequence of variation in the regulatory mechanisms. If adaptation is primarily driven by regulatory mechanisms rather than the appearance of new genes, it could also be expected that some advantageous changes occur on key functional parts of the protein such as DNA- and protein-protein binding domains of GRFs involved in controlling gene transcription. For instance, nucleotide variations on SNPs changing the amino acid sequence of DNA-binding domains (DBDs) of GRFs may alter their binding affinity, thus introducing a source of regulatory functional variation in humans (Barrera et al. 2016). Such variation could result in adaptive source of variation at population level, but also in detrimental traits and risk disease. If the advantageous changes undergo positive selection, it is also possible that these selective events have left footprints in regions where these adaptive variations have occurred. Therefore, a detailed exploration aiming to identify GRFs and particular GRF classes exhibiting footprints of selection in the human genome would be of

great contribution to our understanding of the roles of the regulatory diversity on the evolutionary history of AMH populations.

3.1.2. Genome-wide scans methods for detecting positive selection

With the publication of the large-scale study of human genetic variation (1000 Genomes Project Consortium 2012), humans are plausibly the best model for performing genome wide scans for positive selection. Many different methods for detecting positive selection from polymorphic data have been developed and implemented in studying human variation during the past decades (Tajima 1989; Bustamante et al. 2005; Nielsen et al. 2005; Sabeti et al. 2007; Sabeti et al. 2007; Pickrell et al. 2009; Grossman et al. 2013). Most of them make use of the distortions positive selection causes on the patterns of expected genomic neutral variation. Natural selection also acts with different strengths and depending on the signature it produces on a particular region it is possible, to some extent, to predict the mode of selection. For instance, some of the signatures that selection leaves at population level are consistent with a skew in the allele frequency distribution, reduced levels of haplotype diversity, elevated levels of linkage disequilibrium (LD) (Biswas and Akey 2006), increment in the number of rare alleles, shifts in the allele frequencies between populations (Pritchard et al. 2010), excess of intermediate-frequency alleles, or reduced neutral variation (Figure 10).

3.1.3. Approaches for detecting positive selection at species level

If positive selection has taken place, the strength of this event produces a particular shift in the genome diversity, for instance, by altering the nucleotide diversity or generating extended haplotypes. As previously mentioned (Introduction, figure 3), selective sweeps normally cause a reduction in the genetic variation of the region undergoing selection either at metapopulation (group of spatially separated populations) wide or population-specific level. Although, the levels of diversity gradually return to a baseline overtime due to new mutations generating new alleles (Figure 10a), the signatures of positive selection prevails for thousands of generations (several hundred thousand years in humans) (Vitti et al. 2013), and are thus still possible to detect. Different statistical methods have been designed and implemented for scanning and detecting such patterns of selection in a single

population or in multiple populations in humans (Sabeti et al. 2007; Nielsen et al. 2009; Chen et al. 2010). Based on different aspects of the genetic data, these methods can be classified into different groups (Figure 10).

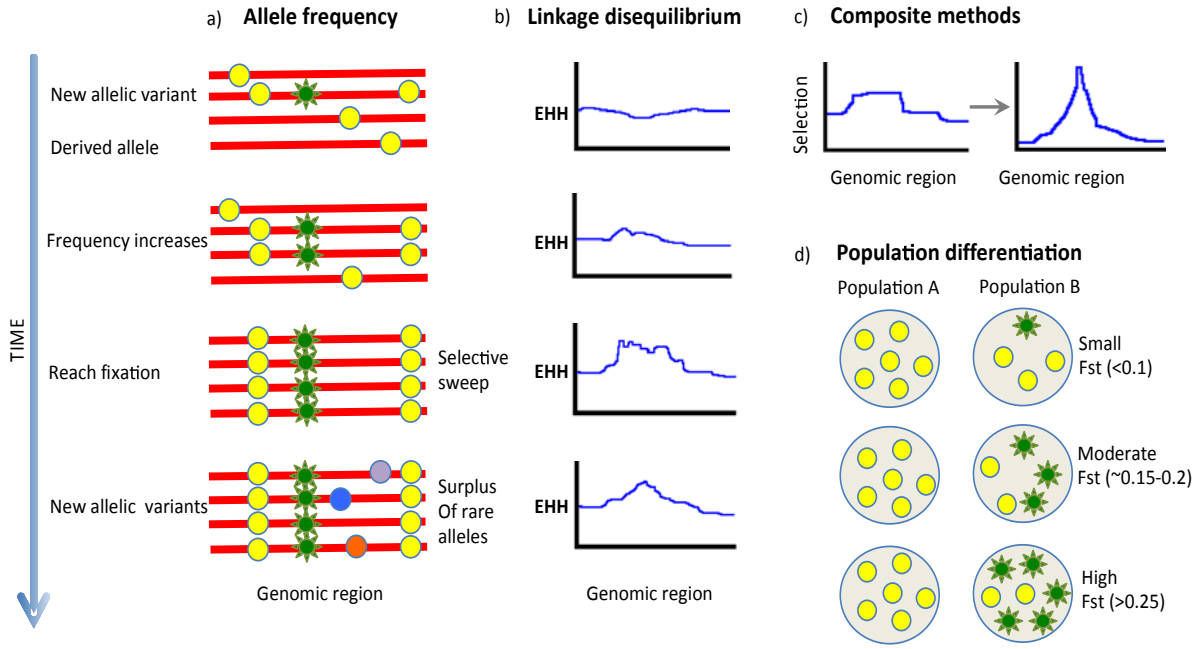


Figure 10. Signatures of selection on the genome used for detecting positive selection. (a) Allele frequency spectrum. A new mutation that improves fitness increases in frequency, and simultaneously brings nearby alleles to high frequency until the region reaches fixation (Selective sweep). This phenomenon is widely known as genetic hitchhiking. After the selective event has elapsed, new mutations create a surplus of rare or low frequency alleles. (b) A selected sweep causes a reduction in the variation of particular regions, thus leaving an Extended Haplotype Homozygosity (EHH). This EHH is conditioned on existing non-random association of alleles due to physical linkage, as well referred as one type of LD, which rises across the genomic region that contains the selected allele (Vitti et al. 2013). Selective sweeps causing EHH look similar to a big city 'skyline'. After a selective sweep has passed, the appearance of new mutations and the restoration of genetic recombination gradually return the diversity to the population. (c) Composite likelihood methods combine (multiplying) a collection of individual scores obtained from one or several tests for detecting selection for all positions (markers) within one region. Thus, the detection methods can be refined by integrating multiple products of the probabilities of all SNPs within one region. (d) Strong variation in the allelic frequencies of particular regions or SNPs between populations might reflect the effect of selection at population-specific level. High levels of population differentiation due to genetic structure increase the fixation index (Wright's F_{st}), between pair populations. Yellow circles correspond to extant allelic variants. Green stars indicate a new allelic variant that increase in frequency. Purple, blue and orange circles indicate new allelic variants that are rare. Figure modified from (Vitti et al. 2013).

Methods based on allele frequency spectrum calculate if the selected allele and it's hitchhiking effect on a particular region has increased in frequency and swept to reach fixation in one population (Tajima 1989; Fay and Wu 2000) (Figure 10a). One widely used method to quantify these selective events is Tajima's D (Tajima 1989). This method quantifies deviation from neutrality by comparing the pair-wise differences between individuals within one population. The Tajima's D test allowed the identification of several genes that have undergone selection in human populations, for instance, the lactose gene (Bersaglieri et al. 2004), Human Leucocyte Antigen (Hughes and Yeager 1998) and the histo-blood group ABO locus (Seltsam et al. 2003). Nonetheless, Tajima's D scores computed on genotypes obtained from SNP data, for instance, next generation sequencing data (NGS), has been proven to lead to high rates of false positives and biased results (Korneliussen et al. 2013).

In addition to site frequency, SNP data also provide information about the LD. As an allele is undergoing selection and its hitchhiking region sweeps through the entire population, it simultaneously causes the appearance of long extended regions characterized for their low variability. These regions are normally referred to as extended haplotypes (Figure 10b). Empirical and simulated data has shown that such regions stay in strong LD (Sabeti et al. 2002; Kim and Nielsen 2004; Stephan et al. 2006; Sabeti et al. 2007; Huff et al. 2010). Therefore, additional information about the LD introduces additional discriminatory power for identifying regions undergoing selection. LD based approaches have also additional power for detecting partial or incomplete sweeps, in which some regions exhibit new mutations that have increased in frequency, but have not yet reached fixation within one population (Pickrell et al. 2009; Messer and Petrov 2013; Vitti et al. 2013). Methods that integrate the information about the haplotypes length between populations, for instance, the widely implemented cross-population EHH (XP-EHH) test, allow to additionally control for the effects of local variation and genetic recombination rates. XP-EHH measures the decrease in the diversity by calculating the probability that any two randomly chosen extended haplotypes around a given locus within the same population are identical by descendent for the entire region (see Methods for XP-EHH model explanation).

Other types of computational approaches for detecting selection include composite methods. These methods have been designed to increase the power for detecting positive selection by combining multiple metrics into a composite score (Figure 10c). Under the assumptions that measuring selection by using one metric can result in a high number of false positives, and that selection affects extended regions, a consecutive genomic region of positive markers may better represent a signature of positive selection. Under this premise, composite multilocus tests of allele frequency differentiation integrate information from the same test across multiple sites to refine the power and reduce the false discovery rate (Nielsen et al. 2005; Vitti et al. 2013). One example of a widely implemented composite method is the Composite Likelihood Ratio test (CLR), initially proposed by Kim and Stephan (2002) and modified by Nielsen et al. (2005). In the new implementation of the CLR test, Nielsen et al. (2005) treated the null hypothesis as not specific to a particular population genetic model, but instead, as derived from the patterns of variation produced in the background from the data itself. This implementation additionally corrects for the SNP ascertainment bias that may be introduced by the nature of the data itself (Nielsen et al. 2005) (see Methods for CLR model explanation). Similarly to XP-EHH, another composite test that incorporates likelihood ratios and information about population differentiation is the XP-CLR test. XP-CLR additionally includes information about recombination rates from the reference population; and similarly identifies genomic regions where shifts in the allele frequency rapidly occurred (as estimated by the extension of the influenced genomic region) to be explained by random drift (Chen et al. 2010) (see Methods for CLR model explanation).

An additional method widely used to measure the effects of selection in a particular population, developed even before the NGS era, is the fixation index (F_{st}) (Weir and Cockerham 1984). Under the assumption that the selective prevalence of a particular allele is affected by the characteristic environmental conditions in which the individuals that carry it live (Vitti et al. 2013), it is expected that selection differentially acts on that locus for different populations (Figure 10d). F_{st} is frequently used to estimate the allelic variance within and between pairs of

populations. High F_{st} scores suggest high genetic differentiation, which may also indicate the effects of directional selection for a particular locus.

By integrating results obtained from the implementation of three different statistical methods for detecting positive selection (Table 3) and considering their statistical attributes (XP-EHH, CLR and XP-CLR) (Pybus et al. 2013), we extensively explored the genomic regions that contain GRFs for signatures of selection in three AMH populations: Utah Residents with Northern and Western Ancestry (CEU), Han Chinese in Bejing (CHB), and Yoruba in Ibadan (YRI). We additionally included F_{st} results as complementary test for measuring if the candidate regions also experience strong genetic differentiation between pairs of populations. Since genetic variability is strongly affected by demographic processes, for instance, bottlenecks or population expansion, and by changes in mutation and recombination rates (Nielsen et al. 2009), it becomes challenging to distinguish between evidence of selection and demography when performing scans for selection genome wide. Considering it is expected that demographic events produce similar patterns in variation genome wide, we decided to use Rank Scores (RS), which are genome based rank "p-values" calculated based on the genome distribution of the raw score for each particular test and population. As a result of implementing this strategy, we expected to reduce the effects demography could have on the analysis and interpretation of the data.

Table 3 Different statistical methods used for identifying signatures of positive selection on GRF genes for three different human populations. cM, centimorgan. CLR and XPCLR tests were performed using slide windows strategy, while XP-EHH and F_{st} were SNP based test (Pybus et al., 2013).

Type	Methods	Reference	Score tail	Feature
Allele frequency spectrum	CLR	Nielsen et al., 2005	Upper	Window (variable size)
Population differentiation	XP-CLR	Chen et al., 2010	Upper	Window (0.1 cM)
	Fst	Weir and Cockerham, 1984	Upper	SNP-specific
Linkage disequilibrium structure	XP-EHH	Sabeti et al., 2007	Upper	SNP-specific

As results, we present a set of candidate GRF genes that might have undergone positive selection in particular AMHs. We also identified several GRF gene families that may have significantly contributed to the evolution and adaptation of three

AMH populations. In addition, we report several regions, KRAB-ZNF gene clusters, which exhibit strong EHH in an Asian population (CHB). Finally, we discuss the potential biological roles of these candidate GRF families and genes, and how these findings could be interpreted in the sense of understanding human evolution.

3.2. Results

3.2.1. GRFs are over represented among genes showing extreme values for selective sweeps in AMHs

We first evaluated if GRF genes exhibit higher scores when compared with other genes by testing if the obtained ranked scores of all four tests are more extreme for all the GRFs than for the rest of the human genes (non-GRFs). We found that GRF genes exhibited higher scores when compared with non-GRFs (Wilcoxon rank test, p-value < 0.02) in all pair comparisons, except for CEU population for the CLR tests (Wilcoxon rank test, p-value > 0.05), suggesting that GRF genes are more often among the candidate regions for selection than other genes. We then chose set of random genes (of the same size as we had of GRF genes) 1000 times and performed the same Wilcoxon rank test to evaluate how often random genes would show higher rank scores. Our results still suggested that GRFs genes have higher rank scores values for the CLR test in the YRI population (p-value < 0.001). Similarly, XP-CLR results also indicated enrichment for GRFs in CEU and YRI populations when using CHB as reference population (p-values 0.03 and 0.006 respectively). F_{st} results also evidenced enrichment for GRFs exhibiting higher scores when testing for genetic differentiation between CEU and YRI versus CHB. In addition, we also tested if rank scores for GRF genes were enriched among the top 5% of the rank score distribution (>1.3 rank score) than all genes. To reduce the confounding demographic effects, we used scores based on the genome wide distribution (rank scores) and threshold corresponding to an empirical p-value of 0.05. Our results show that for XP-CLR and CLR tests, GRF genes were enriched among the top 5% ranks cores in all populations (Fisher's Exact test, p-value < 0.01), except in CHB for the CLR test (Fisher's Exact test, p-value >0.05). In addition, XP-EHH results also indicated that GRF genes are enriched for high scores for all three populations (Fisher's Exact test, p-value < 0.001).

Differences in recombination rates and gene length between GRFs and the rest of the human genes could influence the results. For instance, genetic recombination highly contributes to maintaining genomic diversity by crossing-over and independently assorting new combinations of alleles in the chromosomes of the offspring that were not previously present in the parental generation. Thus, it is expected that a region undergoing positive selection exhibits a reduction in recombination rates as results of the selection acting on it. Consequently, we evaluated if these two measures significantly differ between GRF and non-GRFs genes. We did not find a significant difference between the distributions of the recombination rates between GRF and non-GRF genes (Kolmogorov–Smirnov test; D = 0.019; p-value = 0.18). In addition, despite we found significant correlation between the gene length and the rank score (p-value < 2.2e-16), the correlation value is extremely small (rho = 0.009) and thus, might be negligible.

3.2.2. GRF classes are enriched among candidate regions for positive selection at population-specific level

Multiple GRF genes share similar sequence, structure, gene products and have diverged from a common ancestor. Several of these GRFs classes, or some of their members, have been suggested as important candidates for driving species' diversity, adaptation (Parmacek 2007) and speciation (Nowick et al. 2010; Nowick et al. 2013; Perdomo-Sabogal et al. 2016). We thus tested next, if any particular classes of GRFs are enriched among the GRF genes with high scores (top 5% distribution for all tests). To do this, we performed a Fisher Exact test for all of the 41 DNA-binding GRF classes cataloged so far (Chapter I) and adjusted the significance levels using Bonferroni correction. We found that 40% to 45% of the GRF classes were enriched (adjusted p-value < 0.05), at least in one population, among the upper 5% tail of the distribution for CLR and XP-CLR tests respectively. This number was even higher for the haplotype SNPs based method (77.5%) (Figure 11a and 11b).

Among the GRF classes showing enrichment in the upper 5% tail of the distribution for CLR, we found three of the smallest classes of GRFs (less than 20 genes): AT-

hook, AT-rich interaction domain (ARID), and Cys2-HisCys zinc finger (C2HC), which have six, 12, and 15 members respectively (adjusted p-value < 0.001) (Figure 11a). XP-CLR results also suggest enrichments for two out of these three classes, but for particular populations. For instance, the C2HC GRF class is enriched for CEU while ARID is enriched for CEU and YRI (adjusted p-value < 0.001) (figure 11b). We also found three of the larger GRF classes enriched for higher scores for CLR at population-specific level: Homeo Domain in CEU, and Tryptophan cluster and High-mobility factors in CHB (Figure 11a). In addition, several smaller GRF classes show population-specific enrichments for CLR, for instance, Paired box, NonO domain and CXXC zinc finger factors in CEU and CHB populations, Basic helix-span-helix, and STAT domain factors in CEU, and C3H zinc finger, MADS box and Heteromeric CCAAT-binding factors in YRI (Figure 11a). Likewise, the XP-CLR results also indicated an enrichment for Homeo Domain, Tryptophan cluster, High-mobility, and Paired box. In addition Fork head/winged helix factors class were also enriched with the XP-CLR method (Figure 11b).

Using the same strategy implemented for analyzing CLR and XP-CLR rank scores, we additionally tested if SNPs located in GRF genes are enriched for high XP-EHH rank scores. The results indicated a bigger number of GRF classes enriched for high XP-EHH scores (Figure 11d) when compared to CLR and XP-XLR results, thus suggesting that GRF genes from multiple GRF classes are likely to be located in regions exhibiting EHH. The Forkhead/winged helix factors class, one of the largest classes of GRF genes, was enriched in the 5% upper tail for all three human populations (Figure 11d).

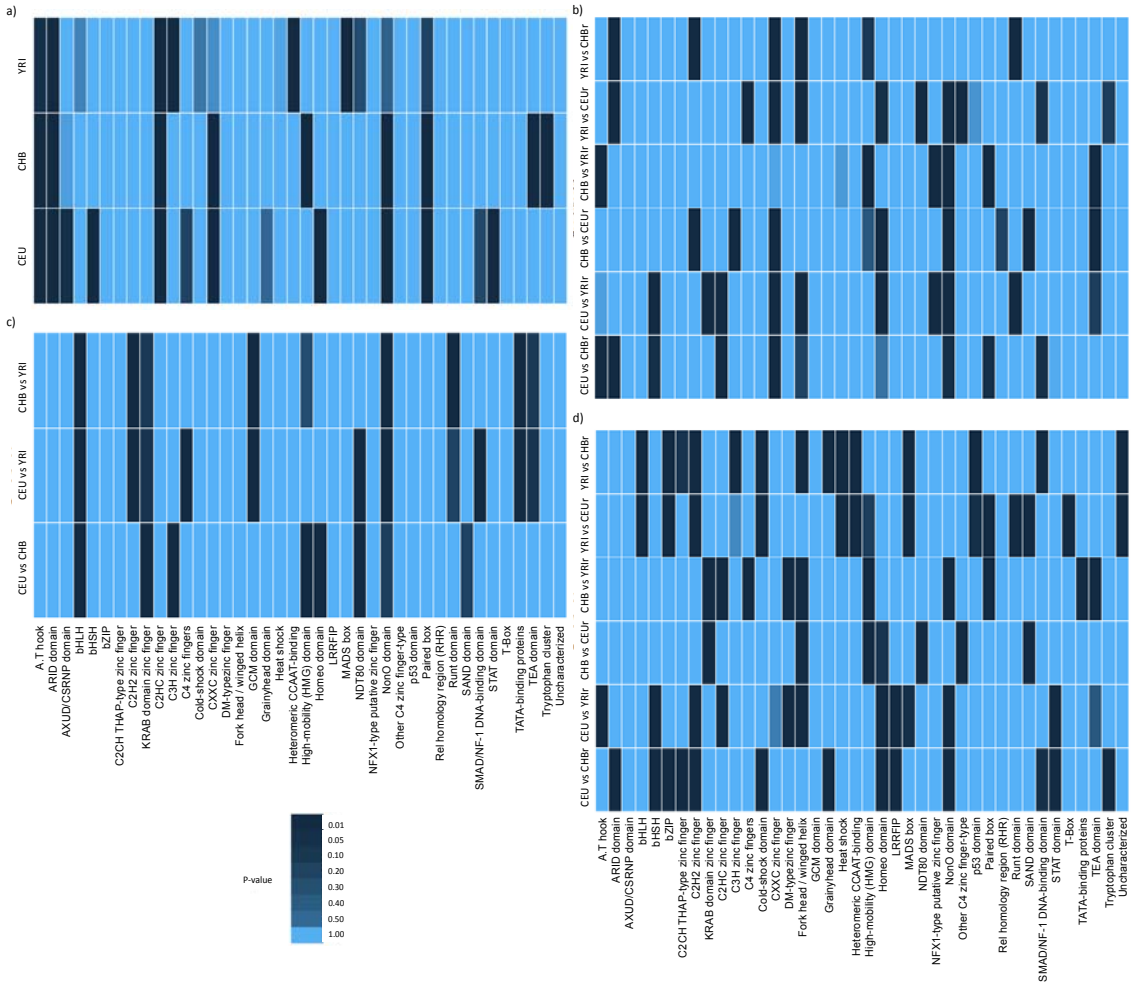


Figure 11. Enrichment analyses for DNA–binding transcription factors classes showing higher scores for three different tests for detecting positive selection and one for detecting genetic differentiation. (a) Composite likelihood ratio test (CLR). (b) Cross-population composite likelihood ratio test (XP-CLR). (c)Fixation index (F_{st}) test. (d) Cross-population EHH detected using XP-EHH method. Lower case r indicates the populations that were used as reference. Darker blue color indicates significant enrichment.

Several classes of GRFs also show population-specific XP-EHH enrichments. For instance, C2H2 zinc finger factors class, excluding KRAB-ZNF, exhibits enrichments for CEU and YRI populations (Figure 11d), while KRAB-ZNF, a subgroup of C2H2 genes that we treated as a separate class (see chapter I), showed enrichment just for CHB population. Two other of the largest classes of GRF genes, Basic helix-loop-helix (bHLH) and Basic leucine zipper factors (bZIP), also show population-specific enrichment for YRI, and CEU and YRI respectively. Similarly to the results obtained with CLR and XP-CLR results, the XP-EHH results also indicated enrichment for the Homeo Domain and Tryptophan cluster (CEU), High-mobility (CHB, YRI), and C2HC

(CEU, YRI) factors. Since methods based on EHH relies on LD, and this linkage break down over time, XP-EHH attributes allow the detection of recent selective sweeps; however, this test provides poor power to detect ancient selective sweeps (Chen et al. 2010). Therefore, the higher number of enriched GRF classes in regions exhibiting EHH may indicate that recent selective sweeps are more common between GRF genes.

To complement these results, we additionally explored which GRF classes are enriched in the 5% upper tail of the F_{st} rank scores distribution. This may indicate which GRF classes are contributing to high genetic differentiation between pairs of populations. Similarly to the results obtained with the other three tests, the F_{st} results suggest that at least three of the larger GRF classes (C2H2, Homeo Domain and High Mobility factors) are significantly contributing to high genetic differentiation among pairs of populations (Figure 11c).

Taken together, there are at least six GRF classes (15%) overrepresented among the genomic regions exhibiting signatures of positive selection: C2H2 zinc finger, KRAB-ZNF zinc finger, Homeo domain, Tryptophan cluster, Fork head / winged helix and,High-mobility (HMG) domain. (Figure 11). The six GRF classes are representatives of the larger DNA-binding transcription factor classes. Other three smaller GRF classes, High Mobility, ARID domain, C2HC and Paired box factors, also show enrichment within the candidate regions. Since each test for detecting positive selection produces a list of candidate regions, we further explored how many candidate genes, out of these six larger GRF families, are candidates in one or more populations. Considering that three of the CLR, XP-CLR and XP-EHH are composite methods, we initially subset those genes that had at least one value in the 5% upper tail of the distribution for each test in each population, and performed an overlap between these gene lists. This produced a set of genes that all three tests suggested as candidates for positive selection at population-specific level (Supplementary Table S4, supplementary data file). We additionally checked if some of these putative GRF candidate genes for selection at population-specific level were also present as candidates in another population (Table 4). Based on the strategy implemented (See Methods), our results suggest that all three tests indicated that

at least 69 (CEU), 64 (CHB) and 98 (YRI) genes from the aforementioned six GRF classes are candidates for positive selection (Table 4). We also detected a group of genes that are candidates between pairs of populations (Table 4). None of the candidate GRF genes was found as candidate for positive selection in all three populations for these six GRF families. The C2H2 and KRAB-ZNF classes exhibited the highest number of candidate genes for positive selection with 68 and 53 respectively, followed by Homeo domain with 31, Tryptophan cluster with 22 genes, Fork head/winged helix with 19, and High-mobility (HMG) domain with nine. Our results also suggest that the majority of the genes reported for the three tests for these six GRF classes presented population-specific signatures of selection (Table 4).

Table 4. Members of six of the larger GRF classes enriched among the high scores for the three tests (CLR, XP-CLR and XP-EHH) for detecting selection analyzed here (upper 5% tail of the distribution). The GRF class of each gene is indicated as following: C2H2 zinc finger (¢), KRAB-ZNF zinc finger (£), Homeo domain (♯), Tryptophan cluster (*), Fork head / winged helix and (¥), High-mobility (HMG) domain (≈).

Population	GRF genes	Total
CEU, YRI	ZNF280D¢, ZFHX3♯, NCOR2*, FOXO1¥, ZNF528£, RFX8¥, PGBD1£, MEOX2♯, FOXP2¥, ONECUT2♯	10
CHB, YRI	ZNF768£, GLI3¢, ZFAT¢, ZNF668£, ELMSAN1*, ZNF521¢	6
CEU, CHB	RFX3¥, VEZF1¢, TRERF1¢, E2F4¥, ZNF511¢, FOXA2¥, MTA1*, GLI2¢, ZNF407¢, SMARCC2*, ZNF844£, FOXK1¥, SALL3¢	13
YRI	ETS1*, MTF1¢, ZNF592¢, PRDM2¢, SOX6≈, ZNF679£, ZNF251£, SPI1*, HIC2¢, KLF12¢, FOX06¥, LHX5♯, FOXJ2¥, ZNF131¢, TFDP1¥, ELK3*, ZSCAN20¢, ZNF83£, CTCFL¢, ZBTB25¢, ZSCAN2¢, ETV6*, ERG*, PHTF1♯, ZBTB46¢, ZBTB41¢, ZNF644¢, HKR1£, SOX15≈, ZNF3£, ZNF827¢, SP8¢, ZNF678£, SOX5≈, ZNF250£, SOX7≈, ZNF319¢, ZBTB40¢, HIVEP2¢, ZNF483£, HBP1≈, ZNF77£, PRDM15¢, HOXB1♯, SATB1♯, RFX7¥, TSHZ2♯, KLF17¢, PRDM4¢, IKZF2¢, ZNF70¢, ZNF181£, ZNF423¢, PBX4♯, ZNF396£, ZNF217¢, ELF2*, TRPS1¢, PATZ1¢, MTA3*, LMX1A♯, HOXD3♯, ZNF354A£, ZSCAN16£, ELF5*, PKNOX1♯, ETV4*, ZNF438¢, IRF1*, ZNF311£, HIVEP1¢, FOXN2¥, E2F2¥, ZNF532¢, ZNF14£, TOX3≈, ZHX2♯, HMX2♯, RREB1¢, ZFP64¢, HMG20A≈, POU3F2♯	82
CEU	LHX1♯, ZNF155£, IRF3*, ZNF653¢, IKZF1¢, ZNF546£, POU2F3♯, ZNF93£, ZNF277¢, ZSCAN25£, RFX2¥, POU6F2♯, ZNF534£, SPDEF*, CDC5L*, ZNF252¢, ZNF341¢, MKX♯, ZNF224£, EVX1♯, ZNF572¢, ZEB1♯, ZNF90£, PRDM9£, ZNF284£, ZNF780B£, PRDM10¢, ZNF707£, ZNF740¢, ZNF76¢, MECOM¢, ZNF253£, ZNF193£, DLX3♯, GABPA*, MIER2*, ZNF280B¢, ZKSCAN5£, ZNF536¢, FOXM1¥, ZNF780A£, GLIS1¢, CUX2♯, ZNF649£, ZKSCAN4£, EGR4¢	46
CHB	ZNF167£, ZNF282£, ZNF197£, RFX4¥, CRX♯, ZNF425£, NCOR1*, ZNF263£, ZNF786£, ZNF660£, FOXJ1¥, ZNF467£, ISX♯, ZNF124£, ZNF512B¢, HESX1♯, ZNF562£, TCF7L1≈, ZNF445£, ZFP161¢, CUX1♯, ZNF646£, ZBTB4¢, BBX≈, ZNF695£, ZBTB20¢, GLIS2¢, FOXR1¥, NKX6-3♯, ZNF317£, EMX1♯, CASZ1¢, NOBOX♯, FOXP1¥, LHX8♯, RERE*, ZNF398£, GFI1¢, ZBTB7B¢, PRDM6¢, ZNF579£, ZNF451¢, ZIC5¢, HIVEP3¢, IRF7*	45

We also found that around 121 C2H2 GRFs, from which 53 are KRAB-ZNF genes, are located in regions that exhibit patterns that are consistent with positive selective events (as mentioned in the general introduction) either at population-specific

level, or between pairs of populations. Out of this set of genes, just nine C2H2 GRF genes have been reported as candidates for positive selection in previous works (*ZFAT, ZBTB41, ZNF827, IKZF2, ZNF438, ZNF546, ZNF780B, ZNF780A, ZBTB20*) (Table 5). In total, out of the 202 GRF genes we found as candidates for positive selection for these six GRF classes, 20 genes have been found in candidate regions for positive selection in previous studies (Table 5).

Table 5. GRF genes reported as candidates for positive selection here for six GRF classes, and that have been previously reported in genome wide scans for selection in humans.

GRF genes	Number of GRF genes	Source
BBX	1	Sabeti et al. 2007
RFX3	1	Pickrell et al. 2009
ISX	1	Metspalu et al. 2011
CUX2, ETV4, FOXP1, FOXP2, IKZF2, LHX8, PHTF1, PKNOX1, POU2F3, ZBTB20, ZBTB41, ZFAT, ZNF438, ZNF546, ZNF780A, ZNF780B, ZNF827	17	Grossman et al. 2013

By using a different composite of multiple signals test (CMS), Grossman et al. (2013) identified that the gene *ZBTB41* is likely to be under selection in YRI, while the genes *ZFAT, IKZF2, ZNF438*, exhibited high CMS scoring SNPs. In addition, Grossman et al. (2013) also suggested that the genes *ZNF827* in YRI, and *ZNF780A, ZNF780B, ZNF546*, and *ZBTB20* in CEU exhibit population-specific localized (genomic regions with a median size of 27 kb) signatures of positive selection.

Among the candidates for positive selection at population level that we detected here, and that have not been previously reported, we identified three GRFs that have been associated with insulin/glucose regulatory pathways, zinc finger protein 407 (*ZNF407*) (Buchner et al. 2015), and forkhead box O1 and A2 (*FOXO1*, *FOXA2* (Kamagate et al. 2008; Wu et al. 2016; Yalley et al. 2016). Our results indicate that *ZNF407,* a C2H2 GRF type, is a candidate gene for positive selection in CEU and CHB populations. Further exploration of this region evidenced that 17 high scoring SNPs for selection (XP-EHH test) result in missense mutations (rs183921097, rs3794942, rs74861823, rs115368653, rs77518676, rs114313623, rs116304324, rs147684864, rs77006793, rs7227263, rs183172085, rs149806516, rs948615, rs73971116, rs185745193, rs34048449, rs34141917). None of these missense mutations have an effect on the amino acid sequence of the protein domains of

ZNF407. We also found that the genes *FOXO1* and *FOXA2* are candidates for positive selection in CEU and YRI, and CEU and CHB, respectively. Further exploration of this genomic regions revealed that *FOXO1* carries three missense mutations, two in YRI (rs148727582, rs70961707) and one that is present in all three populations (rs34733279). Regarding the gene, *FOXA2*, we did not find sequence variation causing missense mutations that could explain functional changes followed by positive selection. We further explored for protein coding genes located around *FOXA2* and that could be hitchhiking this gene; however, there is none. We did find two long non-coding RNAs (lncRNAs) located around 3 kb and 17 kb distant of *FOXA2*, LINC00261 and LINC01384.

3.2.3. Signatures of selection are enriched on protein domains for C2H2, KRAB-ZNF and bHLH GRF classes

Gene regulatory proteins, as single and independent evolutionary units, can consist of either a single functional domain or form more complex multi-domain proteins (Vogel et al. 2004). Non-synonymous changes in the protein domains are expected to have an effect at functional level, either on the specificity the protein binds to the DNA or interacts with other proteins as cofactor. Using information from the SNP based tests (XP-EHH and F_{st}), we explored if higher scores were mostly occurring in functional domains, and if these correspond to synonymous or non-synonymous SNPs. XP-EHH results suggest a significant enrichment for synonymous SNPs with high XP-EHH scores for the GRF classes C2H2, excluding KRAB-ZNF genes, in CEU (Fisher Exact test, p-value 0.02), while KRAB-domain zinc fingers for CHB (Fisher Exact test, p-value 0.04) (Table 5). In both cases YRI was used as the reference population. We also found that the bHLH GRF class was almost significantly enriched within the top 5% for YRI when using CHB as the reference population (Fisher Exact test, p-value 0.057). In addition, the F_{st} results suggest enrichment for synonymous SNPs for the C2H2 zinc finger class, excluding KRAB-ZNFs, between CEU and CHB when compared to YRI (Fisher Exact test, p-value < 0.001) and KRAB-ZNF class between CHB and CEU (Table 6).

Table 6. GRF classes showing enrichment based on the Fisher's Exact test for synonymous SNPs in the 5% upper tail of the distribution for XP-EHH and F_{st} test.

Test	Populations	GRF class	p-value	Nonsynonymous SNPs in regions coding for protein domains	Synonymous SNPs in regions coding for protein domains	Nonsynonymous SNPs in regions not coding protein domains	Nonsynonymous SNPs in regions not coding protein domains
XP-EHH	CEU vs YRI	C2H2 zinc finger	0.021	7	16	55	40
	CHB vs YRI	KRAB domain zinc finger	0.041	68	51	40	14
	YRI vs CHB	bHLH	0,0570	0	6	11	12
Fst	CEU vs CHB	KRAB domain zinc finger	0,0490	41	34	21	1
	YRI vs CHB	C2H2 zinc finger	0,001	2	19	44	46
	YRI vs CEU	C2H2 zinc finger	0,0003	1	18	59	64

3.2.4. KRAB-ZNF gene clusters exhibit regions with EHH and high genetic differentiation in CHB

Considering that KRAB-ZNF genes were enriched for XP-EHH scores in CHB, and that F_{st} results also suggested an enrichment for synonymous SNPs in CHB versus CEU (Table 5), we further explored whether particular KRAB-ZNF clusters exhibit EHH regions. We initially determined where in the human genome KRAB-ZNF genes are located. A particular characteristic of KRAB-ZNF genes is that a majority of them are arranged in gene clusters longer than 150 kb (Supplementary Table S5), with some cases assorted with non-KRAB-ZNF genes. Using the cluster's annotation for KRAB-ZNF genes as described in Huntley et al. (2006) (Supplementary table S5), and the results obtained using XP-EHH and F_{st} tests, we explored if KRAB-ZNF gene clusters exhibit regions with uninterrupted high scores for XP-EHH (5% upper tail of the score distribution) (Supplementary Figure S1), and if that was the case, we also explored if the F_{st} scores suggested high genetic differentiation.

CHB exhibited the longest EHH regions when using CEU and YRI as the reference populations (Supplementary Figure S1); with three out of 25 KRAB-ZNF clusters (one, three, and 14, which are located in chromosomes one, three and 16 respectively) exhibiting uninterrupted number of SNPs with high ranked scores for XP-EHH (>1.3) for regions larger than 100 kb (CHB when using CEU as the reference population)(Table 7, Figure 12). To test whether the number of SNPs with uninterrupted high XP-EHH scores we observed in these regions with EHH>100kb is more than expected by chance, we randomly drew genomic regions of the same size 1000 times and counted the number of SNPs with high XP-EHH rank scores (>1.3) in these regions.

Table 7. KRAB-ZNF gene clusters exhibiting EHH haplotype regions larger than 100 kb. Cluster coordinates based on Huntley et al. 2006.

KRAB-ZNF clusters	Genomic coordinates	Size region (Kb)	EHH coordinates	Size EHH (Kb)	SNP withinthe EHH region	Type of signature (selective sweep)*	Population
1	chr1:247.14-247.24	240	chr1:247.14-247.24	101	644	incomplete ancient	CHB
3	chr3:44.55-44.74	400	chr3:44.55-44.74	188	563	complete recent	CHB
14	chr16:31.01-31.16	864	chr16:31.01-31.16	156	445	incomplete ancient	CHB

* Based on Hierarchical boosting framework and data from Pybus et al. 2015

These results indicated that the KRAB-ZNF gene clusters one, three, and 14, harbor regions exhibiting EHH with higher XP-EHH scoring SNPs than expected by chance (Bonferroni adjusted p-value < 0.04). Considering that genomic regions exhibiting signatures of positive selection also show low recombination rates, we tested if the rates observed for these three EHH regions were smaller than expected by chance. Implementing the same strategy used above, we randomly drew genomic regions of the same size of the EHH (>100 kb) 1000 times, calculated the mean of the recombination rates, and counted how often this values were smaller that the mean of the recombination rates for the EHH regions. The results indicated that the recombination rates observed in the EEH found within these three KRAB-ZNF clusters were smaller than expected by chance (Bonferroni adjusted p-value < 0.009).

In two out of these three KRAB-ZNF clusters (one and three), SNPs from KRAB-ZNF genes exhibited the highest scores (Figure 12a and b), while for the cluster 14, the gene *VKORC1* presented one intronic SNP (rs140525321, first intron) with the highest XP-EHH score (4.25) (Figure 12c). The KRAB-ZNF cluster 14 also seems to have a larger EHH region of 251 Kbps (chr16:30914142-31165239) (Figure 12c). However, the haplotype decayed below the 1.3 rank score threshold at position chr16: 31009343 up stream direction of the EHH, where one SNP (rs74474326) located in the fourth intron of the gene *STX1B* exhibited a lower ranked XP-EHH score (1.08) (Figure 12c). Further inspection of the minor allele frequency (MAF) for this SNP revealed that the alternative variant is fixed for CEU, while the MAF for CHB was 0.11, while the situation for the MAF for the neighboring SNP (rs80168914) was the opposite (fixed in CHB while showing MAF = 0.0051 for CEU). Therefore, it is likely that this lower peak was due to an artifact introduced by the

way in which the MAF was defined in one of the populations. Further exploration of the downstream region of the EHH in the KRAB-ZNF cluster 14 showed continuous XP-EHH haplotype decay (Supplementary Figure S2). Interestingly, we found that six of the genes (*KAT8*, *ZNF646*, *ZNF668*, *FBXL19*, *STX1B* and *VKORC1*) within the EHH from the cluster 14 (251 kb) have been recently associated with obesity (Locke et al. 2015; Yazdi et al. 2015) and anticoagulant response (Patillon et al. 2012) in humans. Two SNPs within this region, one located on the K(lysine) acetyltransferase 8 (*KAT8*) GRF gene (rs9925964), and the other on the gene vitamin K epoxide reductase complex subunit 1(*VKORC1*) gene (rs10871454) have been suggested to have a causative effect on obesity and anticoagulant effect, respectively (Patillon et al. 2012; Locke et al. 2015; Yazdi et al. 2015).

Similarly to the haplotype decay found in the cluster 14, The KRAB-ZNF cluster three also exhibits two EHH regions around the larger EHH. These two regions span between 50 and 80 kb (Figure 12b). The decay on the upstream region of the larger EHH in the KRAB-ZNF cluster three is explained by two SNPs (rs6789709 and rs6441848) occurring on the gene ZNF445, rank scores 1.23 and 1.26 respectively (Figure 12b). These two SNPs exhibit the very same MAF in both populations, rs6789709 and rs6441848 (CEU = 0.1111 and CHB = 0.1117). It possible that these two SNPs causing the haplotype decay in this region of the KRAB-ZNF cluster three obtained a lower rank based on the whole genomic distribution, but they still belong to a larger EHH region in the cluster three. On the downstream region of the larger EHH in the KRAB-ZNF cluster three, the decay is explained by 95 SNPs with score values below the set 1.3 rank score threshold. This region extends around 25 kb (Figure 12b). This suggests that the EHH region on the KRAB-ZNF cluster three extends at least 272 kb.

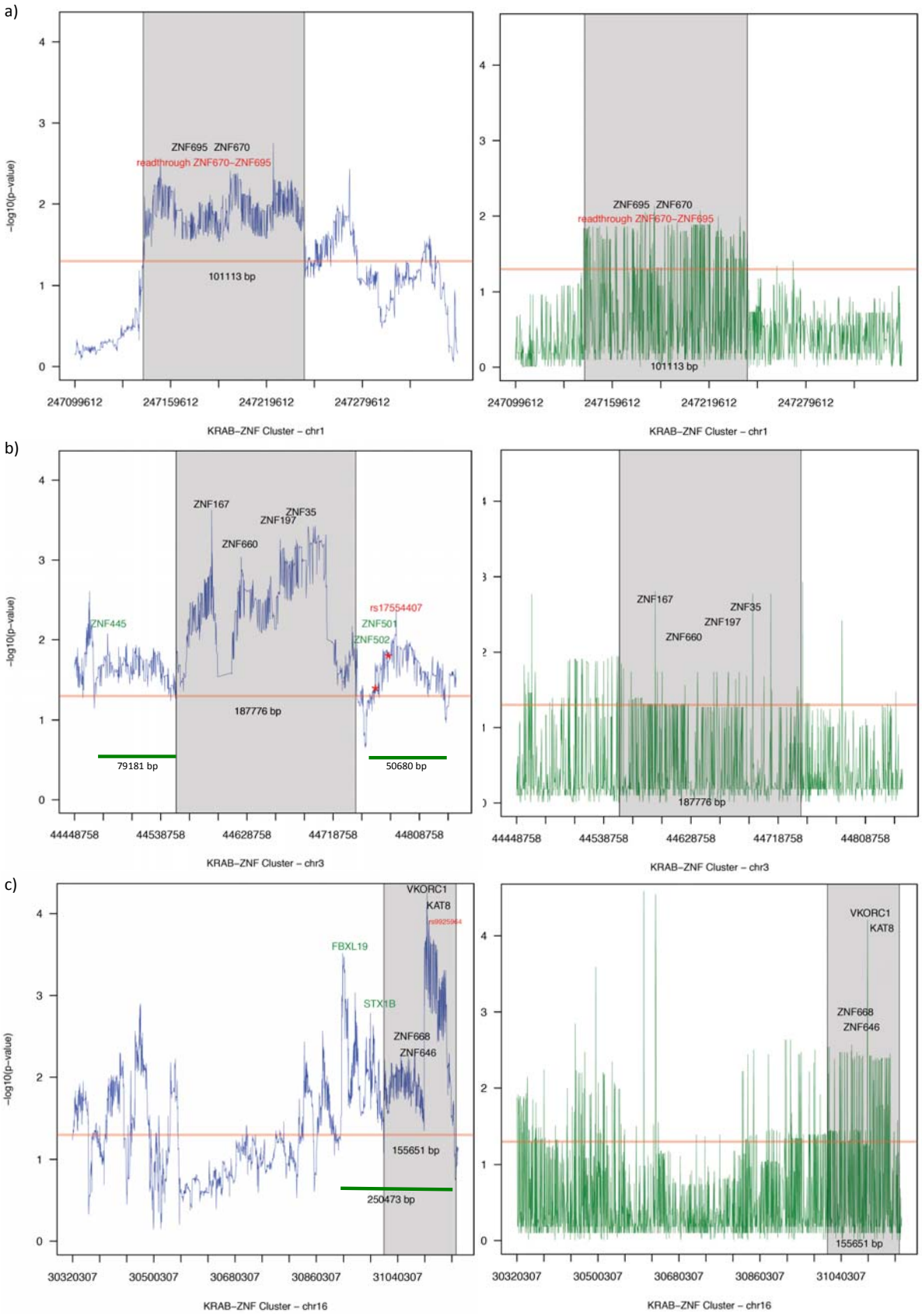


Figure 12. KRAB-ZNF gene clusters showing extended haplotype homozygosity (XP-EHH) regions and high genetic differentiation (F_{st}) for CHB when using CEU as reference population. This indicate that these regions have experienced selective sweeps. (a-left panel) KRAB-ZNF gene cluster one

located on the human chromosome one. The extended haplotype region (grey area) spans around 100 kb (chr1:247142296-247243409) and contains two GRF genes, ZNF670 and ZNF695. Transcription of these two genes also produces readthrough or co-transcribed genes (ZNF670-ZNF695) (in red). (a-right panel) Rank score Fst values for the KRAB-ZNF gene cluster one. The majority of the SNPs showing high genetic differentiation are located within the EHH region. (b-left panel) KRAB-ZNF gene cluster three located on the human chromosome three. The extended haplotype is about 188 kbps long (chr3:44554702-44742478) and harbors four ZNF genes. Vertical green lines indicate other two regions around the larger EHH region that also exhibit EHH. Red stars indicate nonsynonymous SNPs located in genes *ZNF502* (rs185260708, rs181738022) and *ZNF501* (rs150433704). The SNP rs17554407 is located in a non coding region (b-left panel). F_{st} Rank score values for several SNPs suggest high genetic differentiation between CHB and CEU populations in this region. (c-left panel). KRAB-ZNF gene cluster 14 located on human chromosome 16. The extended haplotype is about 156 kbps long (chr16:31009588-31165239) and contains four GRF (three being ZNF) and one non-GRF genes (*VKORC1*). The Non-GRF gene (*VKORC1)* exhibited the highest scores for XP-EHH and F_{st}) within the EHH region. Light horizontal red line indicates the 5% threshold set for the XP-EHH and F_{st} distribution (rank score of 1.3). Green horizontal line (c-left panel) indicates a region of around 251 kb in length showing high XP-EHH scores with one SNP causing the EHH decay at position chr16:31009343 (rs74474326). This EHH decays around the gene STX1B and *FBXL19* (in green), two genes that have also been associated with obesity (Locke et al. 2015; Yazdi et al. 2015). None of these plots shows all the genes located within these three KRAB-ZNF gene clusters.

Similarly to XP-EHH results, F_{st} scores suggested that the extended haplotypes within these three clusters exhibit high genetic differentiation between CHB and CEU (Figure 12). Besides XP-EHH and F_{st} results for these three clusters, CLR and XP-CLR results also indicate the presence of extreme values, however, these do not exhibit a clearly defined pattern within the regions showing the EHH (Supplementary Figure S3).

Considering that genetic variability is strongly affected by demographic processes, we additionally explored the hierarchical boosting simulation framework and its data published by Pybus et al. (2015). This boosting framework is of great utility in uncovering scenarios of complete/incomplete and ancient/recent selective sweeps while controlling demography (Pybus et al. 2015). As result, it was possible to detect that these three KRAB-ZNF clusters (one, three, and 14) contain EHH regions that might have undergone either incomplete and ancient (EHH in clusters one and 14) (Figure 13 a and c) or complete and recent selective sweeps (EHH in cluster 3) (Figure 13b).

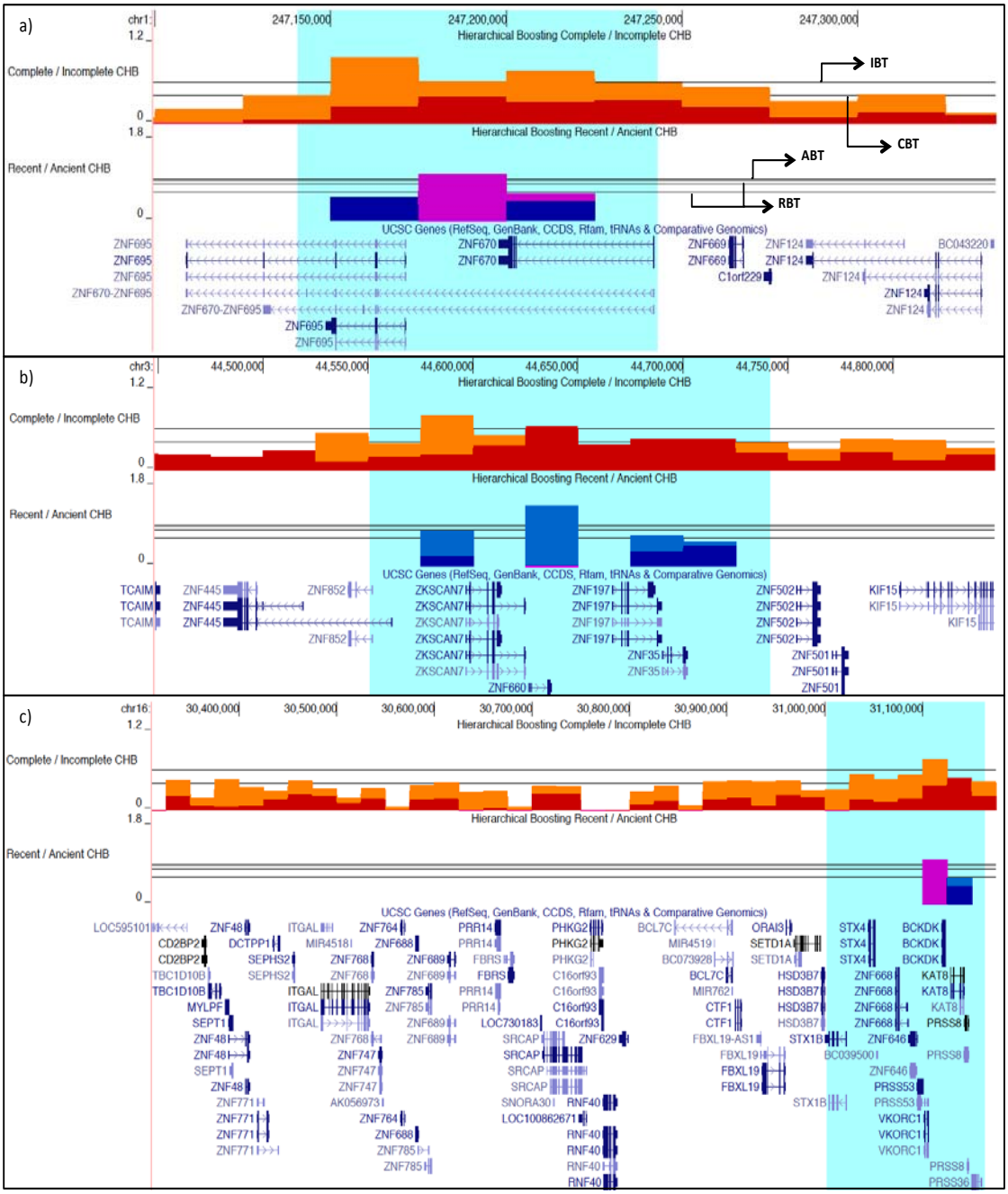


Figure 13. Hierarchical boosting simulation data for three KRAB-ZNF gene clusters in CHB population. (a) KRAB-ZNF cluster one, which is located in human chromosome one. (b) KRAB-ZNF cluster three, which is located in human chromosome three. (c) KRAB-ZNF cluster 14 from human chromosome 16. The four color tracks represent composite scores obtained for these regions: Complete (red), incomplete (orange), recent (blue) and ancient (violet) selective sweeps. As described in Pybus et al. 2015, each track has its own significance threshold (1% false discovery rate (FDR). The recent/ancient tracks reference an approximate age; however, precise dating demands more sophisticated computational approaches. Completeness/incompleteness of the selective sweep is represented by thresholds tracks on the figure. Higher threshold corresponds to incomplete scores (orange). The lower thresholds correspond to complete scores (red). Incomplete boosting (IBT), complete boosting (CBT), Ancient boosting (ABT), and Recent boosting (RBT) thresholds.

3.2.5. SNPs in KRAB-ZNF protein domains as a source of regulatory diversity

Despite the results from the XP-EHH and F_{st} tests indicated that there is an enrichment for synonymous SNPs occurring in regions coding for protein domains of KRAB-ZNF genes, we also found many non-synonymous changes in KRAB-ZNF genes (Tables 5 and 7). In addition, the population-specific putative signatures of selection found for this group of genes suggest that these SNPs may be playing important roles in introducing regulatory diversity at AMHs. Therefore, we further explored the presence of non-synonymous SNPs with high XP-EHH scores in the KRAB-ZNF gene clusters exhibiting EHH and examined potential functional effects of such SNPs using the UCSC genome browser (Karolchik et al. 2014) and UniProtKB (The UniProt Consortium 2015) data bases. Within the whole genomic region for these three KRAB-ZNF clusters, between 43% and 92% of the nonsynonymous SNPs are located in KRAB-ZNF genes (Table 8). Among the regions showing EHH, one of the KRAB-ZNF genes (ZNF646) located in the cluster 14 harbor 11 of the nonsynonymous SNPs that code for amino acids from the protein domains. The other two clusters harbor two (cluster one) and one (cluster three) SNPs introducing missense changes in regions coding for protein domains, all located on three KRAB-ZNF genes and one readthrough region (*ZNF670-ZNF695*) (Table 8).

Table 8. Nonsynonymous SNPs and GRFs located in KRAB-ZNF clusters exhibiting regions of EHH.

KRAB-ZNF clusters	EHH regions	Nonsynonymous SNPs	Nonsynonymous SNPs KRAB-ZNF genes	Nonsynonymous SNPs (EHH)	Nonsynonymous SNPs in regions coding for protein domains (EHH)	ZNF Genes with nonsynonymous SNPs (EHH)
1	chr1:247.14-247.24	13	12	6	2	*ZNF670-ZNF695, ZNF695*
3	chr3:44.55-44.74	32	25	5	1	*ZNF167, ZNF35*
14	chr16:31.01-31.16	83	36	27	11	*ZNF646*

We further explored if some of these non-synonymous SNPs cause to changes in important residues of the protein domains, for instance, residues that are essential for the stability of the zinc finger folds, or important for the interaction between the GRF with other co-factors (KRAB domains) (Figure 14). Out of the 14 nonsynonymous SNPs found for KRAB-ZNF genes in the EHH regions previously

described in Figure 12, three SNPs introduce missense mutations that cause amino acid changes between two histidine residues that provide stability to the fold of two zinc fingers of the gene *ZNF646* (cluster 14): rs35376811 (Arginine to Tryptophan); rs188200157 (Isoleucine to Leucine) and rs3751856 (Arginine to Glutamine). Other nine nonsynonymous SNPs (Supplementary Table S6) were located in regions that code for amino acid residues located between two KRAB-ZNF zinc fingers, from 2 to 16 amino acids distant. Three out of these nine nonsynonymous SNPs are affecting the sequence of the linker regions of zinc fingers located in close proximity, (less than 11 amino acids) before the next finger starts, for two genes: rs140747159 (*ZNF695*), rs141631516 and rs75586809 (*ZNF646*). The gene *ZNF35*, a non-KRAB-ZNF zinc finger gene located on the EHH found in the cluster three, harbors one nonsynonymous SNP (rs191633770) that changes the amino acid sequence from phenylalanine to serine. This change occurs two residues away from N-terminal end of the first Cysteine of the fourth zinc finger, in the linker region.

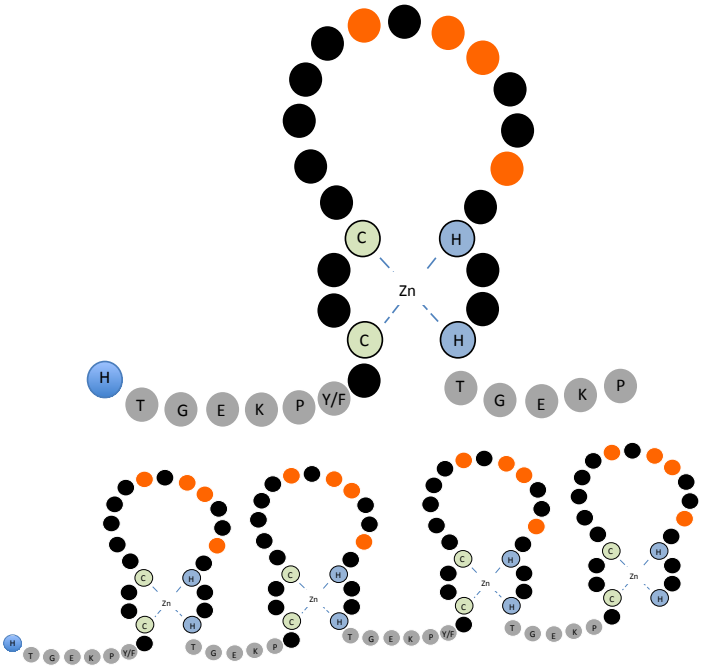


Figure 14. Illustrative representation of a C2H2 zinc finger domain. The two cysteines and histidines, green and blue respectively, bind to the zinc ion, thus providing stability of the fold. The linker sequences that frequently join adjacent fingers (bottom) are shown in grey. Four residues at positions -1, 2, 3 and 6 counted relatively to the alpha helix provide sequence specificity (orange circles). The black residues are considered not structurally relevant. The number of residues between cysteines and histidines (black circles) may vary.

Considering that changes in the functional regions within the three KRAB-ZNF gene clusters could explain the type of selective event causing the EHH, we explored the ancestral states of all nonsynonymous SNPs within the regions exhibiting EHH (>100 kb) in KRAB-ZNF clusters one, three and 14. All nonsynonymous SNPs located in coding regions within the KRAB-ZNF cluster 14 corresponded to nucleotide variants that differed from the ancestral state (Supplementary Table S6). Conversely, KRAB-ZNF cluster one and three presented two and one nonsynonymous SNPs that correspond with the ancestral allele (Supplementary Table S6). In summary, this suggest that 35 out of 37 nonsynonymous SNPs located in these EHH regions may have resulted from *de novo* mutations.

In addition, we also found nonsynonymous SNPs located in the KRAB-ZNF cluster three (chromosome 3, EHH region 50 kb long) that cause changes in residues that are essential for the stability of the C2H2 zinc finger fold of two GRF genes (*ZNF501* and *ZNF502*), and that seem to be population specific (Figure 12b). Two SNPs (rs185260708 and rs181738022) corresponding to the fifth zinc finger of the gene *ZNF502*, which is located in the KRAB-ZNF cluster three, introduce non-synonymous changes that modify the amino acid sequence of the first Histidine to Leucine. These two SNPs showed high XP-EHH for CHB using CEU as reference population (XP-EHH rank score > 1.32). It is likely that these nonsynonymous mutations modify the function of ZNF502 by rendering its fifth finger non-functional. Another example is the SNP (rs150433704) located in the region coding for the sixth zinc finger of the zinc finger gene *ZNF501*. This SNP changes the first Cysteine for Tyrosine, thus similarly indicating that the folding stability of this zinc finger is changed. The SNP rs150433704 also showed a high XP-EHH score (1.71) in CHB when using CEU as reference. These three amino acid changes suggests that changes in the zinc finger domains of particular GRF genes might have an important effect in introducing regulatory diversity between human populations.

3.2.6. Evolutionary older GRFs are main candidates for selection

To identify if human and primate-specific genes are enriched among the genes located in regions with high scores for all tests used, we assigned each gene to their particular evolutionary branch, from those human-specific to those present in vertebrates, by using gene branch assignments (Zhang et al. 2010). We were able to assign 3053 GRF genes to their particular clade (Supplementary Table S7, supplementary data file). In total, we found 15 human and 92 primate-specific GRF genes. our results indicated there is a significant under-representation of human–specific GRF genes among the genes showing high scores for CEU (CLR) and YRI (CLR and XP-EHH) (Fisher Exact test p-values = 0.02, 0.002 and 0.04, respectively). We also found that primate-specific GRF genes were under-represented for CHB (XP-CLR) (Fisher Exact test p-values = 0.01). This suggests that rather old established GRFs are mainly among the candidates for selection within AMHs. Another possible reasons could be that not enough time has passed since the human-specific genes appeared in the human genome and selection still has not left a stronger signature on these genomic regions. It is also likely that new GRFs are neutrally evolving.

3.3. Discussion

Using the results obtained for three different methods, we found that GRF genes are enriched among the candidate genomic regions for positive selection in three human populations. We found that larger groups of GRF classes such as C2H2 zinc finger, KRAB-ZNF zinc finger, Homeo domain, Tryptophan cluster, Fork head/winged helix, and High-mobility (HMG) domain were enriched at least for one population. We also found that at least three KRAB-ZNF clusters have regions bigger than 100 kb exhibiting EHH in CHB population. C2H2 and KRAB-ZNF classes showed enrichment for synonymous SNPs in protein domains. In addition, the KRAB-ZNF class exhibited the major number of nonsynonymous SNPs occurring in protein domains. Further exploration of non-synonymous SNPs with high scores for XP-EHH from GRF classes enriched for positive selection suggest that KRAB-ZNF genes may be important for introducing regulatory diversity in human populations.

3.3.1. GRFs classes enriched among candidate regions for positive selection

The molecular bases to depict how phenotypic differences between human populations have been evolutionary shaped are still far from being fully understood; nonetheless, variation in the transcriptional regulome is likely to play an essential role in fine-tuning the ways in which the current diversity we observe in AMH's traits is expressed (Wray 2007; Nowick et al. 2011; Albert and Kruglyak 2015; Barrera et al. 2016). Genomic sequences carrying the information that codes for GRF proteins hold, to some extent, important clues of the recent human evolutionary adaptation and diversification. For instance, changes introducing subtle variation in the DNA-binding affinities could have substantial functional effect in generating phenotypic variation without being deleterious (Barrera et al. 2016). Genetic variation, especially the variation that confers particular advantages for the individual, is expected to be subjected to natural selection (Gillespie 1991). our exploration of genome wide scans for signatures of positive selection suggest that particular GRF classes have widely contributed to this adaptive variation at population-specific level. Among the whole set of genes we explored here, we identified that six of the larger GRF classes are enriched among the regions exhibiting high scores for tests for detecting positive selection in humans: C2H2 zinc finger, KRAB-ZNF zinc finger, Homeo domain, Tryptophan cluster, Fork head/winged helix and, and High-mobility (HMG) domain. Twenty out of the 202 genes we found as candidates for positive selection from six GRF classes have also been reported as candidate regions for positive selection in previous studies. It is likely that methodological criteria and levels of significance used as threshold in previous works on genome-wide scans for selection (Sabeti et al. 2007; Pickrell et al. 2009; Metspalu et al. 2011; Grossman et al. 2013) left some many of these 182 GRF genes out of the candidate lists. It is also plausible that some of the genes we are reporting here correspond to new candidates for positive selection not previously described.

Changes in diet during human dispersals Out of Africa were likely to go together with Adaptive molecular variations in metabolic pathways. One remarkable example of such metabolic adaptations is the lactase persistence in European and

some African populations (Bersaglieri et al. 2004). Changes in carbohydrate intake in the form of starch as a source of glucose are thought considerable boosted the energy resources in human tissues with high glucose requirements such as the brain, blood cells, and developing embryos (Hardy et al. 2015). Such variations in diet may have been accompanied by adaptive changes in gene regulatory pathways. Among the set of candidates GRF genes for positive selection, we identified three genes that play key roles in insulin/glucose regulatory pathways: zinc finger protein 407 (*ZNF407*) (Buchner et al. 2015), and forkhead box O1 and A2 (*FOXO1*, *FOXA2*) (Kamagate et al. 2008; Wu et al. 2016; Yalley et al. 2016). Insulin-mediated glucose absorption in adipocytes in essential for keeping the glucose homeostasis and insulin sensitivity in the whole body (Bogan 2012). This metabolic pathway is greatly mediated by the solute carrier family 2, member 4 (*SLC2A4*; *GLUT4*) gene. It was recently discovered that ZNF407 GRF protein controls the transcription and mRNA stability of *SLC2A4*, and that Knockdown experiments on gene *ZNF407* resulted in 30– 40% reduction in insulin-stimulated glucose uptake (Buchner et al. 2015). In spite the biological roles of the majority of C2H2 proteins are still unknown, those C2H2 genes that have been functionally characterized participate in a wide repertoire of molecular regulatory roles such as protein-protein interaction, RNA binding, sequence-specific binding to DNA. Some DNA-binding C2H2, for instance, KRAB-ZNF genes, are involved in genomic recombination and chromosome segregation (Stubbs et al. 2011). Although none of these 17 missense SNPs showing signatures of selection on *ZNF407* is affecting the residues that are essential for the stability of the zinc finger folds and the affinity of the DNA binding protein domains, we suggest they could be introducing subtle regulatory variation in the ZNF407 protein structure, thus altering the pathways this gene is associated to; for instance, insulin-mediated glucose uptake.

The genes *FOXO1* and *FOXA2*, other two GRF candidate genes for positive selection, have been recently associated with the regulation of Insulin-sensitive pathways (Puigserver and Dominy 2010; Cheng et al. 2016; Yalley et al. 2016). These two genes are members of another large GRF classes, the Fork head/winged helix, an evolutionary conserved GRF class that gathers genes with key functional roles in glucose metabolic enzymes, apoptotic factors, and cell cycle regulators in multiple

tissues(Ho et al. 2008). FOXO1 binds the promoter region of insulin-like growth factor-binding protein 1 (IGFBP1) and glucose-6-phosphatase (G6Pase) in response to insulin signaling in liver, thus regulating their expression (Yalley et al. 2016). By opening and remodeling the chromatin on the *IGFBP1* promoter region, FOXO1 significantly increases the binding activity of RNA polymerase II and other two pioneer GRF proteins (FOXA1/A2) (Hatta and Cirillo 2007). By using knock down experiments in FOXO1 and FOXA2, Yalley et al. (2016) revealed that these two GRF proteins interdependently bind and regulated the expression of IGFBP1. In addition, Yalley et al. (2016) also revealed that changes in binding affinity result in alterations in chromatin structure and reduction in the acetylation of H3K27 histone mark. We found three missense SNPs, from which one causes a mutation (rs34733279) that modifies the amino acid residue from aspartic acid to asparagine. This nucleotide variant found is present just in CHB (allele frequency 0.11). It is likely that such missense change produces a functional changes in the regulatory mechanisms of FOXO1, for instance, by decreasing or increasing the affinity in which FOXO1 interact with chromatin, DNA, with other pioneer GRFs or cofactors to fine tune insulin pathways regulation. Single nucleotide variations in *FOXO1* have also been associated with an ectatic disease of the cornea "keratoconus" in a Saudi Arabian population (Abu-Amero et al. 2015).

To sum up, based on the signatures of positive selection we detected for these two genes, we suggest that some of the nonsynonymous variation we observed in *ZNF407* and *FOXO1* are introducing subtle changes in the gene regulatory activity of these two genes. We also think it is likely that such variation may have differentially contributed into the regulatory pathways of insulin and glucose of these three human populations, and thus, into human adaptation.

3.3.2. EHH haplotypes in KRAB-ZNF gene clusters suggest selection on specific traits that swept in AMH populations

The occurrence of long extended regions exhibiting low variability and recombination rates are normally considered as candidate regions for selective sweeps (Vitti et al. 2013). Three out of 25 KRAB-ZNF gene clusters exhibited EHH regions larger than 100 kb for XP-EHH and low recombination rates, which suggest

the occurrence of selective sweeps (Sabeti et al. 2007; Vitti et al. 2013). The additional results obtained from the F_{st}, a test based on population genetic differentiation (Weir and Cockerham 1984), also suggests that these EHH regions carry highly differentiated SNPs between CHB and CEU populations.

Multiple genome wide scans for positive selection in humans available in published literature (Sabeti et al. 2007; Pickrell et al. 2009; Metspalu et al. 2011; Grossman et al. 2013) resulted in only a few overlapping regions (Perdomo-Sabogal et al. 2014). Out of all the genes located within the three KRAB-ZNF clusters we identified with EHH, two genes have been previously reported as candidate regions for selective sweeps, *ZNF501* (cluster three) (Grossman et al. 2013) and *VKORC1* (non-GRF in cluster 14) (Ross et al. 2010; Patillon et al. 2012). By using the CMS test, Grossman et al. (2013) identified one high scoring candidate SNP (rs2257995) for positive selection located in the promoter region of the gene *ZNF501* (179 bp from the start of the gene). By using a similar complementary analytic strategy to the one we implemented here, Patillon et al. (2012) suggested the gene *VKORC1* was located in a region exhibiting EHH in all Asian populations, as was initially put forward by Ross et al. (2010).

Despite that the overlap between the genes located in the other two KRAB-ZNF clusters (one and three) and the genes previously reported in literature was small, the results we obtained from CLR, XP-CLR and F_{st} tests also showed regions with higher rank scores (above the threshold) within the EHH regions located in the three KRAB-ZNF clusters. Taken together, the results from these three different tests for detecting positive selection in AMH populations, added to the patterns of variation observed within these regions (increased haplotype homozygosity, high genetic differentiation, and number of nonsynonymous SNPs with likely functional effect), we suggest these three KRAB-ZNF gene cluster have undergone a selective sweep at least in one human population (CHB).

The demographic processes that humans have experienced during the migration out of Africa have influenced their genetic diversity in many different ways. For instance, migration, changes in population size, bottlenecks, interbreeding with

archaic humans, are just a handful of examples of how demography could have shaped human variability (Pybus et al. 2015). It is likely that such demographic events mimic the signatures that selection could have left on the genome of different populations, thus leading to misinterpretations (Nielsen et al. 2009). Using the analyses recently published by Pybus et al. (2015), who implemented a hierarchical boosting algorithm for refining the detection of signatures of selective sweeps on three human populations, we were able to cross-validate the signatures of selective sweeps observed in the three KRAB-ZNF gene clusters one, three and 14. In addition, these results also allowed us to establish the type and approximated age of the sweeps observed here. It is likely that the clusters one and 14 experienced an incomplete ancient selective sweep in CHB, which probably occurred between 45 to 30 kilo years ago (kya), while the cluster three may be the result of a recent and rapid complete selective sweep occurred between 25 to 10 kya as well in CHB. It is important to highlight that calculating a more precise dating requires more sophisticated computational approaches (Pybus et al. 2015). Considering that a majority of the nonsynonymous SNPs located within these regions differ from their ancestral states, we suggest that the EHH observed in these three KRAB-ZNF clusters may be the result of selective sweeps from de novo mutations, instead of sweeps from extant genetic variation (as described in the introductory chapter).

As a final observation, one of the EHH haplotypes seems to extend farther within the KRAB-ZNF gene cluster 14; however, the presence of one SNP in an intronic region of a non-GRF gene causes the EHH decay. Detailed exploration of the surrounding SNPs evidenced that this lower peak could be explained by the way the MAF was defined in one of the populations. Thus, it is likely that the EHH region from this KRAB-ZNF cluster spans 261 kb and not only 156 kb as we initially thought. Patillon et al. (2012) suggested that the genomic region exhibiting signatures of selective sweep in all East Asian populations spans around 505 kb; however, our analyses for EHH showed that the haplotype decays around the 251 kb up and downstream the EHH region in CHB. Based on the genes that displayed most extreme XP-EHH scores, Patillon et al. (2012) also suggested that the most likely targets for selection in this region were the adjacent genes *VKORC1*, *BCKDK*, *MYST1*, and *PRSS8*. None of them is being catalogued as GRF gene. However,

genome-wide association study on dependency of phenprocoumon, a long-acting oral anticoagulant drug, indicated that the genes *STX4A* and *ZNF646* carry variants that displayed strongest association with phenprocoumon maintenance dosage variation in European individuals from Rotterdam (Teichert et al. 2011). Our results showed that despite *ZNF646* does not carry the SNPs with most extreme values for the tests implemented here, it does explain 51% of the nonsynonymous SNPs (14 variants) exhibiting higher scores in this EHH region (KRAB-ZNF cluster 14) for CHB. Therefore, it is plausible that *ZNF646* has also introduced regulatory diversity in the Asian regulome, for instance, by regulating pathways implicated in blood coagulation. In addition, *ZNF646*, together with another five genes located in the 251 kb long EHH region, four GRFs (*KAT8*, *ZNF646*, *ZNF668*, *FBXL19*) and two non-GRFs (*STX1B* and *VKORC1*) have recently been associated with obesity (Locke et al. 2015; Yazdi et al. 2015). One SNP located on *KAT8*, a gene that codes for a GRF protein, has been associated with body mass index in European individuals. In addition, the association and expression quantitative trait loci (eQTL) data suggest that this SNP also affects the gene expression of *ZNF646*, *VKORC1* and *ZNF668* (Locke et al. 2015). Consequently, we suggest it is likely that the selective sweep that occurred in this genomic region where this KRAB-ZNF gene cluster is located have probably influenced the regulatory pathways associated with both phenotypical conditions, obesity and anticoagulant response. In special, considering that four genes in this region are GRFs.

3.3.3. Positive selection of C2H2 genes as a potential source for regulatory diversity

C2H2 genes, including KRAB-ZNF genes, have experienced independent expansions in primates (Nowick et al. 2011; Najafabadi et al. 2015). Proteins of this DNA-binding transcription factor class typically contain modular Cys2-His2-ZNF domains joined together in tandem arrays (Huntley et al. 2006). Each C2H2 zinc finger contacts three or more nucleotides. Four amino acids at positions -1, 2, 3 and 6 counted relatively to the alpha helix provide, to a large extend, the sequence specificity for the C2H2 zinc fingers (Wolfe et al. 2000). The C2H2 motif folds into a ββα structure where two highly conserved cysteine and histidine residues provide the stability to the fold (Lee et al. 1989; Brayer et al. 2008). This finger-like structure

is essential for binding to the DNA; therefore, changes in these structures are likely to be a source of human regulatory diversity. Out of the 121 C2H2 GRF genes (among them 53 KRAB-ZNF genes) that we found in regions exhibiting signatures of positive selection, just nine have been previously reported as candidates (Grossman et al. 2013). By exploring nonsynonymous changes occurring in the cystidine or histidine residues of KRAB zinc fingers within KRAB-ZNF clusters exhibiting EHH, we detected three nucleotide changes that may alter the functionality or specificity of the zinc fingers of two KRAB-ZNF genes (*ZNF501* and *ZNF502*) in the Asian population (CHB). Non-synonymous SNPs on the fifth and sixth zinc finger of the genes *ZNF502* and *ZNF501*, respectively, change the Cysteine of the C2H2 structure. Protein (Intensity-based absolute quantification, iBAQ) and mRNA expression (Fragments Per Kilobase of transcript per Million mapped reads, FPKM) data for *ZNF502* and *ZNF501* suggest that these two GRFs are highly expressed in testis (iBAQ and mRNA), ovary (mRNA) and Lymphoid/Immune cells/system (*ZNF501)* (Frézal 1998; Wilhelm et al. 2014). Previous studies found that sequence differences at the positions coding for the DNA-binding amino acids of PRDM9, a zinc finger protein, seem to be connected to male sterility risk in Asians (Miyamoto et al. 2008). Despite additional information about the functional roles of *ZNF502* and *ZNF501* genes is not known yet, we speculate that the SNPs here described for *ZNF502* and *ZNF501* may be connected with regulatory differences at reproductive (fertility) and immune system level among human populations.

We additionally detected nonsynonymous changes in the linkers of three KRAB-ZNF genes *ZNF35*, *ZNF646*, *ZNF695*. KRAB-ZNF genes, together with the rest of C2H2 genes, are rapidly evolving within mammals (Tadepally et al. 2008). This rapid evolution has not been restricted to mechanisms such as segmental and partial duplications, rearrangements and accretion of C2H2 zinc finger motifs (Tadepally et al. 2008; Nowick et al. 2010; Najafabadi et al. 2015). Variation and diversification of the amino acid residues that are also essential for the folding of the DNA binding domains and regulatory activity, could also have an effect in the regulatory activity. For instance, amino acid substitutions in the linker region have been previously found to strikingly affect DNA specificity and affinity in one GRF, the general transcription factor IIIA (Ryan and Darby 1998), thus probably indicating the

regulatory effects of amino acid changes in such regions. Another recent example of amino acids changes affecting the affinity of zinc fingers is the c.C5054G/p.S1685W mutation, which affects 2 of the 3 ZNF407 isoforms (Kambouris et al. 2014). This Serine to Tryptophane mutation is located in the peptide that links the fingers C2H2 types 18 and 19, causing cognitive impairment in the individual that carry it (Kambouris et al. 2014). *ZNF35*, *ZNF646*, *ZNF695* harbor nonsynonymous SNPs in the linker regions of several zinc fingers. *ZNF35* gene has been found as one of two non-KRAB-ZNF genes that can bind and regulate specific classes of EREs (Najafabadi et al. 2015). ERE families contain the hominid-specific retrotransposon SINE-VNTR-Alu (SVA) family (Wang et al. 2005), endogenous retroviruses (ERVs), and long interspersed nuclear elements (LINEs). It has been suggested that variation in the tandem array of zinc finger domains may results in a diversifying mechanisms to down regulate the expression of newly evolved ERVs in humans (Lukic et al. 2014). Therefore, considering that changes in the residues located in the linker regions of particular zinc fingers may also introduce a source of population-specific mechanism, we think the changes found in *ZNF35*, *ZNF646* and *ZNF695* may be introducing regulatory mechanisms to repress ERVs' expression.

Finally, considering the multiple missense mutations occurring in the genes *ZNF695* and *ZNF646*, six and 14 nonsynonymous SNPs respectively, and that these genes are located in a region with strong signatures of positive selection (KRAB-ZNF clusters one and 14) in CHB, we also suggest that these genes have experienced rapid evolution at least in this human population.

3.4. Conclusions

Using the most recent catalog for GRFs, the information from the 1000 genomes project, and data obtained for multiple tests for detecting positive selection in humans, we identified a group of candidate GRF genes that might have undergone positive selection in three human populations. These results present several scenarios where multiple classes of GRFs may have contributed to the regulatory diversity and adaptation of humans. At least six of the larger GRF classes are enriched for regions exhibiting signatures of positive selection in humans. Further inspection of single nucleotide variants in several GRF genes suggest how genetic

variation could be introducing regulatory diversity in humans, thus possibly leading to the evolution of particular traits associated with body mass index, blood coagulation, reproduction, and insulin/glucose regulatory pathways in in humans.

3.5. Methods

3.5.1. Identifying candidate GRFs for selection in three AMHs populations

Using whole genome sequencing data from 1000G project (1000 Genomes Project Consortium 2012) and the data from 1000 Genomes Selection Browser 1.0 (Pybus et al. 2013), we extensively looked for candidate genomic regions that exhibit signatures of positive selection, and where GRFs are placed, for three AMHs populations: Utah Residents with Northern and Western Ancestry (CEU), Han Chinese in Bejing (CHB), and Yoruba in Ibadan (YRI).

To better understand how natural selection may have shaped the diversity we currently observe for GRF genes in these three different human populations, we first analyzed the results obtained from four different statistical population genetics methods, which were initially implemented by Pybus et al. (2013). Based on our particular interest, we selected three tests that implement different strategies for detecting signatures of positive selection in human, CLR, XP-CLR and XP-EHH, as previously described in the introductory section of this chapter. As additional strategy to complement our findings, we also included the F_{st} statistics; a test is of utility for estimating population genetic differentiation. By analyzing the results obtained from multiple methods, we expected to better describe how positive selection may have influenced the genetic variation of GRF genes. The resulting data covers around 83% of the GRF genes we cataloged, while for the remnant 17% there was not information available.

One of our main interests was to identify GRFs located in regions that have significantly higher scores compared to the rest of the genomic distribution for all the four tests analyzed here (CLR, XP-CLR, XP-EHH and F_{st}) (Table 3). Considering that scores found in the upper tail of the distribution of all these four tests indicate deviations from neutrality, and suggests regions/genes that are likely candidates

for being under positive selection, we explored GRFs having scores in the high 5% upper tail of the distribution. By using ranked score values, raw scores ranked based in the genome-wide distribution of scores obtained for each population (Pybus et al. 2013), we expected to identify those GRF genes with higher rank score for each test. Considering that a rank score of 1.3 will correspond to the 5% of the upper tail of the distribution, we used this value as threshold. We considered all genes with scores larger than this threshold as "GRF candidates for positive selection".

3.5.2. GRF overrepresentation

We first evaluated if GRFs are enriched among the top 5% of all set of human genes for the four tests on the CEU, CHB and YRI populations. From the whole set of human genes, we selected all the rank scores corresponding to GRF genes. To do this, we used the catalog of GRF (Chapter I), thus making possible to generate two set of genes. We then performed a Wilcoxon-rank-test to test if GRF genes presented more extreme rank scores than for the rest of the human genes (non-GRFs) in all four tests. Subsequently, we performed permutation test based on the Wilcoxon rank-test (1000 permutations) to evaluate if by assigning the same size of GRFs genes to random data the number of scores being on the 5% upper tail of the distribution were exchangeable under the null hypothesis of no difference between our initial observation and the sampled data.

3.5.3. Recombination rates difference quantification

To evaluated if there the recombination rates of GRF genes significantly differ from those found for non-GRF genes, we used the standardized recombination maps and rates from deCODE for the human genome reference GRCh37/hg19 (Masson et al. 2010). We assigned the recombination rates to each gene within the two groups, GRFs and non-GRF genes. Then, we quantified the distance between these two empirical distributions by implementing a two-sample Kolmogorov–Smirnov test (KS).

3.5.4. GRFs distribution of length

Gene length can influence the number of GRF regions that can be present in the upper tail of the distribution. Knowing that this variation in length is important to

properly identify true biological signatures, instead of rather than length-dependent artifacts, we assessed if the distribution of the gene lengths between GRFs and non-GRFs was significantly different. If the distribution of lengths of GRF and non-GRF genes were by chance the same, it would imply that sampling GRF genes from the upper tail would not be biased towards the distribution of their lengths. We performed a Spearman's rank correlation test to measure it there was statistical dependence between gene length and the rank scores obtained for each of the four tests.

3.5.5. Details of XP-EHH, CLR and XP-CLR calculation

3.5.5.1. XP-EHH method (Sabeti et al. 2007)

For computing XP-EHH, the initial step requires calculating EHH for each population. Therefore, for a bi-allelic SNP with alleles b and B, the EHH is computed as follows:

$$EHH(x) = \frac{\sum_{i=1}^{h_x}\binom{n_i}{2}}{\binom{n_b}{2}+\binom{n_B}{2}} \qquad (1)$$

In equation (1) one $\mathbf{n_b}$ and $\mathbf{n_B}$ represent the amount of haplotypes with alleles b and B respectively; $\mathbf{n_i}$ is the sum of the $\mathbf{i^{th}}$ haplotype in a particular population and $\mathbf{h_x}$ correspond to the number of different haplotypes in the core region up to a distance **x** from the locus. The unstandardized XP-EHH methods is then defined as:

$$unstandirized\ XP - EHH = \log\left(\frac{\int_D EHH_{pop1}(x)dx}{\int_D EHH_{pop2}(x)dx}\right) \qquad (2)$$

In equation (2), *pop1* and *pop2* correspond to the two populations, the integration values *D* corresponds to the cutoff over the *x* integration and is used for determining when EHH has decayed to sufficiently small values. The raw XP-EHH scores from Eq. (2) are then standard normalized and a p-value cut off is obtained. This also requires correcting for multiple testing. Taking into account that XP-EHH test is not sensitive to allele frequencies, the data does not required to be fit into frequency bins before calculating the significance.

3.5.5.2. CLR method (Nielsen et al. 2005)

Considering that the (unknown) probability of detecting a derived allele which frequency is *j* in the sample be p_j, j=1, 2, …, n – 1, assuming independence genome wide Pr(X_i=*j*)= p_j.. Let **p** = (p_1, p_2, …, p_{n-1}). For k SNPs, the composite likelihood function is formed by combining (multiplying) the sampling probabilities across the chromosome:

$$CL_1(\boldsymbol{p}) \equiv \prod_{i=1}^{k} p_{x_i} = \prod_{j=1}^{n-1} p_j^{k_j} \qquad (3)$$

in equation (3), k_j is the number of SNPs with derived allele frequency j in the sample. The maximum composite of **p** is then given by $\widehat{p} = \frac{k_j}{k}, j = 1, 2, \dots, n-1$.

3.5.5.3. XP-CLR method (Chen et al. 2010)

To reduce the effect s that SNP ascertainment bias could introduce when detecting candidate regions for positive selection, Chen et al. (2010) built up the XP-CLR method based on the multiplelocus composite likelihood ratio method previously described by Nielsen et al. (2005). The XP-CLR likelihood is given by:

$$CL(r, w, s) = \prod_{i=1}^{k} \int_0^1 f(p_1^i | p_2^i, w, s, r^i) \binom{n}{m_i} (p_1^i)^{m_i} \times (1 - p_1^i)^{n-m_i} \, dp_1^i \qquad (4)$$

in equation (4), $\boldsymbol{r}$ corresponds to the recombination rate: $\{r^1, r^2, \cdots, r^k\}$, $\boldsymbol{n}$ is the sample size, m_i corresponds to the counts of neutral alleles at locus i, s is the selection coefficient, k is the size defined for the sliding window, w is defined as the weight factor which is based on information about the linkage disequilibrium and p represents the allele frequency.

Chapter 3

Human lineage-specific transcriptional regulation through GA binding protein transcription factor alpha (GABPa)

4.1. Introduction

Compared with other primate species, humans exhibit distinctive physical and behavioral characteristics, such a full striding bipedalism, more specialized limbs and cerebral cortex, a brain increased in size, and declarative language. Despite it has been long suggested that changes in the ways gene expression is controlled play essential roles in the evolution of such morphological traits (Wray 2007), the majority of the regulatory elements that control human-specific traits still remain unidentified.

GRF proteins directly and indirectly interact with DNA to mediate gene transcriptional control (Iyengar and Farnham 2011; Perdomo-Sabogal et al. 2014). A particular group of GRFs, DNA-binding transcription factors, recognize and directly interact with specific DNA sequences to regulate gene expression. These specific DNA sequences, as well known as transcription factor binding sites or CREs, are normally around 10 base pairs (bp) long in eukaryotes, even though their length can range from five to 30 bp (Galas and Schmitz 1978; Stewart et al. 2012). Sequence variations in CREs can considerable alter the binding specificity and recognition of particular TFs and hence, switch on/off the transcriptional regulation of a particular target gene (Lin et al. 2007). Therefore, genetic changes altering the configuration of CREs result in new mechanisms to control gene transcriptional regulation, and thus, they are expected to greatly contribute on species phenotypical differentiation and evolutionary history of lineage-specific traits.

A handful of comparative genomics analyses, mainly focused on particular genes, have aimed to identify and characterize human- and primate-specific changes in CREs (Huby et al. 2001; Rockman et al. 2005; Romanelli et al. 2009). Nonetheless, changes in the activity of a myriad of CREs during human evolution still remain to be functionally characterized, and most importantly, experimentally validated by exploring different molecular and computational approaches.

By integrating the results of different experimental and computational methods such as chromatin immunoprecipitation followed by high-throughput sequencing ChIP-seq, comparative genomics, reporter gene assays, RNA interference siRNA, and gene ontology analyses, we explored the evolution and functionality of GABPa CREs in humans. GABP is a member of the E-twenty six (ETS) transcription factors group class and regulates many genes that have been associated with cell migration and differentiation, cell cycle control and fate, hormonal regulation and apoptosis. This protein consists of two subunits, alpha (GABPa) and beta (GABPB1). GABPa carries the ETS domain, while GABPB1 harbors the transcriptional activation domain (Figure 15) (Ripperger et al. 2015). Together with GABPB1, GABPa forms a complex capable of binding the core consensus motif GGAA to regulate gene expression (Batchelor et al. 1998).

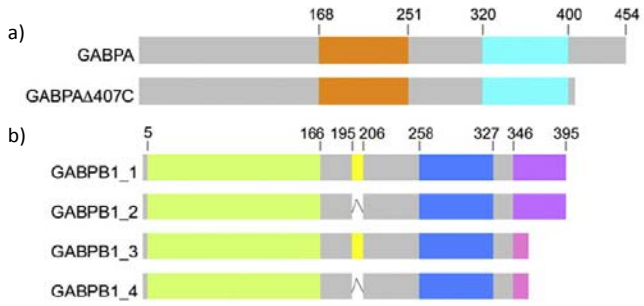


Figure 15. Representation of GABP subunits corresponding to GABPA and GABPB1. **(a)** GABP subunit alpha (GABPA), the colored boxes depict the pointed (orange, 168–251 aa) and ETS domains (cyan, 320–400 aa). **(b)** GABP subunit beta, the colored boxes show the ankyrin repeats (green, 5–34, 37–66, 70–99, 103–132, and 136–166), the 12-aa-isoform-1-and-3-specific insert (yellow, 195–206 aa), the transcriptional activation domain (blue, 258–327 aa), and several C-termini of the long and short isoforms (violet/pink, N346 aa). Modified from (Ripperger et al. 2015).

As transcriptional regulator, GABPa has also been associated with neuromuscular function (Rosmarin et al. 2004; Yang et al. 2011), and coordinating the expression of many cytochrome c oxidase genes (COX) (Guo et al. 2000). GABPa is also involved

in mitochondrial synthesis and biogenesis (Yang et al. 2014) in early stages of development (Ristevski et al. 2004), and its silencing of GABPa has been proved to induce premature embryonic lethality in rodents (Ristevski et al. 2004; Jaworski et al. 2007).

GABPa also has the ability to bind promoters harboring bi-directional gene pairs coded on opposite strands (Collins et al. 2007), thus regulating the expression of two downstream genes (Lin et al. 2007). Importantly, GABPa protein and its binding domain are highly conserved in primate species and other mammals (Figure 16), suggesting it is possible that GABPa binds the same core motif in many of these species. Therefore, it would be possible to identify genetic changes occurring on CREs of GABPa by using comparative genomics, making possible to explore CREs evolutionary changes along the lineages that lead to humans.

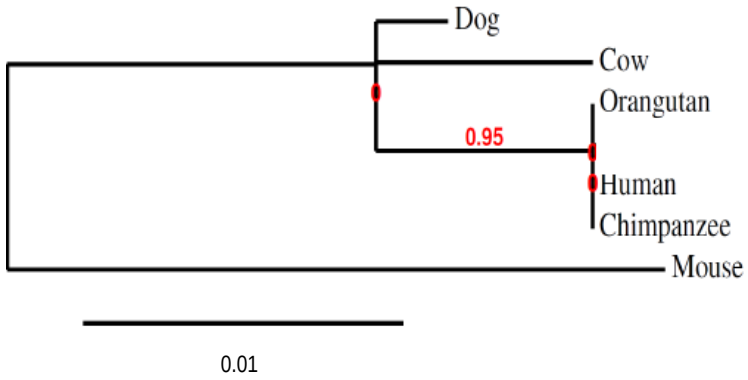


Figure 16. Genetic distances tree showing the protein conservation level of GABPa in different mammalian clades. The tree was produced using BLAST pairwise alignments. The evolutionary distance between pairs of sequences was calculated using Kimura's model of neutral evolution (Kimura 1968). Fast Minimum Evolution algorithm (Desper and Gascuel 2004) was used to model the distances between the sequences.

Our methodological design integrated three experimental (ChIP-Seq, siRNA and reporter gene assays) and two computational (comparative genomics by using multiple sequence alignment and gene ontology analysis) approaches (Figure 17). By using data generated from genome-wide assays of protein-DNA interaction in human embryonic kidney cells (HEK293T), we identified a set of putative GABPa target genes. Additionally, using data derived from GABPa knock-down experiments in HEK293T cells, we independently confirmed the functionality of a representative number of the CREs that were initially detected. To identify GABPa CREs with

human and primate specificity, we also performed ancestral reconstruction of the GABPa CREs using multiple sequence alignment. Finally, performed reporter gene assays in human and non-humans primate cells. The regulatory activity of the ancestral states of promoters missing the core consensus sequence of GABPa was measured by introducing nucleotide changes in human/hominid-specific GABPa binding sites. Taken together the results from our four previously mentioned experiments, we performed gene ontology enrichment analyses to functionally characterize those genes that are putative targets of GABPa.

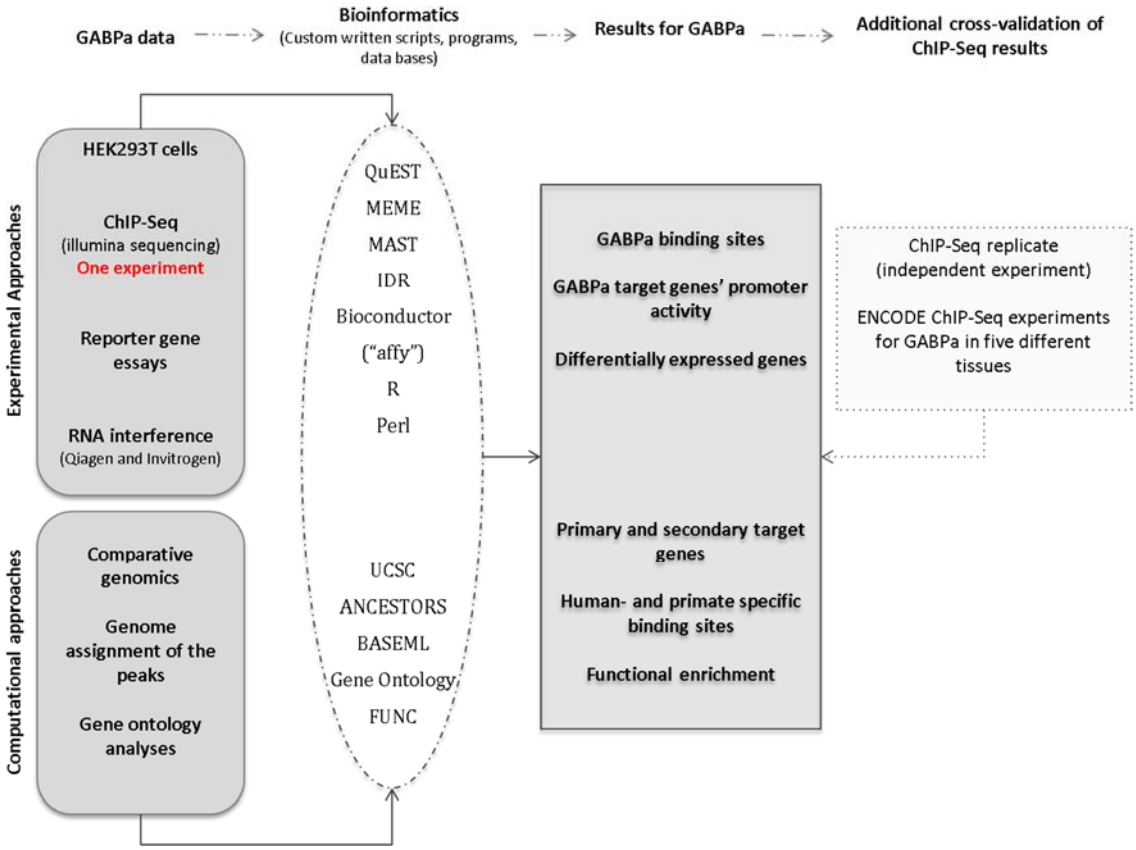


Figure 17. Schematic representation of the multiple strategies implemented here for detecting human-specific GABPa CREs. (Left) main experimental and computational strategies defined as starting points. (Middle) shows the main bioinformatics tools, packages and databases that were used for data processing and analyzing. (Right) Main findings and scientific contributions generated in this research project.

4.2. Results

4.2.1. Identification of GABPa binding sites by chromatin immunoprecipitation

To identify where GABPa binds in the human genome, we performed ChIP-Seq experiments with a GABPa-specific antibody in HEK293T cells. The efficiency of this antibody was previously validated by proving its ability to detect highly specific GABPa peaks (Valouev et al. 2008). Our initial ChIP-Seq experiments only included one ChIP-Seq experiment, which produced a set of 6,208 putative GABPa ChIP-Seq peaks (Supplementary Table S8, supplementary data file) (see Methods). This experiment was posteriorly validated by using information from a second ChIP-Seq experiment and additional comparisons with ENCODE reported datasets for the same protein in different cell lines (see below).

To identify the GABPa consensus core motif, we selected 200 bp regions up and downstream the center of the ChIP-Seq the peak as input for the *de novo* motif discovery algorithm available in the Multiple EM for Motif Elicitation (MEME) suite (Bailey et al. 2009). We generated an 11 bp consensus binding sequence and the position-specific weight matrix (PWM) for GABPa (Figure 18). The majority of the sites contributing to the PWM (93%) were located adjacent to the peaks' center (Figure 19a; Supplementary Table S9, supplementary data file), thus indicating consistency in our peak calling results and giving validity to our ChIP-Seq results. Taken together, the binding sites and motif identified here seem to be reliable.

Figure 18. Sequence logos representing the GABPa PWM generated HEK293T.

The algorithm implemented in the MEME suite, when the default options are used, considers that each peak contains zero or one sequence motif. Under this notion, this particularity results favorable for detecting non-repetitive motif elements. However, as it is possible that more than one GABPa motif is located within the same peak region (Yu et al. 1997) (Figure 19b), we searched for additional binding sites by using the motif sequence alignment and the Motif Alignment and Search Tool (MAST) (Bailey et al. 2009)(Figure 19c). As result, we obtained 11,619 PWM hits in 5,797 peak regions of 200 bp in length. The majority of the peaks contained between one and two binding sites (Figure 19d, Supplementary Table S10, supplementary data file).

By using the gene annotation from the UCSC genome browser and the genomic coordinates of the 6,208 peaks for GABPa, we were able to identify 4,277 (69%) peaks located within 300 bp up- and downstream of the transcriptional start sites (TSSs). These TSS correspond to 11,848 transcripts from 3,994 putative target genes (Entrez IDs). The number of peaks mapping in the neighborhood of genes increased to 5321 (86%), when we extended the window to ± 5 kb centered to the TSSs. These TSS correspond to 15,046 transcripts from 5,218 putative targets (Supplementary Table S11, supplementary data file). If we further extended the window to ± 10 kb, the number of peaks mapping in the proximity of genes 5,465 peaks (88%) mapping to 18,730 transcripts and corresponding to 5,784 putative genes. Thus, the majority of peaks reside close to the TSS (Figure 19d). For downstream analyses we used mappings within ± 5 kb centered on UCSC-annotated TSSs.

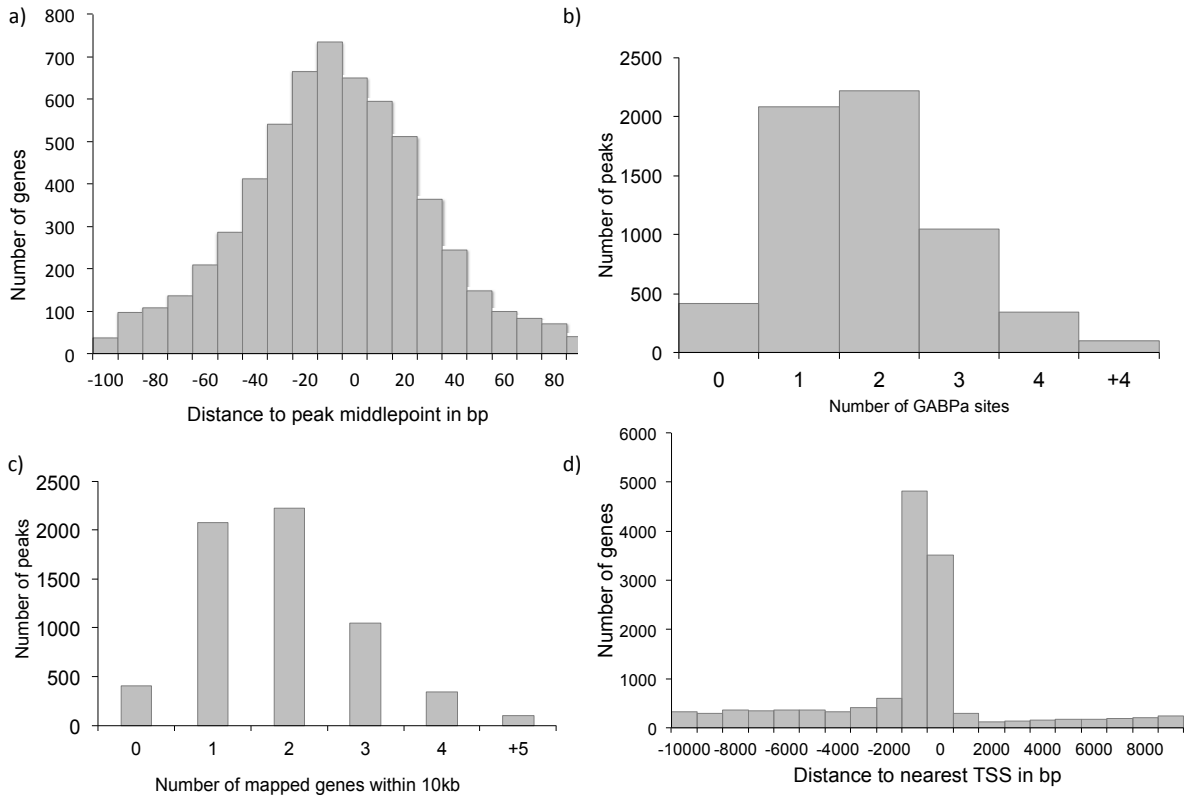


Figure 19. GABPa binding sites are located in the proximity of the peak centers and the TSS sites. **(a)** Number and distance to the peak center of the sites contributing to the MEME motif discovery (6,031 of 6,208 in total). **(b)** GABPa motif distribution per each ChIP-Seq peak. **(c)** Motif presence distribution in a window of 200 bp around the ChIP peak centers. **(d)** Peak call distances from the closest TSS of UCSC genes in a window of 10 kb centered to the TSS. TSS distance given is base pairs. Negative values represent upstream, positive values downstream regions.

To further validate these results, we compared our results with a second ChIP-Seq experiment for GABPa in HEK293T cells. Overlaps between the peak calling results from both ChIP-Seq experiments showed that 91% of the peaks (5,647 peaks) were present in both (Figure 20a). Following the guidelines suggested by the ENCODE consortium, we also calculated the irreproducible discovery rate (IDR) for both replicates (Landt et al. 2012), finding that 3,677 peaks (~60%) overlapped at an IDR <0.05 (Figure 20b), thus demonstrating reasonable consistency among the two replicates.

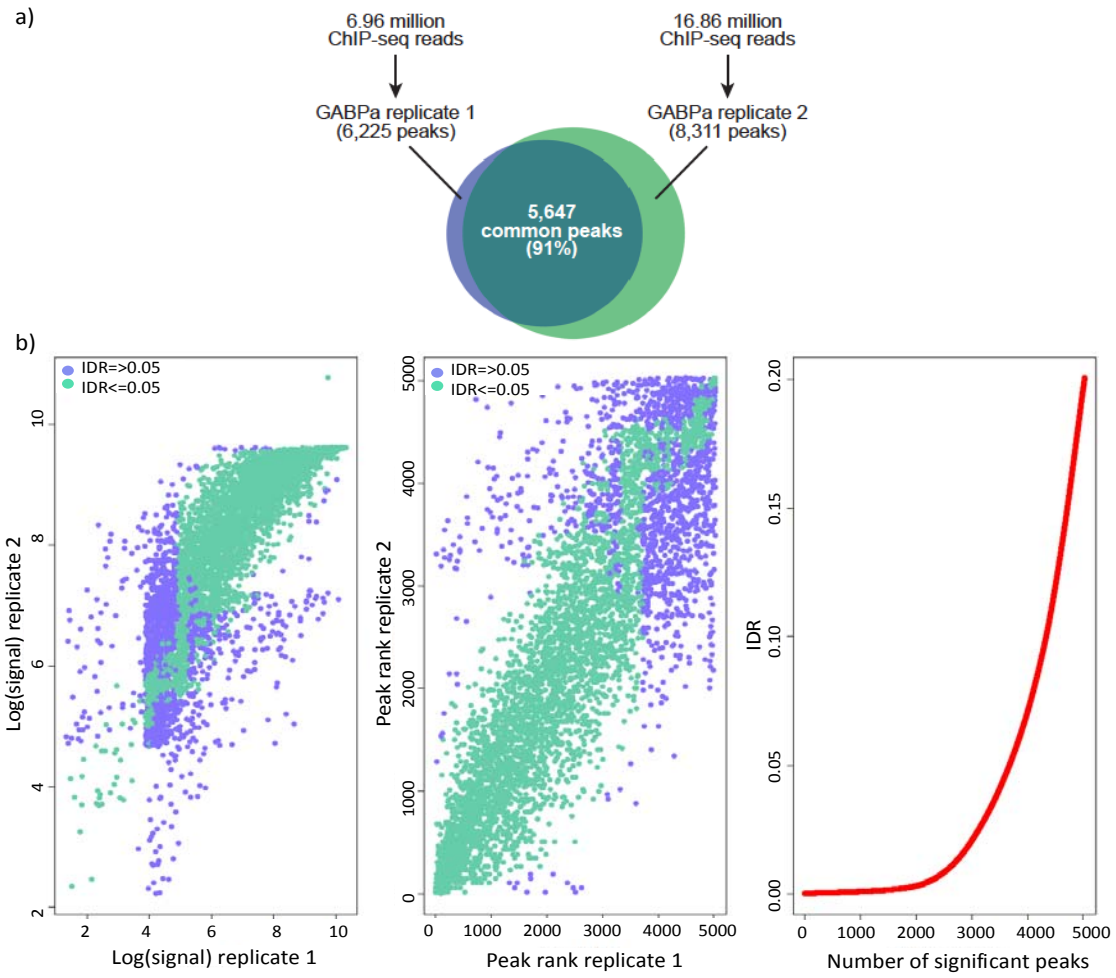


Figure 20. Cross validation of our initial peak call by comparing it with a second ChIP-Seq experiment for GABPa. **(a)** Overlap between the peaks called for two GABPa ChIP-Seq replicate experiments. Out of the 6,288 binding peaks obtained for the Replicate 1 (~7 million reads) by using QuEST 2.4, 91% (5,647 binding peaks) overlapped with the peaks obtained for the second ChIP-Seq replicate (~17 million reads, 8,311 binding peaks). **(b)**. Left panel, IDR scatter plots of the signal scores. Middle panel, IDR scatter plots showing the ranks of the peaks that overlap in both replicates. Right panel corresponds to the calculated IDR as function of different rank thresholds. Light green dots show the pairs of peaks that passed an IDR threshold of 5 percent,

In addition, we assessed if the peaks we found for GABPa in HEK293T cells were also detected for the same protein in another five human cell lines: H1 human embryonic stem cells (H1hesc), cervical ectoderm (HeLa-S3), Lymphoblastoid (GM12878), liver (HepG2) and Myelogenous leukemia (K562) (Wang et al. 2013). This overlap corroborated that at least 3645 (~60%) of GABPa peaks we reported here were also detected in other human cell lines from different tissues (Wang et al. 2013) (Figure 21). We additionally verified that the PWM generated for GABPa here

was highly consistent with the ones reported in the ENCODE datasets for the aforementioned mentioned cell lines (Figure 22). In addition, we also designed and implemented GABPa knock-down experiments and reporter gene assays in the same cell line (see below). Taken together, the ChIP-Seq experiments, GABPa binding sites and motif identified here seem to be reliable for performing down stream analyses.

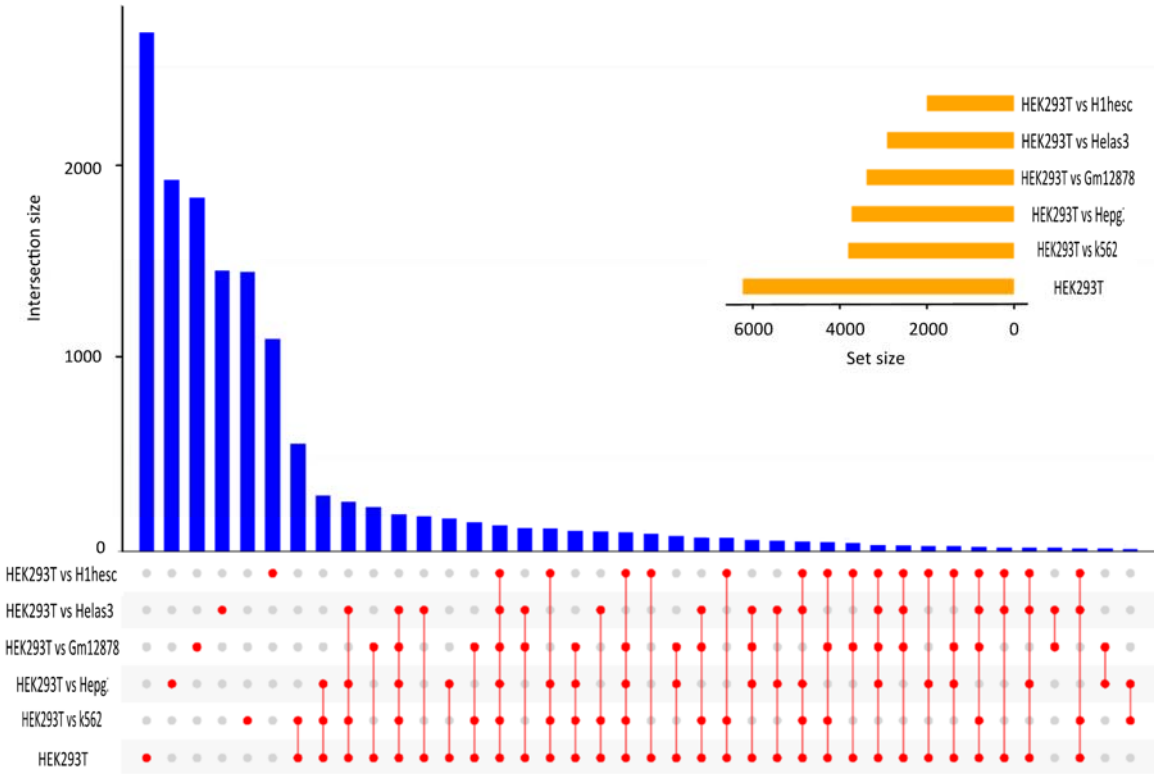


Figure 21. Overlap between the GABPa peaks reported in this study versus the data available from the ENCODE project for the same protein in different cell lines. Cell lines: H1 human embryonic stem cells (H1hesc), cervical ectoderm (HeLa-S3), Lymphoblastoid (GM12878), liver (HepG2) and Myelogenous leukemia (K562) (Wang et al. 2013). Set size was defined using the number of peaks in our dataset (6208) and the overlap of peaks between ours and the other cell lines.

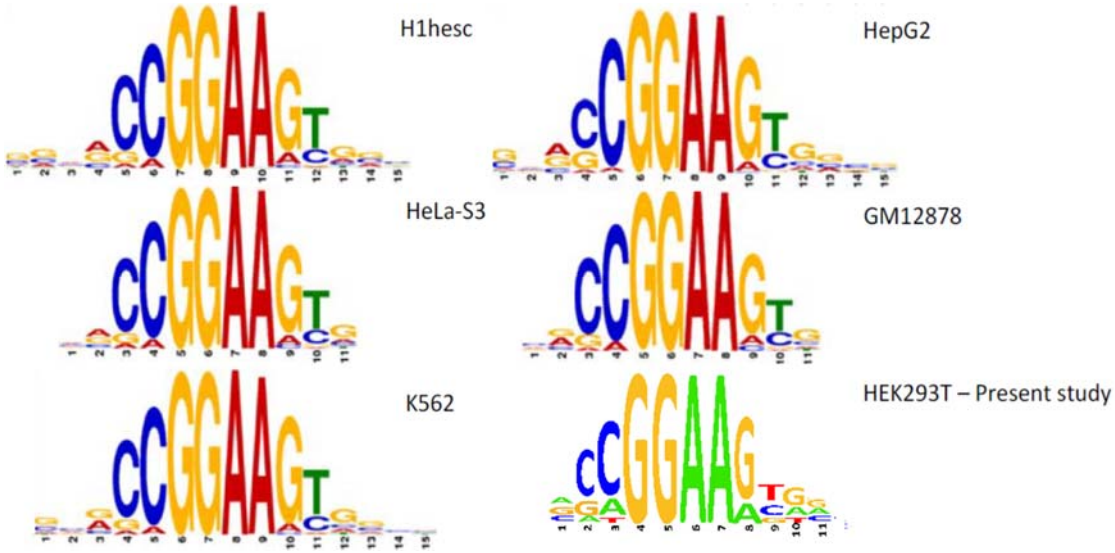


Figure 22. Comparison of GABPs motifs identified de novo from different cell lines. All consensus core sequences were identified using the MEME-ChIP suite of tools. Sequence logos represent different PWM. Cell lines: H1 human embryonic stem cells (H1hesc), cervical ectoderm (HeLa-S3), Lymphoblastoid (GM12878), liver (HepG2) and Myelogenous leukemia (K562).

4.2.2. Identification of newly evolved GABPa binding sites

By combining the information from the 11,619 GABPa binding sites and the multiple sequence alignments (MultiZ 44) available in UCSC, we extracted 11,008 alignments corresponding to the CREs found within the peaks for GABPa (see methods). For the remaining 611 genomic locations we either detected that the sequence had gaps or that there was no sequence alignment available. To specifically identify GABPa CREs that are human-specific, we first sourced eight available genomes for non-human primates. As part of our investigative interests, we also analyzed three additional clades to detect hominini- (Human and Chimpanzee), Homininae- (Homini and gorilla) and Hominidae- (Homininae and Orangutan) specific GABPa CREs. We obtained hominid ancestral sequences for 10,943 binding sites. The remaining sequences were miss-aligned (Supplementary Table S12, supplementary data file). We then proceeded to identify GABPa binding sites within these ancestral sequences (Supplementary Table S13, supplementary data file).

Out of the sequence alignment reconstruction we performed here, we discovered 224 GABPa binding sites with human specificity. By using the coordinates from the ChIP-Seq peaks, we identified that these 224 GABP binding sites contained 219 peaks located in the promoter region (± 5 kb centered on TSSs) of 217 genes

(Supplementary Table S14, supplementary data file). We additionally detected that three of these genes are human-specific (*CEP170, RPL41* and *GUSBP4*), while other three are Hominidae-specific (*STAG3L4, USP6* and *ZNF383*) (Zhang et al. 2010). Using UCSC genome browser, we manually inspected the genomic regions of those peaks containing binding sites and that did not map to know genes. One of them is located 317 nucleotides upstream of the transfer RNA Phe (anticodon GAA) gene (uc021qjx.1), while the other one is 3,810 nucleotides upstream of a human cDNA (uc021suf.1). Among the 217 promoters that gained human-specific GABPa binding sites, we detected 12 KRAB-ZNF transcription factors. Subsequent test revealed that KRAB-ZNFs are indeed significantly overrepresented among genes with human-specific GABPa binding sites in their promoters (p-value= 0.01, Fisher's exact test). Using the ancestral reconstruction results, we identified 57 Hominini-, 244 Homininae- and 310 and Hominid-specific binding sites (Supplementary Table S15, supplementary data file) mapping to 44, 240 and 326 genes respectively (Supplementary Table S16, supplementary data file). Binding site appearances for all ancestral branches leading to human are shown in Figure 23.

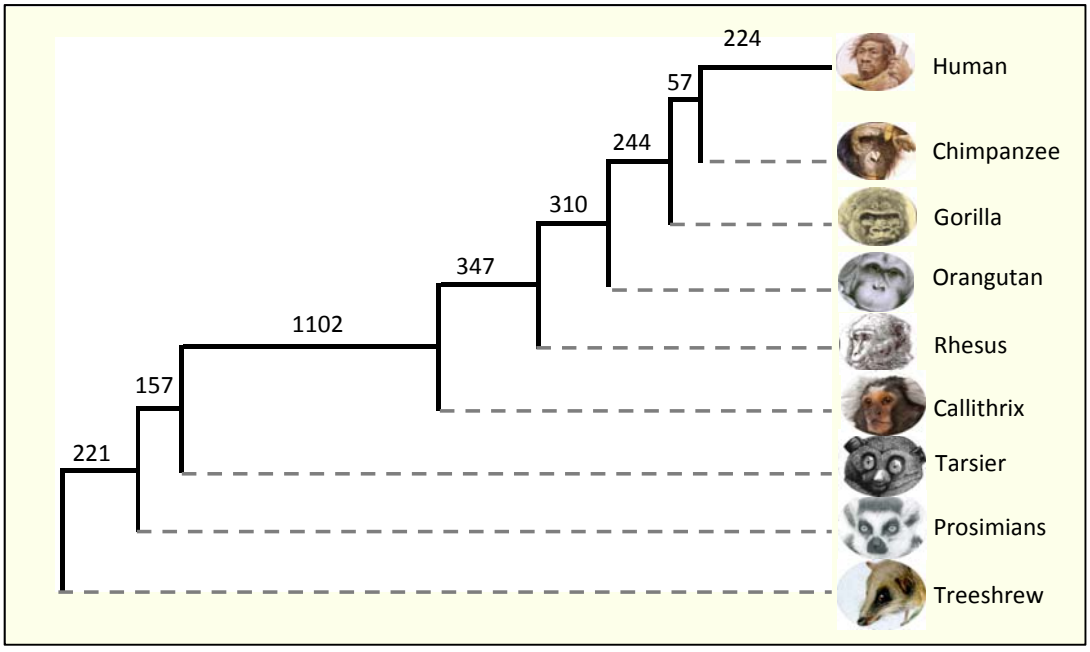


Figure 23. Phylogenetic three showing GABPa CREs gained on the ancestral lineages leading to humans.

To explore which are the biological and molecular functions of those genes harboring human-specific GABPa binding sites, we further performed enrichment tests based on their Gene Ontology annotation (GO). Several GO terms were enriched for biological and molecular functions associated with heart development, RNA processing of tRNAs and mRNAs, mammary gland morphogenesis and development, metabolism and biosynthesis of lipids, and signaling pathways involved in neurological development such as ventral spinal cord interneuron and spinal cord motor neuron cell fate specification, and dorsal/ventral neural tube patterning (Supplementary Table S17, supplementary data file).

4.2.3. Identification of differential gene expression after GABPa knock-down

To independently verify the functionality of the GABPa binding sites we detected by using ChIP-Seq, we performed siRNA experiments to knock-down GABPa protein in HEK293T cells. The knock-down was performed through transfection of cells with two independent silencing molecules followed by genome-wide expression profiling at two different time points (at 24 and 72 hours after transfection) (see Methods). For the analyses of the effects of the siRNA on the expression patterns, we exclusively accounted genes that were differentially expressed (DE) in both knock-down experiments. In total, we had expression data available for 14,873 genes. The 24h time point rendered 1,156 DE genes, while the number of DE genes was higher for the 72h time point (3,238 genes). The number of DE genes with a GABPa binding site in the promoter region was higher than expected by chance at the 72h time point (pvalue<0.001) (see Methods). Consequently, we used the results we obtained for the 72h time point for downstream analyses of DE genes. In total, there were 4,531 genes with expression data and harboring at least one GABPa binding site in their promoter region. About a quarter of these genes (1,215) presented significant changes in expression after GABPa knock-down at the 72h time point. These genes are strong putative primary target genes of GABPa (Supplementary Table S18, supplementary data file).

4.2.4. Analyses of differentially expressed genes after GABPa siRNA interference

To explore the main biological, cellular and molecular functions of those genes that are putative primary targets of GABPa, we performed a GO enrichment analyses (see methods). Functional GO categories associated with RNA processing, ribosome biogenesis, regulation of lipid metabolism, protein ubiquitination, neuron projection development and innate and adaptive immune responses, among others, were significantly enriched (Supplementary Table S19, supplementary data file). We additionally tested for GO enrichment among those genes that were DE, but did not harbor a GABPa binding site in their promoter region, since it is likely they are secondary targets of GABPa (2,023 DE, 72h time point, see methods) (Supplementary Table S18, supplementary data file). We found very similar enriched GO terms between both groups, DE genes with ChIP-Seq peaks and DE genes without ChIP-Seq peaks for GABPa. This suggests that among the DE expressed genes we found a collection of putative primary and secondary targets for GABPa protein, but also that both groups are likely to be involved in very similar functional pathways. For instance, both groups showed significant enrichments for lipid and fat biosynthesis and metabolism such as lipoprotein transport, glycosphingolipid metabolic process, lipid biosynthetic process, fatty acid transmembrane transport, and regulation of fatty acids, mitochondrial biogenesis, among others (Supplementary Table S20, supplementary data file). We additionally found an overrepresentation of functional groups associated with, for instance, regulation of endothelial tube morphogenesis and endothelial cell proliferation, and epithelial cell differentiation involved in mammary gland development.

A further inspection of genes having at least one GABPa binding site in their promoter region revealed 13 genes that encode for subunits of the Cytochrome c oxidase (COX) enzyme (COX10, COX15, COX16, COX17, COX18, COX4I1, COX4NB, COX5B, COX6A1, COX6B1, COX7A2, COX7C, COX8A) (Supplementary Table S11, supplementary data file). Ten of them showed significant changes in expression after GABPa knock-down with at least one of the siRNA molecules. Several COX subunit genes were also differentially expressed besides not having a GABPa

binding site in their promoter region (COX1, COX11, COX7BP1 COX14, COX20, COX4I2, COX5A, COX6CP1, COX6CP2, COX7A2P2).

4.2.5. Differentially expressed genes with human-specific GABPa binding sites

We first verified if genes carrying newly evolved GABPa binding sites showed significant changes in expression levels after GABPa knock-down. Out of the 217 genes with human-specific binding sites, there were 177 with expression data in our siRNA experiments. We found that 52 out of the 177 were significantly DE in both siRNA experiments (72h time point). Genes carrying human-specific GABPa binding sites were also enriched among the genes exhibiting significant changes in expression (Fisher's Exact test, p-value= 5.79^{-10}). Out of the six human- and great ape-specific genes carrying human-specific binding sites, four (*CEP170, RPL41, GUSBP4*, *STAG3L4*) were found DE after knocking-down GABPa with at least one of the siRNA molecules (Supplementary Table S16, supplementary data file). This finding suggests that newly evolved GABPa binding sites in human and apes are functional.

GABPa regulatory activity has been associated with several medical consequences in humans (Piñero et al. 2015). We explored if the 52 DE expressed genes carrying human-specific GABPa binding sites share common disease associations with GABPa. By using the database "DisGeNET" (Piñero et al. 2015), we identified 17 out of the 52 genes that have been associated with at least one of the diseases GABPa has been associated with as well: diabetes (*PCMT1, UGGT2*), Parkinson disease (*RPL6, TDP2, HSPA8*) and breast cancer (*ACOT13, ANTXR1, BAG4, EMG1, HSPA8, NEK7, YPEL3, ZNF398*), among other diseases (Supplementary Table S21, supplementary data file).

4.2.6. Functional analysis of newly evolved GABPa binding sites using reporter gene assays

Taking into consideration we identified genes harboring human-specific binding sites; we selected four promoter regions for five genes (*ZNF197, ZNF398*, *ZNF425, ANTXR1* and *TMBIM6*) to further validate GABPa regulatory activity using reporter

gene assays. In addition, since we found enrichment for members of the KRAB-ZNF class, and knowing that GABPa binds and regulates bi-directional promoters, we first chose two promoter regions corresponding to three KRAB-ZNF genes, *ZNF197* and bi-directional promoter *ZNF398* and *ZNF425*. The gene *ZNF197* has two GABPa binding sites in its promoter region, one being specific to humans, while the other one is conserved among the mammalian clade. The bi-directional promoter carries two overlapping binding sites with two human-specific nucleotide changes located ~130 bp apart from the TSSs. To account for the bi-directional functionality of this promoter, we cloned it in both directions. We also considered the promoter regions of the anthrax toxin receptor-1 gene (*ANTXR1*) and the Transmembrane BAX Inhibitor Motif Containing 6 (*TMBIM6*) genes. The promoter region of the human *ANTXR1* gene carries three GABPa binding sites, while chimpanzee and rhesus it carries just two. In addition, this gene is overexpressed in HEK293T cells. Despite the fact our main interests was aiming to find and understand the regulatory evolution of GABPa at human-specific level, we also considered the promoter region of *TMBIM6*, a gene that harbors a hominid-specific GABPa binding site. In addition, *TMBIM6* exhibited a strong GABPa ChIP-Seq peak in its promoter region and high expression levels (Sultan et al. 2008). We also identified a highly conserved GABPa binding site located in close proximity to the hominid-specific binding site for *TMBIM6*, nonetheless, this did not entirely match the GABPa core consensus sequence "GGAA" (Figure 24) and was under represented in the whole set of binding sites found here (0.94%).

Two fragments of each orthologous promoter, for human, chimpanzee and rhesus macaque genomic DNA, were cloned. One fragment corresponded to the wild type (*wt*) and the other one to the mutated version (*mut*). In the mutated types, one or two nucleotides were changed in the human binding sequence to simulate the ancestral state. Conversely, chimpanzee and macaques binding sites were modified to mimic the human-specific GABPa binding sites (Figure 25). We measured expression changes for all constructs, enabling us to quantify the effects of different GABPa binding sites on the promoter activity (Figure 25). We detected significant differences in the promoter activity between human wt promoter and the chimpanzee and rhesus macaque wt promoters for *ZNF197*, and *ZNF398*/*ZNF425*

genes (Figure 25). In the case of the bi-directional promoter for the genes *ZNF398*/*ZNF425*, the directionality had a cell–specific effect on the reporter activities. The wt promoters, for all three species, exhibited significant differences in direction *ZNF398* in activity in COS-1 cells and *ZNF425* in HEK293T cells (Figure 25). The insertion of one nucleotide mutation into the chimpanzee's and rhesus macaque's *ZNF197* binding sites to recreate the human version yield a significant increase in the promoter activity in both cell lines. However, the modification of the human *ZNF197* wt promoter into the ancestral state did not cause significant changes in the promoter activity

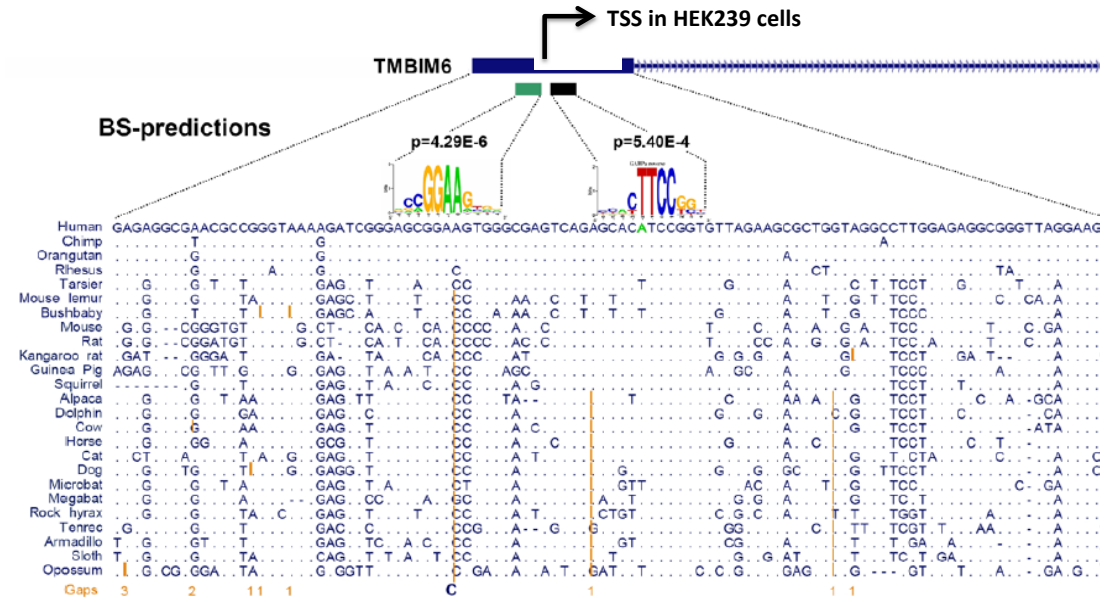


Figure 24. View of the promoter region of the gene TMBIM6, the GABPa ChIP-Seq reads and the GABPa binding site predictions for multiple sequence alignment. The first exon (5'UTR) is shown as blue bar with a black arrow indicating the transcription start site (TSS) in HEK293T cells. GABPa binding sites are represented as green and black boxes. GABPa logos and their counterpart positions are shown over the multi species sequence alignment. For the sequence alignments, dots represent sequence conservation to the human reference. Orange bars indicate bases that are not depicted. Orange numbers below the sequence alignment corresponds to the sum of bases not depicted. The blue (C) illustrates the presence of a single cytosine in all non-haplorhini. Taken from (Perdomo-Sabogal et al. 2016).

This suggests that the sequence around the core-binding motif might also contribute to maintain the promoter activity. Likewise, the introduction of the human-specific GABPa binding site in the wt promoter of *ZNF398*/*ZNF425* for chimpanzee and rhesus macaque resulted in a significant increase in the activity in

both cell lines, while in both cases the ancestral mutation in the human binding site significantly reduced, more than a two-fold, the promoter activity.

The *TMBIM6* and *ANTXR1* wt promoter activities were significantly higher as well, in at least one cell line, when compared with the wt promoters from chimpanzee and rhesus macaque. Interestingly, the promoter activity of the hominid-specific GABPa binding site, promoter *TMBIM6*, was higher in humans when compared with chimpanzee (Figure 25), which suggests there is an effect of the sequence differences between human and chimpanzees on the activity of this promoter. The mutation of the rhesus macaque *TMBIM6* wt promoter by inserting the hominid-specific binding site resulted in a significant change in the promoter activity. The insertion of the hominid-specific binding site in the human promoter significantly decreased the reported activity below the initially reported activity for the wt in chimpanzee. The effect was similar for the chimpanzee promoter activity, which fell below the one reported for rhesus macaque's wild type promoter. For the *ANTXR1* promoter, the mutation of the wt promoters from chimpanzee and rhesus macaque by incorporating the human GABPa binding site significantly increased the reporter activity in three out of four cases, for chimpanzee in both cell lines and macaque in HEK293T cells (Figure 25).

Promoter (length)	# of BS per peak	Species	Wild type (wt)	Mutated (mut)	Log2 ratios (mut/wt) HEK293	COS-1
ZNF398	2	HSA	CTCGGAAGCG---GAAGCCG	CTCGGCAGCG---GAGGCCG	↓ -1.16 ***	↓ -1.69 ***
(329bp)	0	PTR	CTCGGcAGCG---GAcGCCG	CTCGGAAGCG---GAAGCCG	↑ 0.92 ***	↑ 2.45 ***
	0	MAC	CcCGGcAatGgctGggGCCG	CTCGGAAGCG---GAAGCCG	↑ 2.22 ***	↑ 2.28 ***
ZNF425	2	HSA	CTCGGAAGCG---GAAGCCG	CTCGGCAGCG---GAGGCCG	↓ -0.32 ***	↓ -0.76 ***
(329bp)	0	PTR	CTCGGcAGCG---GAcGCCG	CTCGGAAGCG---GAAGCCG	↑ 0.34 ***	↑ 0.44 **
	0	MAC	CcCGGcAatGgctGggGCCG	CTCGGAAGCG---GAAGCCG	↑ 0.35 ***	↑ 0.69 ***
ZNF197	2	HSA	TGCCGGAAGGGC	TGCCGCAAGGGC	→ 0.08 ns	→ -0.04 ns
(430bp)	1	PTR	TGCCGcAAGGGC	TGCCGGAAGGGC	↑ 0.25 *	↑ 0.47 ***
	1	MAC	TGCCGcAAGGGC	TGCCGGAAGGGC	↑ 0.35 ***	↑ 0.64 ***
ANTXR1	3	HSA	GCGAGGAAGGGC	GCGAGGAGGGGC	↓ -0.11 ns	↓ -0.34 *
(633bp)	2	PTR	GCGAGGAgGGGC	GCGAGGAAGGGC	↑ 0.29 **	↑ 0.54 ***
	2	MAC	GCGAGGAgGGGC	GCGAGGAAGGGC	↑ 0.14 *	↑ 0.19 ns
TMBIM6	1 (2)	HSA	GAGCGGAAGTGG	GAGCGGACGTGG	↓ -0.45 ***	↓ -0.93 ***
(576bp)	1 (2)	PTR	GAGCGGAAGTGG	GAGCGGACGTGG	↓ -0.49 ***	↓ -0.81 ***
	0 (1)	MAC	GAGCGGAcGTGG	GAGCGGAAGTGG	↑ 0.30 ***	↑ 0.70 ***

Figure 25. Reporter gene assay constructs and promoter activities given as log2 ratios of average normalized firefly luciferase gene expression. The number of predicted binding sites is indicated for each gene. HSA: *Homo sapiens*, PTR: *Pan troglodytes* (chimpanzee) and MAC: *Macacca mulatta*

(rhesus macaque). *wt* and *mut* correspond to wild type and mutated sequences respectively. Underlined bases indicate differences from the human wt sequence. In green mutated bases and red generated or disrupted of GABPa binding site, respectively. Increase or decrease in reporter activity are represented by green or red arrows, respectively. Yellow arrows represent no change. Significance levels, as determined by Welch's t-test for unequal variances, are indicated as (*) p-value < 0.05, (**) p-value < 0.01, (***) p-value < 0.001 and (ns) not significant. Taken from (Perdomo-Sabogal et al. 2016).

In summary, altering the GABPa non-human wt promoter sequence by inserting the human-specific mutation resulted in a significant increase in the reporter activity, in both cell lines, in 17 out of 18 cases. Conversely, the disruption of the GABPa binding site in human and chimpanzee promoters resulted in a significant decrease of the reporter activity, 9 out of 12 cases. We did not detected conflicting effects, meaning that the insertion of binding sites did never result in a significant decrease in activity, and the disruption did never result in significant increases in activity.

4.3. Discussion

Integrating data from different experiments allowed us to determine that GABPa protein regulates a considerable number of human genes. In total, we identified GABPa binding sites located in the promoter region of 5,321 genes, from which 217 have a newly evolved binding site in humans. By using expression data from siRNA knock-down experiments, we also detected that GABPa binds and regulates expression of almost one third (31%) of the genes in HEK293T cells. In addition, 1,215 genes harboring at least one GABPa binding site exhibited significant changes in expression after GABPa knock-down, suggesting this set of genes constitutes a list of potential primary targets of GABPa. Taken together, our results suggest that GABPa is involved in regulating the expression of genes that have relevant functions at neurological, metabolic, endothelial, epithelial, and mammary morphogenetic level.

4.3.1. Newly evolved GABPa binding sites are functional

To detect newly evolved human- and hominid-specific GABPa binding sites in HEK293T cells, we applied ancestral sequence reconstruction approach for 11,008

human core consensus sequences. We used whole genome sequence alignments of 44-vertebrates. The accuracy of this strategy relies on UCSC alignments, which for human-chimpanzee and human-macaque results suspicious for 0.004 and 0.02% of the sequences aligned (Prakash and Tompa 2007). Therefore, this suggests that out of the 11,008 human-macaque alignments of 11 bp each, 24 nucleotides were ambiguously aligned. However, a majority of the suspicious alignments occur on intronic and intergenic regions (Prakash and Tompa 2007), which also suggests that the fraction of problematic alignments is likely to be smaller in our results. In addition, genic promoter regions are highly conserved in mammals and other vertebrates (Mahony et al. 2007). In this sense, a preponderant number of GABPa binding sites were located in the close proximity of the TSSs, 300 bp up- and downstream, which also suggest higher accuracy in the alignments.

To further validate if the GABPa binding sites we found are functional, we experimentally validated the regulatory roles of four of them, three carrying newly evolved GABPa human binding sites, and one hominoid specific one. We applied dual reporter gene assays for human, chimpanzee and macaque promoters in human HEK293T and African green monkey-derived COS-1 cell. For both species-specific cellular backgrounds the reporter activity was similar, thus suggesting independence of the cell line species type. This indicates that promoter activity of human-specific GABPa binding sites can also be measured in the biological background of an old world monkey cell line.

Our results indicate that the insertion of human-specific GABPa binding sites into the promoter region of *ANTXR1* and *ZNF197* for chimpanzee and macaque predominantly resulted in significant increases in the reporter gene expression, while the disruption of the newly evolved binding site in human did not result in a significant decrease of the promoter activity. From a general perspective, this should not be appraised as the irrelevance of human-specific binding sites, since it is likely that under different biological scenarios these newly evolved binding site may still have a functional effect. For instance, the presence of sequence variations in the promoter region to facilitate the settlement of the transcriptional machinery could be understood as a compensatory effect that switches the activity of the new

GABPa binding site. The promoter regions of both genes (*ZNF197* and *ANTXR1*) carry human-specific variants located less than 100 bp of the newly evolved GABPa binding site. We found similar results for the promoter of *TMBIM6*, where despite human and chimpanzee carrying the very same GABP core consensus sequence, the human wt promoter drives comparably higher reporter activity. This could be the effect of two single nucleotide variants on the human promoter that are closely located to the GABPa binding site (Figure 24), TMBIM6 sequence alignments. This suggest an evolutionary framework where increases in gene expression of TMBIM6 was likely beneficial, possibly followed by a reduction in the selective pressure.

Contrary to what was observed by Collins, et al (2007), who introduced GABPa binding sites in the promoter regions of genes that are not targets of GABPa protein, we did detect significant change in the regulatory activity after inserting human-specific GABPa binding sites into promoters of chimpanzee and macaque. Our main conclusion here is that binding sites require a particular genomic context to produce an effect on their regulatory activity.

Out of the 4,531 genes with expression data in our GABPa siRNA experiments and that carry at least one GABPa binding site, we found that 1,215 of the genes with at least one GABPa binding site significantly changed in expression after GABPa knock-down. Interestingly, genes harboring newly evolved binding sites in human were enriched among the genes with DE after GABPa knock-down. This highlights that a representative number of newly evolved binding sites in humans in involved in the regulatory activity of GABPa in humans. This includes, *ANTXR1* and *ZNF398*, genes that we also tested in our reporter assays.

4.3.2. Human-specific GABPa binding sites regulate genes that are potentially important for human evolution and human diseases

Newly evolved binding sites might result in changes in the dynamics in which genes are regulated at species-specific level. Among the set of genes harboring a human-specific GABPa binding sites, we found that KRAB-ZNF transcription factors genes were significantly overrepresented. KRAB-ZNFs, considered a relatively young

group of genes with many fast evolving members (Hamilton et al. 2006; Huntley et al. 2006; Nowick et al. 2010) and seem disposed to gain novel GABPa binding sites. This group of genes has been associated with important biological process, including regulatory functions at brain and developmental level (Najmabadi et al. 2011; Zhang et al. 2011), thus becoming excellent candidates for studying their roles in post-zygotic isolation, speciation in mammals (Nowick et al. 2013) and brain function in humans. Therefore, it would be interesting, for instance, to study the regulatory roles that newly evolved GABPa binding sites have introduced, via regulation of KRAB-ZNFs, in the neurological pathways at human brain level.

We also found that genes carrying newly evolved GABPa binding sites in humans were significantly enriched for molecular and biological processes associated with RNA processing, especially tRNAs, and signaling pathways involved in ventral spinal cord interneuron and spinal cord motor neuron cell fate specification and dorsal/ventral neural tube patterning (Supplementary Table S17, supplementary data file). As an example, *DARS* and *PARS2* genes carry human-specific GABPs binding sites in their promoter regions, and both were DE after GABPa knock-down. These two genes encode for aminoacyl tRNA synthetases (aspartyl- and prolyl-tRNAs) and catalyze the incorporation of amino acids to their cognate tRNAs. Changes affecting the molecular functionality of *DARS* and *PARS2* have resulted in neuronal disorders of the peripheral nervous system, amyotrophic lateral sclerosis, and ataxia (Park et al. 2008), brain stem and spinal cord leukoencephalopathy with cerebellar and dorsal column dysfunctions (Sissler et al. 2007; Taft et al. 2013), and Alpers syndrome (Sofou et al. 2015). GABPa has also been associated with spinocerebellar ataxia (Lee et al. 2014; Piñero et al. 2015), a disease that similarly to leukoencephalopathy causes the progressive deterioration of locomotion during childhood or adolescence in humans. Therefore, we suggest that changes affecting *GABPa* expression could be associated with progression deterioration of certain areas of the brain.

Further examples of genes harboring newly evolved GABPa binding sites and exhibiting significant changes in expression after GABPa knock-down are *ALDOA, HSPA8,* and *TP73*. We found that these three genes, and also *TMBIM6*, which carries

a hominid-specific GABPa binding site and was tested in our reporter assays experiment, have been linked to the presence of medical disabilities at cognitive level, for instance, Alzheimer's disease (Harris et al. 2007; Kumar et al. 2009) and Parkinson's diseases (Naidoo 2009; Lauterbach 2013), neocortical regionalization and loss of Cajal-Retzius cells (Meyer et al. 2004), irregular accumulation of cerebrospinal fluid in the brain (Yang et al. 2000) and neuronal abnormal apoptosis (Pozniak et al. 2002). We also found that GABPa is involved in the regulation of at least 20 genes coding for COX subunits, with half of them being putative primary targets. COX subunits may modulate the synthesis of the cytochrome c oxidase, an enzyme that plays essential roles in controlling oxidative phosphorylation in eukaryotes (Li et al. 2006). Dysfunctional activity of the COX enzyme has been associated with mitochondrial defects in humans (Barrientos et al. 2002). In correspondence with previous studies, our findings suggest that expression changes in *GABPa* might lead to mitochondrial dysfunction in patients with Alzheimer's disease, (Sheng et al. 2012).

Genes carrying human-specific GABPa binding sites and being DE after we knocked-down GABPa protein have also been associated with the very same diseases that GABPa has been linked to. Decrease in the expression levels of protein-L-isoaspartate (D-aspartate) O-methyltransferase (*PCMT1*), UDP-glucose glycoprotein glucosyl-transferase 2 (*UGGT2*) and *TMBIM6,* have been linked to diabetes types I and II. Expression levels of *PCMT1*, a gene that encode for a repair enzyme, seem to hamper the development and acuteness of diabetes type we (Wägner et al. 2007; Wägner et al. 2008). Similarly, *UGGT2*, which encodes for a protein of the endoplasmic reticulum (ER) that regulates protein transport out of the ER, was significantly down regulated in patients afflicted by diabetes type II when compared to non-diabetics (Marchetti et al. 2007). We also detected enrichment for processes associated with the appearance and frequency of other metabolic disorders. For instance, triglyceride biosynthesis and cholesterol transport, both processes being linked to usually high blood sugar levels. Taken together, we suggest that newly evolved GABPa binding sites might be introducing novelty in metabolic pathways, thus possibly increasing the susceptibility for developing diabetes and other metabolic disorders.

GABPa protein has been found to play essential roles, for instance, as co-regulator, in controlling breast epithelial cell migration (Odrowaz and Sharrocks 2012). Noteworthy, we found that genes carrying newly evolved GABPs binding sites are enriched for processes associated with cell epithelial and endothelial migration, and mammary gland development. For instance, we found that the Prohibitin 2 gene (*PHB2*), also known as Repressor of Estrogen Receptor Activity (*REA*), carries a GABPa binding site 26 bp downstream of its TSS. This gene is essential for mammary gland development including mammary gland cell proliferation, breast alveolus development and postnatal breast development (Mussi et al. 2006). We additionally detected that *PHB2* displayed significant changes in expression after GABP knock-down. Given the evolutionary singularity of the human breast, we suggest that this GABPa binding site is a very promising candidate for a functional contribution to this human trait.

4.4. Conclusions

Using a combinatorial strategy that involved the integration of ChIP-Seq data and comparative genomics, followed by experimental validation of human-specific GABPa binding sites, we identified and verified the functionality of newly evolved GABPa binding sites. Our contributions portray a scenario that bridges transcriptional regulation by GABPa with the evolution of particular human traits and speciation. These human-specific phenotypes include brain functions, breast morphology, and metabolic pathways among others. The integrative strategy implemented here as well serves as an example for combining bioinformatics approaches with the constantly increasing number of publicly available datasets, especially for studying CREs evolution on a "TF-ome"-wide level.

4.5. Methods

4.5.1. Chromatin immunoprecipitation-sequencing

ChIP-Seq experiments were performed according to the protocol established by Warnatz et al. (2011). In summary, we cross-linked $5x10^8$ HEK293T cells for 10 min

at room temperature with 1% formaldehyde, we prepared the nuclei following the published protocol and fragmented the chromatin to 100-500 bp size by 45 cycles of 30 sec on/off at the highest amplitude using a Bioruptor water bath sonicator (Diagenode). We immunoprecipitated the nucleic acids with 10µg rabbit anti-GABP-α (H-180X, Santa Cruz Biotechnology sc-22810) and 70 µl Protein G-Dynabeads (Invitrogen). After washing of beads, we eluted the protein-DNA complexes, the crosslinks were reversed overnight, and the DNA was purified according to the manufacturer's protocol. For sequencing library preparation, we subjected 2 ng ChIP DNA and 10 ng Input DNA to end-repair, addition of adenine bases and ligation of sequencing adapters, followed by DNA amplification through PCR and subsequent gel purification for sequencing on an Illumina Genome Analyzer GAII according to the manufacturer's protocol for 36 bp reads. We mapped the reads the human reference genome version hg18 using Eland. Mapping resulted in 6,955,499 GABPa ChIP reads uniquely mapped (allowing up to two mismatches) and 2,948,346 corresponding reads from the input DNA. A replicate ChIP-seq experiment performed later for validation of the initial experiment resulted in 16,856,422 GABPa ChIP reads and 26,104,399 reads from the input DNA.

4.5.2. Peak calling, gene mapping, MEME and MAST analysis

To analyze the ChIP-Seq data, we followed three steps as previously suggested by Valouev (2008). We used the peak-calling package Quantitative Enrichment of Sequence Tags (QuEST) to identify genomic regions with high density read coverage (peaks) within the 6.96 million mapped reads from GABPa ChIP-Seq. In total, we found 6,208 genomic peaks for GABPa (Supplementary Table S8, supplementary data file). Analysis of the replicate ChIP-Seq experiment with 16.8 million reads resulted in 8,311 GABPa peaks that were used for determining the overlap between both replicates, and for calculating the irreproducible discovery rate (IDR) as described before (Landt et al. 2012). Using the 6,208 peaks obtained for the first replicate and the UCSC annotation for 82,961 known transcripts for the human genome, we assigned peaks to the nearest TSS (±5 kb of the transcripts) (Karolchik et al. 2014). Transcript IDs were converted to Entrez gene IDs using UCSC's knownToLocusLink table. Out of all the peaks, 1,116 mapped to UCSC transcripts that do not have an Entrez gene ID assigned (Supplementary Table S11,

supplementary data file), with 281 corresponding to putative nucleic acids and 20 to human cDNAs.

We extracted 200 bp peak-associated sequences around the peak center via UCSC table browser (Karolchik et al. 2014) and using the default parameters we applied MEME algorithm (Bailey et al. 2009) to identify over-represented GABPa consensus core motifs. Out of the 6,208 peaks, 97% contributed to the motif found by MEME (Supplementary Table S9, supplementary data file Since MEME assumes there is none or one motif within a particular sequence, we then used MAST and the MEME derived position weight matrix, to identify peaks containing more than one GABPa consensus motif within each peak (Supplementary Table S10, supplementary data file).

4.5.3. Multiple sequence alignment extraction and conversion

Using the whole genome alignment for 44 vertebrates from UCSC and the UCSC table browser, we retrieved the alignments corresponding to the GABPa binding sites within the ChIP peak regions. The genome alignments consists of 1-200 bp of multiple sequence alignments that can be concatenated. We then converted multiple alignment formatted alignments into FASTA format, and excluded the non-syntenic blocks and species with missing sequence data. For instance, the insertions were not included in the alignments (Supplementary Table S13, supplementary data file).

4.5.4. Ancestral sequence reconstruction

For determining the ancestral sequences we used the program ANCESTORS (Blanchette et al. 2004). This program allows, among other features, to rebuild the most likely scenario of insertions and deletions within the sequence alignment. It also takes care of preserving high accuracy in the sequence alignment. To use ANCESTORS, we first input the information from the multiple sequence alignment and the phylogenetic tree, which already included the branch lengths. Phylogeny was sourced from UCSC (phyloP44wayPlacMammal) (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/phastCons44way/vertebrate.mod). The branch lengths were estimated by using baseml program included in the

Phylogenetic Analysis by Maximum Likelihood (PAML) package (Yang 2007). The Hasegawa, Kishino and Yano nucleotide substitution model (HKY) (Hasegawa et al. 1985) was applied in both programs. Therefore, we were able to reconstruct the ancestral sequences of 34 mammalian species using the genome alignment for 44 vertebrates (Diallo et al. 2007).

4.5.5. Cloning and plasmid preparation

We designed the PCR primers using the Primer3 online service and extended by 29 bp Gateway attB tails (Invitrogen) at the 5' end of each primer. We performed the PCR Touch-down PCR as described previously (Ralser et al. 2006), except for the supplementation of each reaction with 0.001U *Pfu* polymerase. We introduced the mutations by using primer-mediated mutagenesis. To facilitate cloning, we amplified the Gateway cloning cassette (Invitrogen) with the forward primer attP1 and the reverse primer attP2 and cloned into the pGL3 reporter vector (Promega). We purified the PCR products and cloned upstream of the luciferase gene in the modified pGL3 vector using BP Clonase II Enzyme Mix (Invitrogen) following the manufacturer's instructions. We transformed the plasmids into the methylation-deficient Dam^- *E. coli* strain GM2929. We validated the inserts of positive clones using capillary Sanger sequencing (Services in Molecular Biology, Berlin, Germany). The concentration of DNA was measured on a Nanodrop UV spectrophotometer (NanoDrop Technologies) and standardized to 50 ng/µL for transfections.

4.5.6. Cell culture, transient transfection, and reporter gene activity assays

Reporter genes are frequently used to measure gene expression and molecular processes coupled to changes in gene expression. In this method, the reporter gene is cloned together with the DNA sequence of interest, in this case, with the CREs of GABPa, and then transfected into cells. Subsequently, cells are assayed for the activity of the reporter gene by directly measuring the reporter expression itself. In our experiments, the HEK293T and COS-1 cells were cultivated in Dulbecco's modified Eagle's medium (DMEM, Gibco) supplemented with 100 U/ml penicillin/G-streptomycin (Biochrom) and 10% heat-inactivated fetal bovine serum (Biochrom) at 37°C and 5% CO_2. We seeded ~15,000 (HEK293T) and ~5,000 (COS-

1) cells per well in clear-bottom 96-well plates (Costar). We co-transfected co-transfected 150 ng of experimental firefly luciferase plasmid twenty-four hours after seeding together with 10 ng of *Renilla* luciferase control plasmid (pRL-TK, Promega) in five replicates using Lipofectamine 2000 following the manufacturer's recommendations. Cells were lyzed 24 hours post-transfection.

We measured firefly luciferase and *Renilla* luciferase activities using the Centro LB960 luminometer (Berthold) and the Dual Luciferase Kit (Promega). We followed the protocol suggested by the manufacturer with the exception of injecting 25 µl each of the firefly luciferase and *Renilla* luciferase substrate reagents. All measurements were performed at least in three technical and two biological replicates, including new dilution and concentration adjustments of reporter plasmids.

4.5.7. Inhibition of GABPa Expression by RNA Interference

GABPa knock-down experiments were performed using two independent siRNA molecules, specifically one unmodified synthetic small interfering RNA (Qiagen SI00423311) and one chemically modified synthetic small interfering RNA (Invitrogen HSS103907). We seeded HEK293T cells in 12-well plates together with siRNA-HiPerFect complexes according to the HiPerFect fast forward protocol (Qiagen). For the mock transfections, we treated the cells with HiPerFect reagent only. We performed knock-down transfections in triplicates, and mock transfections were performed in quadruplicates. We extracted the total RNA from cultured cells at 24h and 72h post-transfection using the RNeasy mini kit (Qiagen) following the manufacturer's instructions. All RNA samples were DNase-treated, purified, quantified, and inspected for integrity. For hybridizations on microarrays, biotinylated cRNA was synthesized using the GeneChip expression 3' amplification one-cycle target labeling and control reagents (Affymetrix). Following integrity control, we hybridized the cRNA to the Affymetrix GeneChip HG-U133Plus2. Then we washed and stained the arrays, and scanned following the recommended protocols from Affymetrix.

To analyze gene expression we used the "affy" package from Bioconductor (Gautier et al. 2004). We then calculated Robust Multi Array (rma) normalized expression values for each probe set (Bolstad et al. 2003), included those probe sets with a reliable detection (p-value < 0.05), and merged the rma values of probe sets belonging to the same gene by obtaining one mean expression value (Nowick et al. 2009). For detecting genes that were differentially expresses between the knock-down experiments and the mock transfected samples at two different time points (24 and 72 hours), we used the package "multtest" (Pollard et al. 2005) from Bioconductor. We considered as DE all those genes that exhibited significant changes in expression after GABPa knock-down using both molecules for each time point (24 and 72 hours, p-value < 0.05). We then overlapped the GABPa candidate target genes from our ChIP-Seq experiments with the DE genes, obtaining 392 and 1,215 overlapping genes for 24 and 72 hours time point, respectively. To analyze if these overlaps were higher than expected by chance, we performed a permutation test, which indicated that the detected overlap was higher than expected by chance for the 72 hours time point (p-value = 0.001). This was not the case for the 24 hours treatment (p-value = 0.137). Consequently, we selected the data obtained for the 72 hours treatment for downstream analyses.

4.5.8. Gene ontology enrichment analyses

Taking into account that we obtained three different gene sets out of the ChIP-Seq and siRNA knock-down experiments (72 hours time point), we performed three independent GO enrichment analyzes, one per gene list, to functionally characterize these groups of genes. To do so, we used the hypergeometric test implemented in FUNC (Prüfer et al. 2007). With the first GO enrichment analysis we aimed to functionally characterize the 217 genes that harbor human-specific GABPa CREs by comparing them with the whole set of genes that have at least one GABPa binding site in their promoter region (TSS ± 5 kb) (Supplementary Tables S14 and S17, supplementary data files). In the second GO enrichment analysis we tested for enrichment of GO groups among the genes that have at least one GABPa CRE in their promoter regions and that were DE after GABPa knock-down experiments with both silencing molecules (Qiagen: SI00423311 and Invitrogen: HSS103907). We compared this set of GO groups against the set of DE genes that do not have ChIP-

Seq peaks in their promoter region (Supplementary Table S18 and S19, supplementary data files) In the third analysis, we tested if DE genes after GABPa knock-down that do not have a GABPa CRE in their promoter region were enriched for particular GO groups (Supplementary Tables S18 and S20, supplementary data files). In all three cases, we refined the initial results of the enriched GO groups as implemented in FUNC (Prüfer et al. 2007) applying a p-value cutoffs of $p<0.05$ for the first and third tests test and of $p<0.01$ for the second test after refinement.

Conclusions

Deciphering how different mechanisms are integrated to regulate gene expression, and thus, define phenotypical differences between and within species still represents a challenge with an infinite world of possibilities to be studied. By putting together all the extant information about genes involved in the regulation of the expression of other genes, we built an up to date catalog of human GRF genes. This catalog enabled us to identify several GRFs that are good candidates for playing a master key role in regulating human physiology, and hence, potentially in human adaptive evolution and speciation. By analyzing extant information about genetic changes that have occurred since humans split from their last common ancestor with chimpanzee, we were able to portray how changes in GRF genes may have contributed to the appearance of new functions and thus, new species-specific traits. In addition, by identifying human-specific cis-regulatory elements for the GRF protein GABPa, we also suggested how newly evolved genomic regions might have driven towards the evolution of human-specific traits; for instance, cognitive abilities and neurological development; but also to particular medical conditions such as diabetes, Alzheimer and Parkinson diseases.

Subtle changes in GRF proteins may also result in fine regulatory changes between individuals and populations within species. It is expected that slight molecular changes that result in a functional improvement of an individual's fitness, will increase in frequency at population level due to positive selection. In humans, such changes may have significantly contributed to human adaptation while expanding to colonize different latitudes. By exploring data from genome-wide scans for detecting positive selection, we draw several scenarios in which changes in GRF sequences may have led to adaptive responses at population-specific level in humans. Our results suggest that at least six of the larger classes of GRF genes may have differentially contributed to the diversification of the human regulome during the Out of Africa human expansion by slight changes in DNA sequences. For instance, we found that the C2H2-ZNF GRF class, including KRAB-ZNF genes,

displays an enrichment of genes in regions exhibiting signatures of positive selection at population specific level. Further exploration allowed us to detect that several C2H2-ZNFs are located in regions with reduced haplotype heterozygosity and contain many mutations causing nonsynonymous changes in their coding regions. Despite the number of amino acid changes in the GRF's protein domains was rather small, we suggest that such changes could explain subtle regulatory functional variation. This subtle variation could have led to differences in reproduction, insulin/glucose and lipids metabolism at population-specific level.

We conclude that identifying genetic variation that modifies the ways in which gene expression is fine-tuned within and between species becomes essential to understand how phenotypical differences and human-specific traits have been shaped during human evolution. While it is widely acknowledged that genetic variation in cis-regulatory elements has played a key role in the evolution and diversification of humans, we think that changes in GRF genes have also significantly contributed with the subtly tweaking of adaptive regulatory pathways, and thus may provide key clues about human speciation and human adaptive evolution.

Appendices

Appendix A

Supplementary tables

Supplementary Table S2. GRF genes located in region exhibiting signatures of positive selection in AMH populations. The different symbols indicate population specific signatures of positive selection. Candidate genes identified by (*) Grossman et al, 2013; (¥) Sabeti et al, 2007; (£) Pickrell et al, 2009. CEU: Utah Residents (CEPH) with Northern andWestern European ancestry, CHB: Han Chinese in Bejing, China, YRI: Yoruba in Ibadan, Nigeria; JPT: Japanese in Tokyo, Japan; Bantu: Bantu-speaking populations. Table taken from Perdomo-Sabogal et al. 2014.

GRF gene	Populations										
	CEU	YRI	CHB	JPT	Oceania	South Asia	East Asia	Mideast	Native Americans	Bantu	Biaka Pygmy
POGZ,MCM6, PCGF1, KCNH7, RFX5	*										
ACTR5, ADNP2, ANKRD45, PAWR, HIPK1, HIRA, RRN3, RNF135, SIN3A, SLC30A9, USF1, TAX1BP3, ZBTB41, EBF1, KCNIP4, NCOA1		*									
ASXL2, FMNL2, FOXP1, LHX8, PAPOLA, CHD2, TERF2IP			*	*							
DPF1		*								£	
HIF1A, SNAPC1	£	*			£	£		£			
ZMYM6	*¥										
WWOX	*¥£				£						
CTNND2	¥										
BMI1, NFE2L2	¥		¥	¥							
AFF2, BBX			¥	¥							
FBN1, MYEF2	¥ £					£		£			
APC	£					£	£	£			
KCNH5, PHF19	£					£		£	£		
ERBB4	£				£	£	£	£			
RFX3	£					£					£
RGS9	£					£		£			
RHOA											
SETBP1	£						£	£		£	£
ARIH2											
ATF6						£			£		
CIITA									£		£
CLOCK											
DUSP12, FOXE1, TRIM14										£	
HEY2							£			£	£
HSF2									£		
NCOA7							£			£	£
POLR2K					£				£		
PPARA					£	£	£	£	£		
SFPQ											
YTHDC1									£		

Supplementary Table S5. KRAB-ZNF cluster defined as in Huntley et al, (2006). Genomic coordinates are based on the human genome reference, version hg19.

KRAB-ZNF clusters (Huntley et al, 2006)	Genomic coordinates	Size cluster (Kb)	KRAB-ZNF genes in the cluster	GRF genes in the cluster	All genes
1	chr1:247099483-247339322	240	4	4	5
2	chr3:40383122-40692996	310	3	3	6
3	chr3:44448334-44848334	400	7	8	9
4	chr4:39541-483540	444	4	5	8
5	chr5:178285305-178462544	177	4	5	5
6	chr6:28040489-28430489	390	10	16	17
7	chr7:63499554-64463192	964	9	9	10
8	chr7:99049770-99236476	187	5	5	9
9	chr7:148669510-149580510	911	11	13	20
10	chr8:145945215-146215827	271	6	6	8
11	chr10:38050199-38512208	462	4	4	4
12	chr12:133481895-133851895	370	4	5	8
13	chr16:3259999-3509999	250	6	8	12
14	chr16:30311000-31175239	250	12	17	43
15	chr18:32806333-32980333	174	5	5	5
16	chr19:2805474-2965474	160	5	5	5
17	chr19:9239000-9889000	650	13	13	16
18	chr19:11569273-12519495	950	17	20	23
19	chr19:19639000-24314329	4675	41	44	79
20	chr19:35156419-35458419	302	5	5	5
21	chr19:36640679-38318696	1678	29	30	31
22	chr19:44008483-45018484	1010	21	24	24
23	chr19:52350013-54088308	1738	37	44	50
24	chr19:55978188-56158188	180	5	9	12
25	chr19:56568969-57378969	810	12	17	19

Supplementary Table S6. Nonsynonymous alleles located on the coding region of genes located within three KRAB-ZNF clusters (whole cluster) exhibiting EHH regions. Shaded green indicates nucleotide variation for nonsynonymous SNPs where the alternative allele is equal to the ancestral allele.

Cluster 1	rs140747159	T	C	T	2.21	ZNF695
	rs2642993	T	G	T	2.22	
	rs117437844	C	T	C	2.21	
	rs55762230	C	T	C	2.22	
	rs2642992	A	G	G	1.91	
Cluster 3	rs34437520	C	T	C	1.6	ZNF167
	rs2034476	C	G	G	1.73	
	rs148794995	G	C	G	3.17	
	rs2272044	C	G	C	3.41	ZNF35
	rs191633770	T	C	T	3.2	
Cluster 14	rs184925516	C	T	C	2.09	STX4
	rs144183945	A	G	a	2.20	
	rs8046978	G	A	G	1.74	ZNF668
	rs749670	A	G	A	1.64	ZNF646
	rs77579502	G	A	G	1.63	
	rs35041466	A	G	A	1.64	
	rs141631516	C	T	C	1.66	
	rs78522165	G	A	G	1.71	
	rs75586809	C	T	C	1.67	
	rs149125224	T	A	T	1.65	
	rs147316630	C	A	C	1.67	
	rs35713203	G	C	G	1.52	
	rs113926102	C	T	C	1.83	
	rs35376811	C	T	C	1.88	
	rs188200157	A	C	A	1.85	
	rs3751856	G	A	G	1.82	
	rs7196726	G	A	G	1.62	
	rs7199949	G	C	C	1.46	PRSS53
	rs188342896	G	A	G	1.74	
	rs11150606	T	C	T	1.34	
	rs191369353	T	C	T	3.96	VKORC1
	rs138110276	C	A	C	3.86	
	rs17855606	A	G	a	2.6	MYST1
	rs142789229	C	T	C	3.31	PRSS8
	rs117442264	C	T	C	2.16	PRSS36
	rs145749002	G	A	G	2.13	
	rs188497178	G	T	G	2.27	

Appendix B

Supplementary Data Files

Description

The following list Excel spreadsheets correspond the supplementary material that was cited in the in text and that could not be included within the main document.

Supplementary Table S1.xlsx (Catalog of gene regulatory factors)
Supplementary Table S3.xlsx (List of GO terms that were used for identifying gene regulatory factors)
Supplementary Table S4.xlsx (Workbook containing information about candidate GRF genes for positive selection.
Supplementary Table S7.xlsx (GRF evolutionary branch assignment)
Supplementary Table S8.xlsx (Genomic peaks found after performing ChIP-Seq experiments with a GABPa specific antibody in HEK293T cells)
Supplementary Table S9.xlsx (Distance of the de novo motif discovery algorithm MEME identified sites from Chip-Seq peak middlepoints)
Supplementary Table S10.xlsx (11,619 PWM hits in 5,797 peak regions of 200 bp found by using the motif alignment and scan tool MAST)
Supplementary Table S11.xlsx (Genomic peak mapping to UCSC transcripts and entrez Ids)
Supplementary Table S12.xlsx (11008 alignments obtained after performing multiple species alignments from the UCSC MultiZ 44 vertebrate alignments)
Supplementary Table S13.xlsx (GABPa binding sites ancestral sequences reconstruction)
Supplementary Table S14.xlsx (217 Genes mapping ± 5kb around Human specific GABPa binding sites)
Supplementary Table S15.xlsx (Hominini-, Homininae- and Hominid-specific GABPa binding sites)
Supplementary Table S16.xlsx (Genes harbouring Hominini-, Homininae- and Hominid-specific GABPa binding sites in their promoter regions)
Supplementary Table S17.xlsx (Gene ontology enrichment analysis for genes located ± 5 kb around Human specific GABPa TFBS. All the other genes located around ±5 kb ChIP-Seq GABPa peaks were used as background set)
Supplementary Table S18.xlsx (Deferentially expressed genes after GABPa knock-down (72h treatment)).
Supplementary Table S19.xlsx (Gene ontology enrichment analyses for 1280 deferentially expressed genes having at least one GABPa binding site in our ChIP-Seq data)
Supplementary Table S20.xlsx (Gene ontology enrichment analyses for 1934 deferentially expressed genes without GABPa binding site in our ChIP-Seq data)
Supplementary Table S21.xlsx (Disease associations for genes with a human specifc GABPa binding site that changed in expression after GABPa knock down. Information sourced from DisGenet)

Appendix C

Supplementary figures

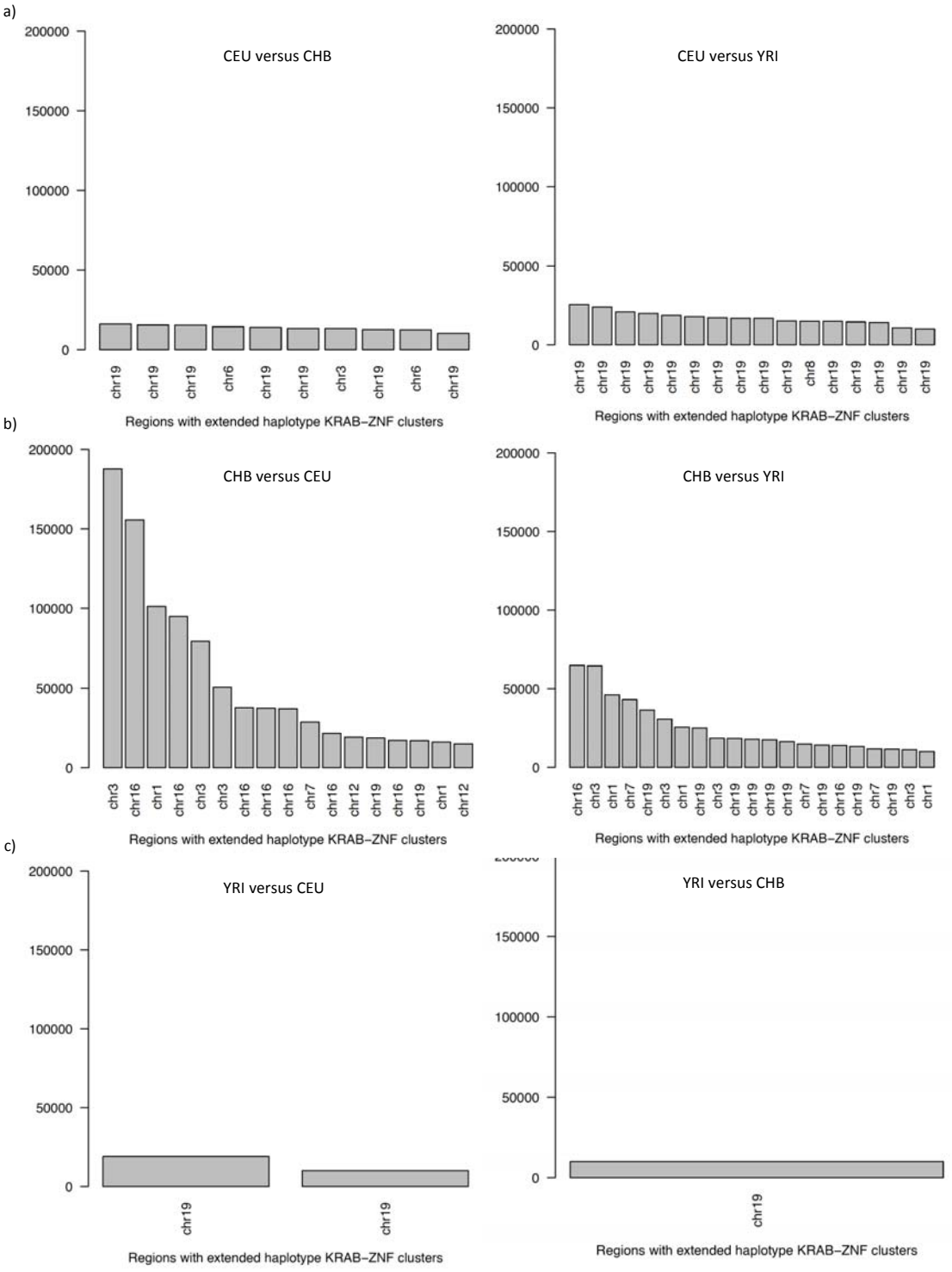


Figure supplementary S1. Distribution of regions exhibiting EHH with consecutive XP-EHH scores higher on the upper 5% tail of the distribution. Regions larger than 10 kb in the three human populations are shown. (a) CEU versus CHB (left panel) and YRI (right panel). (b) CHB versus CEU (left panel) and YRI (right panel). (c) YRI versus CEU (left panel) and CHB (right panel). XP-EHH results indicate that just three KRAB-ZNF gene clusters exhibit EHH larger than 100 kbps (b-left panel).

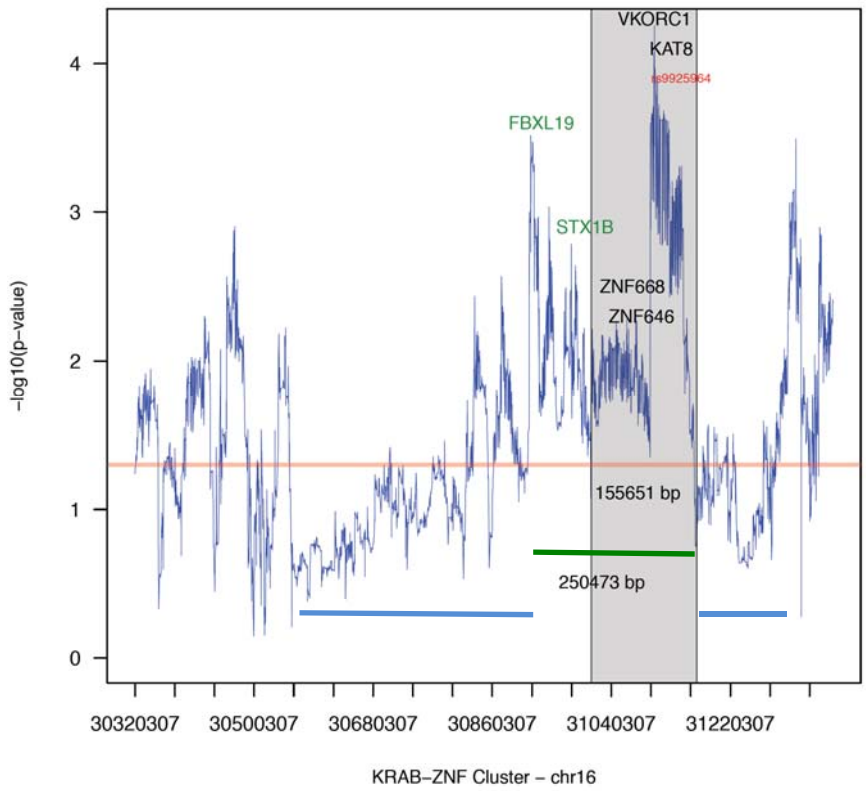


Figure supplementary S2. XP-EHH rank scores indicating the presence of EHH in the KRAB-ZNF cluster 14. Shaded grey region indicates the region that the EHH span within each cluster. Light horizontal red line indicates the 5% threshold set for the XP-EHH distribution (rank score of 1.3). Green horizontal line (c-left panel) indicates a region of around 251 kb in length showing high XP-EHH scores with one SNP causing the EHH decay at position chr16:31009343 (rs74474326). Blue horizontal lines show regions with haplotype decay below the threshold.

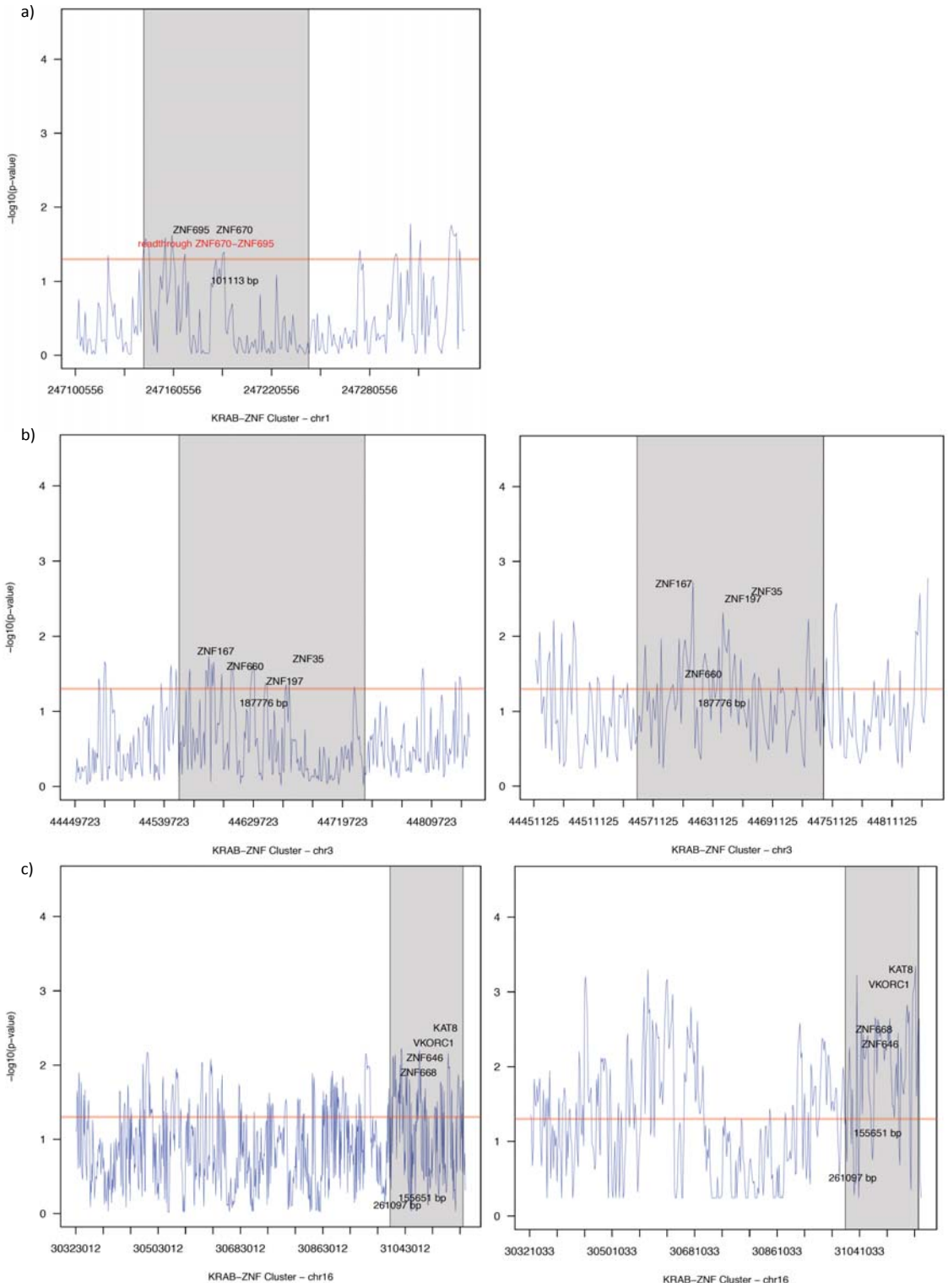


Figure supplementary S3. CLR and XP-CLR rank scores for three KRAB-ZNF clusters exhibiting extended haplotypes. Shaded grey region indicates the region that the EHH span within each cluster. (a-left) CLR scores for the KRAB-ZNF cluster 1. There was no data available for this region for the XP-CLR test. (b) Scores obtained for the KRAB-ZNF cluster 3 for CLR (left) and XP-CLR (right) tests. (c) Scores obtained for the KRAB-ZNF cluster 14 for CLR (left) and XP-CLR (right) tests. Despite the EHH regions within these three KRAB-ZNF clusters exhibit high scores for CLR and XP-CLR, these results do not reveal a clear signature of selection for these regions.

Bibliography

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

Abu-Amero KK, Helwa I, Al-Muammar A, Strickland S, Hauser MA, Allingham RR, Liu Y. 2015. Case-control association between CCT-associated variants and keratoconus in a Saudi Arabian population. J. Negat. Results Biomed. 14:10.

Albert FW, Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. 16:197–212.

Alves I, Srámková Hanulová A, Foll M, Excoffier L. 2012. Genomic data reveal a complex making of humans. PLoS Genet. 8:e1002837.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37:W202–W208.

Barrera LA, Vedenko A, Kurland JV, Rogers JM, Gisselbrecht SS, Rossin EJ, Woodard J, Mariani L, Kock KH, Inukai S, et al. 2016. Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science 351:1450.

Barrientos A, Barros MH, Valnot I, Rötig A, Rustin P, Tzagoloff A. 2002. Cytochrome oxidase in health and disease. Gene 286:53–63.

Batchelor AH, Piper DE, de la Brousse FC, McKnight SL, Wolberger C. 1998. The structure of GABPalpha/beta: an ETS domain- ankyrin repeat heterodimer bound to DNA. Science 279:1037.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. Am. J. Hum. Genet. 74:1111–1120.

Berto S, Perdomo-Sabogal A, Gerighausen D, Qin J, Nowick K. 2016. A consensus network of gene regulatory factors in the human frontal lobe. Front. Genet. [Internet] 7. Available from: http://www.frontiersin.org/bioinformatics_and_computational_biology/10.3389/fgene.2016.00031/abstract

Biswas S, Akey JM. 2006. Genomic insights into positive selection. Trends Genet. TIG 22:437–446.

Blanchette M, Green ED, Miller W, Haussler D. 2004. Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res. 14:2412–2423.

Bogan JS. 2012. Regulation of Glucose Transporter Translocation in Health and Diabetes. Annu. Rev. Biochem. 81:507–532.

Bolstad BM, Irizarry R., Åstrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–193.

Bornberg-Bauer E, Huylmans A-K, Sikosek T. 2010. How do new proteins arise? Curr. Opin. Struct. Biol. 20:390–396.

Bräuer G, Mbua E, Yokoyama Y, Falguères C. 1997. Modern human origins backdated. Nature 386:337–338.

Brayer KJ, Kulshreshtha S, Segal DJ. 2008. The Protein-Binding Potential of C2H2 Zinc Finger Domains. Cell Biochem. Biophys. 51:9–19.

Briers S, Crawford C, Bickmore WA, Sutherland HG. 2009. KRAB zinc-finger proteins localise to novel KAP1-containing foci that are adjacent to PML nuclear bodies. J. Cell Sci. 122:937–946.

Britten RJ, Davidson EH. 1971. Repetitive and Non-Repetitive DNA Sequences and a Speculation on the Origins of Evolutionary Novelty. Q. Rev. Biol. 46:111–138.

Brivanlou AH, Darnell JE. 2002. Signal Transduction and the Control of Gene Expression. Science 295:813–818.

Buchner DA, Charrier A, Srinivasan E, Wang L, Paulsen MT, Ljungman M, Bridges D, Saltiel AR. 2015. Zinc finger protein 407 (ZFP407) regulates insulin-stimulated glucose uptake and glucose transporter 4 (Glut4) mRNA. J. Biol. Chem. 290:6376–6386.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157.

Cheatle Jarvela AM, Hinman VF. 2015. Evolution of transcription factor function as a mechanism for changing metazoan developmental gene regulatory networks. EvoDevo 6:3.

Cheng M, LIU X, YANG M, HAN L, XU A, HUANG Q. 2016. Computational analyses of type 2 diabetes-associated loci identified by Genome-wide association studies. J. Diabetes:n/a – n/a.

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

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. Nat. Rev. Genet. 14:645.

Collins PJ, Kobayashi Y, Nguyen L, Trinklein ND, Myers RM. 2007. The ets-Related Transcription Factor GABP Directs Bidirectional Transcription: e208. PLoS Genet. 3.

Corsinotti A, Kapopoulou A, Gubelmann C, Imbeault M, Santoni de Sio FR, Rowe HM, Mouscaz Y, Deplancke B, Trono D. 2013. Global and stage specific patterns of Krüppel-associated-box zinc finger protein gene expression in murine early embryonic cells. PLoS ONE 8:e56721.

Crew AJ, Gill S, Cooper CS, Gusterson BA, Chan AM-L, Rocques PJ, Shipley J, Clark J. 1994. Identification of novel genes, SYT and SSX , involved in the t(X;18)(p11.2;q11.2) translocation found in human synovial sarcoma. Nat. Genet. 7:502–508.

Desper R, Gascuel O. 2004. Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol. Biol. Evol. 21:587–598.

Diallo AB, Makarenkov V, Blanchette M. 2007. Exact and heuristic algorithms for the Indel Maximum Likelihood Problem. J. Comput. Biol. J. Comput. Mol. Cell Biol. 14:446–461.

Dixit M, Ansseau E, Tassin A, Winokur S, Shi R, Qian H, Sauvage S, Mattéotti C, van Acker AM, Leo O, et al. 2007. DUX4, a Candidate Gene of Facioscapulohumeral Muscular Dystrophy, Encodes a Transcriptional Activator of PITX1. Proc. Natl. Acad. Sci. U. S. A. 104:18157–18162.

Dobzhansky T. 1941. Genetics and the origin of species. New York: Columbia University Press

Durrani K, Papaliodis GN. 2008. The Genetics of Adamantiades-Behcet's Disease. Semin. Ophthalmol. 23:73–79.

Edfeldt K, Ahmad T, Åkerström G, Janson ET, Hellman P, Stålberg P, Björklund P, Westin G. 2014. TCEB3C a putative tumor suppressor gene of small intestinal neuroendocrine tumors. Endocr. Relat. Cancer 21:275.

Eldredge N. 1989. Macroevolutionary dynamics: species, niches, and adaptive peaks. New York: McGraw-Hill Available from:

Enard W, Przeworski M, Fisher SE, Lai CSL, Wiebe V, Kitano T, Monaco AP, Pääbo S. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869–872.

Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ. 2006. The HUGO Gene Nomenclature Database, 2006 updates. Nucleic Acids Res. 34:D319–D321.

Fay JC, Wu C-I. 2000. Hitchhiking Under Positive Darwinian Selection. Genetics 155:1405–1413.

Fisher SE, Tyler-Smith C, Ayub Q, Chen Y, Vernes SC, Yngvadottir B, Xue YL, Hu M. 2013. FOXP2 Targets Show Evidence of Positive Selection in European Populations. Am. J. Hum. Genet. 92:696–706.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. Nucleic Acids Res. 42:D749–D755.

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, et al. 2004. Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol. 2:E207.

Frézal J. 1998. Genatlas database, genes and development defects. Comptes Rendus Académie Sci. Sér. III Sci. Vie 321:805–817.

Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, Bollongino R, Sun C, Giemsch L, Schmitz R, Burger J, et al. 2013. A revised timescale for human evolution based on ancient mitochondrial genomes. Curr. Biol. CB 23:553–559.

Galas DJ, Schmitz A. 1978. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 5:3157–3170.

Gautier L, Cope L, Bolstad BM, Irizarry RA. 2004. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinforma. Oxf. Engl. 20:307–315.

Geschwind DH, Konopka G. 2012. Neuroscience: Genes and human brain evolution. Nature 486:481.

Gibbons A. 2011. A New View Of the Birth of Homo sapiens. Science 331:392–394.

Gillespie JH. 1991. The causes of molecular evolution. New York: Oxford University

Green RE, Fritz MH-Y, Hansen NF, Durand EY, Malaspinas A-S, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, et al. 2010. A draft sequence of the Neandertal genome. Science 328:710–722.

Grossman SR, Wong SH, Cabili M, Adegbola RA, Bamezai RNK, Hill AVS, Vannberg FO, Rinn JL, Lander ES, Schaffner SF, et al. 2013. Identifying recent adaptations in large-scale genomic data. Cell 152:703–713.

Guo A, Nie F, Wong-Riley M. 2000. Human nuclear respiratory factor 2α subunit cDNA: Isolation, subcloning, sequencing, and in situ hybridization of transcripts in normal and monocularly deprived macaque visual system. J. Comp. Neurol. 417:221–232.

Hamilton AT, Huntley S, Tran-Gyamfi M, Baggott DM, Gordon L, Stubbs L. 2006. Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. Genome Res. 16:584–594.

Hardy K, Brand-Miller J, Brown KD, Thomas MG, Copeland L. 2015. The Importance of Dietary Carbohydrate in Human Evolution. Q. Rev. Biol. 90:251–268.

Harris SE, Fox H, Wright AF, Hayward C, Starr JM, Whalley LJ, Deary IJ. 2007. A genetic association analysis of cognitive ability and cognitive ageing using

325 markers for 109 genes associated with oxidative stress or cognition. BMC Genet. 8:43–43.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Hatta M, Cirillo LA. 2007. Chromatin Opening and Stable Perturbation of Core Histone:DNA Contacts by FoxO1. J. Biol. Chem. 282:35583–35593.

Ho KK, Myatt SS, Lam EW-F. 2008. Many forks in the path: cycling with FoxO. Oncogene 27:2300–2311.

Huby T, Dachet C, Lawn RM, Wickings J, Chapman MJ, Thillet J. 2001. Functional analysis of the chimpanzee and human apo(a) promoter sequences: identification of sequence variations responsible for elevated transcriptional activity in chimpanzee. J. Biol. Chem. 276:22209–22214.

Huff CD, Harpending HC, Rogers AR. 2010. Detecting positive selection from genome scans of linkage disequilibrium. BMC Genomics 11:8–8.

Hughes AL, Yeager M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. Annu. Rev. Genet. 32:415–435.

Hughes TR. 2011. Introduction to "A Handbook of Transcription Factors." In: Hughes TR, editor. A Handbook of Transcription Factors. Vol. 52. Subcellular Biochemistry. Springer Netherlands. p. 1–6. Available from: http://dx.doi.org/10.1007/978-90-481-9069-0_1

Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, Kim J, Gordon L, Branscomb E, Stubbs L. 2006. A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res. 16:669–677.

Iyengar S, Farnham PJ. 2011. KAP1 protein: an enigmatic master regulator of the genome. J. Biol. Chem. 286:26267–26276.

Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature 516:242–245.

Jaworski A, Smith CL, Burden SJ. 2007. GA-binding protein is dispensable for neuromuscular synapse formation and synapse-specific gene expression. Mol. Cell. Biol. 27:5040–5046.

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

Jonnalagadda M, Norton H, Ozarkar S, Kulkarni S, Ashma R. 2016. Association of genetic variants with skin pigmentation phenotype among populations of west Maharashtra, India. Am. J. Hum. Biol.:n/a – n/a.

Kamagate A, Qu S, Perdomo G, Su D, Kim DH, Slusher S, Meseck M, Dong HH. 2008. FoxO1 mediates insulin-dependent regulation of hepatic VLDL production in mice. J. Clin. Invest. 118:2347–2364.

Kambouris M, Maroun RC, Ben-Omran T, Al-Sarraj Y, Errafii K, Ali R, Boulos H, Curmi PA, El-Shanti H. 2014. Mutations in zinc finger 407 [ZNF407] cause a unique autosomal recessive cognitive impairment syndrome. Orphanet J. Rare Dis. 9:80–80.

Karlsson EK, Kwiatkowski DP, Sabeti PC. 2014. Natural selection and infectious disease in human populations. Nat. Rev. Genet. 15:379–393.

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. 2014. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 42:D764–D770.

Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217:624–626.

Kim Y, Nielsen R. 2004. Linkage Disequilibrium as a Signature of Selective Sweeps. Genetics 167:1513–1524.

Kim Y, Stephan W. 2002. Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome. Genetics 160:765–777.

King M-C, Wilson AC. 1975. Evolution at Two Levels in Humans and Chimpanzees. Science 188:107–116.

Konopka G, Geschwind DH, Bomar JM, Winden K, Coppola G, Jonsson ZO, Gao F, Peng S, Preuss TM, Wohlschlegel JA. 2009. Human-specific transcriptional regulation of CNS development genes by FOXP2. Nature 462:213–217.

Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R. 2013. Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. BMC Bioinformatics 14:289–289.

Kuhlwilm M, Gronau I, Hubisz MJ, de Filippo C, Prado-Martinez J, Kircher M, Fu Q, Burbano HA, Lalueza-Fox C, de la Rasilla M, et al. 2016. Ancient gene flow from early modern humans into Eastern Neanderthals. Nature [Internet] advance online publication. Available from: http://dx.doi.org/10.1038/nature16544

Kumar RA, Karamohamed S, Sutcliffe JS, Cook EH, Geschwind DH, Dobyns WB, Scherer SW, Christian SL, Marshall CR, Badner JA, et al. 2009. Association and mutation analyses of 16p11.2 autism candidate genes. PloS One 2009 42e4582 4:e4582.

Lai CSL, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413:519–523.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22:1813–1831.

Lauterbach EC. 2013. Psychotropics regulate Skp1a, Aldh1a1, and Hspa8 transcription — Potential to delay Parkinson's disease. Prog. Neuropsychopharmacol. Biol. Psychiatry 40:236–239.

Lee L-C, Weng Y-T, Wu Y-R, Soong B-W, Tseng Y-C, Chen C-M, Lee-Chen G-J. 2014. Downregulation of proteins involved in the endoplasmic reticulum stress response and Nrf2-ARE signaling in lymphoblastoid cells of spinocerebellar ataxia type 17. J. Neural Transm. 121:601–610.

Lee MS, Gippert GP, Soman KV, Case DA, Wright PE. 1989. Three-Dimensional Solution Structure of a Single Zinc Finger DNA-Binding Domain. Science 245:635–637.

Lin JM, Collins PJ, Trinklein ND, Fu Y, Xi H, Myers RM, Weng Z. 2007. Transcription factor binding and modified histones in human bidirectional promoters. Genome Res. 17:818–827.

Li Y, Park J-S, Deng J-H, Bai Y. 2006. Cytochrome c oxidase subunit IV is essential for assembly and respiratory function of the enzyme complex. J. Bioenerg. Biomembr. 38:283–291.

Locke AE, Goel A, Ehret GB, Winkler TW, Schmidt EM, Yengo L, Demirkan A, Vedantam S, Berndt SI, Bragg-Gresham JL, et al. 2015. Genetic studies of body mass index yield new insights for obesity biology. Nature 518:197–206.

Lukic S, Nicolas JC, Levine AJ. 2014. The diversity of zinc-finger genes on human chromosome 19 provides an evolutionary mechanism for defense against inherited endogenous retroviruses. Cell Death Differ. 21:381–387.

Mahony S, Corcoran DL, Feingold E, Benos PV. 2007. Regulatory conservation of protein coding and microRNA genes in vertebrates: lessons from the opossum genome. Genome Biol. 8:R84–R84.

Marchetti P, Bugliani M, Lupi R, Marselli L, Masini M, Boggi U, Filipponi F, Weir GC, Eizirik DL, Cnop M. 2007. The endoplasmic reticulum in pancreatic beta cells of type 2 diabetes patients. Diabetologia 50:2486–2494.

Maricic T, Lalueza-Fox C, de la Rasilla M, Rosas A, Gajovic S, Kelso J, Enard W, Schaffner W, Pääbo S, Günther V, et al. 2013. A recent evolutionary change affects a regulatory element in the human FOXP2 gene. Mol. Biol. Evol. 30:844–852.

Masson G, Gylfason A, Gudjonsson SA, Thorleifsson G, Jonasdottir A, Jonasdottir A, Sigurdsson A, Frigge ML, Walters GB, Helgason A, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467:1099–1103.

McGowen MR, Gatesy J, Wildman DE. 2014. Molecular evolution tracks macroevolutionary transitions in Cetacea. Trends Ecol. Evol. 29:336–346.

Mellars P, Gori KC, Carr M, Soares PA, Richards MB. 2013. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. Proc. Natl. Acad. Sci. U. S. A. 110:10699–10704.

Messer PW, Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol. Evol. 28:659.

Messina DN, Glasscock J, Gish W, Lovett M. 2004. An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. Genome Res. 14:2041–2047.

Metspalu M, Remm M, Pitchappan R, Singh L, Thangaraj K, Villems R, Kivisild T, Romero IG, Yunusbayev B, Chaubey G, et al. 2011. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. Am. J. Hum. Genet. 89:731–744.

Meyer G, Socorro AC, Garcia CGP, Millan LM, Walker N, Caput D. 2004. Developmental Roles of p73 in Cajal-Retzius Cells and Cortical Patterning. J. Neurosci. 24:9878–9887.

Miyamoto T, Koh E, Sakugawa N, Sato H, Hayashi H, Namiki M, Sengoku K. 2008. Two single nucleotide polymorphisms in PRDM9 (MEISETZ) gene may be a genetic risk factor for Japanese patients with azoospermia by meiotic arrest. J. Assist. Reprod. Genet. 25:553–557.

Mueller KL, Murray JC, Michaelson JJ, Christiansen MH, Reilly S, Tomblin JB. 2016. Common Genetic Variants in FOXP2 Are Not Associated with Individual Differences in Language Development. PLoS One 11:e0152576.

Mussi P, Liao L, Park S-E, Ciana P, Maggi A, Katzenellenbogen BS, Xu J, O'Malley BW. 2006. Haploinsufficiency of the Corepressor of Estrogen Receptor Activity (REA) Enhances Estrogen Receptor Function in the Mammary Gland. Proc. Natl. Acad. Sci. U. S. A. 103:16716–16721.

Naidoo N. 2009. ER and aging—Protein folding and the ER stress response. Ageing Res. Rev. 8:150–159.

Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A, Albu M, Weirauch MT, Radovani E, Kim PM, et al. 2015. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat. Biotechnol. 33:555–562.

Najmabadi H, Jamali P, Zecha A, Mohseni M, Püttmann L, Vahid LN, Jensen C, Moheb LA, Bienek M, Larti F, et al. 2011. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature 478:57–63.

Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, et al. 2009. Darwinian and demographic forces affecting human protein coding genes. Genome Res. 19:838–849.

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

Nowick K, Carneiro M, Faria R. 2013. A prominent role of KRAB-ZNF transcription factors in mammalian speciation? Trends Genet. TIG [Internet]. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0168952512001916

Nowick K, Fields C, Gernat T, Caetano-Anolles D, Kholina N, Stubbs L. 2011. Gain, loss and divergence in primate zinc-finger genes: a rich resource for evolution of gene regulatory differences between species. PLoS ONE 6:e21553.

Nowick K, Gernat T, Almaas E, Stubbs L, Robinson GE. 2009. Differences in Human and Chimpanzee Gene Expression Patterns Define an Evolving Network of Transcription Factors in Brain. Proc. Natl. Acad. Sci. U. S. A. 106:22358–22363.

Nowick K, Hamilton AT, Zhang H, Stubbs L. 2010. Rapid sequence and expression divergence suggest selection for novel function in primate-specific KRAB-ZNF genes. Mol. Biol. Evol. 27:2606–2617.

Odrowaz Z, Sharrocks AD. 2012. The ETS transcription factors ELK1 and GABPA regulate different gene networks to control MCF10A breast epithelial cell migration. PloS One 7:e49892.

Oppenheimer S. 2009. The great arc of dispersal of modern humans: Africa to Australia. Quat. Int. 202:2–13.

Park SG, Schimmel P, Kim S. 2008. Aminoacyl tRNA Synthetases and Their Connections to Disease. Proc. Natl. Acad. Sci. U. S. A. 105:11043–11049.

Parmacek MS. 2007. Myocardin-Related Transcription Factors: Critical Coactivators Regulating Cardiovascular Development and Adaptation. Circ. Res. 100:633–644.

Patillon B, Luisi P, Blanché H, Patin E, Cann HM, Génin E, Sabbagh A. 2012. Positive Selection in the Chromosome 16 VKORC1 Genomic Region Has Contributed to the Variability of Anticoagulant Response in Humans: e53049. PLoS One

Perdomo-Sabogal A, Kanton S, Walter MBC, Nowick K. 2014. The role of gene regulatory factors in the evolutionary history of humans. Curr. Opin. Genet. Dev. 29:60.

Perdomo-Sabogal A, Nowick K, Piccini I, Sudbrak R, Lehrach H, Yaspo M-L, Warnatz H-J, Querfurth R. 2016. Human lineage-specific transcriptional regulation through GA binding protein transcription factor alpha (GABPa). Mol. Biol. Evol. [Internet]. Available from: http://mbe.oxfordjournals.org/content/early/2016/01/21/molbev.msw007.abstract

Pickrell JK, Feldman MW, Pritchard JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 19:826–837.

Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI. 2015. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database J. Biol. Databases Curation 2015:bav028.

Pollard KS, Dudoit S, van der Laan MJ. 2005. Multiple Testing Procedures: the multtest Package and Applications to Genomics. In: Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S, editors. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health. Springer New York. p. 249–271. Available from: http://dx.doi.org/10.1007/0-387-29362-0_15

Pozniak CD, Barnabe-Heider F, Rymar VV, Lee AF, Sadikot AF, Miller FD. 2002. p73 Is Required for Survival and Maintenance of CNS Neurons. J. Neurosci. 22:9800.

Prakash A, Tompa M. 2007. Measuring the accuracy of genome-size multiple alignments. Genome Biol. 8:R124–R124.

Pritchard JK, Pickrell JK, Coop G. 2010. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Curr. Biol. 20:R208–R215.

Prüfer K, Muetzel B, Do H-H, Weiss G, Khaitovich P, Rahm E, Pääbo S, Lachmann M, Enard W. 2007. FUNC: a package for detecting significant associations between gene sets and ontological annotations. BMC Bioinformatics 8:41–41.

Puigserver P, Dominy JE. 2010. Nuclear FoxO1 inflames insulin resistance. EMBO J. 29:4068–4069.

Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, Pavlidis P, Laayouni H, Bertranpetit J, Engelken J. 2013. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res. [Internet]. Available from: http://nar.oxfordjournals.org/content/early/2013/11/24/nar.gkt1188.abstract

Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, Engelken J. 2015. Hierarchical boosting: a machine-learning framework to detect and

classify hard selective sweeps in human populations. Bioinformatics:btv493.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Ralser M, Querfurth R, Warnatz H-J, Lehrach H, Yaspo M-L, Krobitsch S. 2006. An efficient and economic enhancer mix for PCR. Biochem. Biophys. Res. Commun. 347:747–751.

Ravasi T, Bertin N, Carninci P, Daub CO, Forrest ARR, Gough J, Grimmond S, Han J-H, Hashimoto T, Hide W, et al. 2010. An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. Cell 140:744–752.

Reich D, Johnson PLF, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060.

Reimers-Kipping S, Hevers W, Pääbo S, Enard W. 2011. Humanized Foxp2 specifically affects cortico-basal ganglia circuits. Neuroscience 175:75–84.

Reyes-Centeno H. 2016. Out of Africa and into Asia: Fossil and genetic evidence on modern human origins and dispersals. Quat. Int.

Reznick DN, Ricklefs RE. 2009. Darwin's bridge between microevolution and macroevolution. Nature 457:837–842.

Ripperger T, Manukjan G, Meyer J, Wolter S, Schambach A, Bohne J, Modlich U, Li Z, Skawran B, Schlegelberger B, et al. 2015. The heteromeric transcription factor GABP activates the ITGAM/CD11b promoter and induces myeloid differentiation. Biochim. Biophys. Acta BBA - Gene Regul. Mech. 1849:1145–1154.

Ristevski S, O'Leary DA, Thornell AP, Owen MJ, Kola I, Hertzog PJ. 2004. The ETS Transcription Factor GABPα Is Essential for Early Embryogenesis. Mol. Cell. Biol. 24:5844–5849.

Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, Wray GA. 2005. Ancient and recent positive selection transformed opioid cis-regulation in humans. PLoS Biol. 3:e387–e387.

Romanelli MG, Lorenzi P, Sangalli A, Diani E, Mottes M. 2009. Characterization and functional analysis of cis-acting elements of the human farnesyl diphosphate synthetase (FDPS) gene 5' flanking region. Genomics 93:227–234.

Rose JE, Behm FM, Drgon T, Johnson C, Uhl GR. 2010. Personalized smoking cessation: interactions between nicotine dose, dependence and quit-success genotype score. Mol. Med. Camb. Mass 16:247–253.

Rosmarin AG, Resendes KK, Yang Z, McMillan JN, Fleming SL. 2004. GA-binding protein transcription factor: a review of GABP as an integrator of intracellular signaling and protein-protein interactions. Blood Cells. Mol. Dis. 32:143–154.

Ross KA, Bigham AW, Edwards M, Gozdzik A, Suarez-Kurtz G, Parra EJ. 2010. Worldwide allele frequency distribution of four polymorphisms associated with warfarin dose requirements. J. Hum. Genet. 55:582–589.

Ryan RF, Darby MK. 1998. The role of zinc finger linkers in p43 and TFIIIA binding to 5S rRNA and DNA. Nucleic Acids Res. 26:703–709.

Ryan RF, Schultz DC, Ayyanathan K, Singh PB, Friedman JR, Fredericks WJ, Frank J. Rauscher III. 1999. KAP-1 Corepressor Protein Interacts and Colocalizes with Heterochromatic and Euchromatic HP1 Proteins: a Potential Role for Krüppel-Associated Box–Zinc Finger Proteins in Heterochromatin-Mediated Gene Silencing. Mol. Cell. Biol. 19:4366–4378.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918.

Schwartz JJ, Roach DJ, Thomas JH, Shendure J. 2014. Primate evolution of the recombination regulator PRDM9. Nat. Commun. 5:4370.

Seltsam A, Hallensleben M, Kollmann A, Blasczyk R. 2003. The nature of diversity and diversification at the ABO locus. Blood 102:3035–3042.

Sesink Clee P, Gonder MK. 2012. Macroevolution: Examples from the Primate World. Nat. Educ. Knowl. 3(12):2.

Sheng B, Wang X, Su B, Lee H, Casadesus G, Perry G, Zhu X. 2012. Impaired mitochondrial biogenesis contributes to mitochondrial dysfunction in Alzheimer's disease. J. Neurochem. 120:419–429.

Sissler M, Andel RJ van, Scheper GC, Krageloh-Mann I, Uziel G, Coster R van, Muravina TI, Bugiani M, Pronk JC, Florentz C, et al. 2007. Mitochondrial aspartyl-tRNA synthetase deficiency causes leukoencephalopathy with brain stem and spinal cord involvement and lactate elevation. Nat. Genet. 39:534–539.

Soejima M, Koda Y. 2007. Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. Int. J. Legal Med. 121:36–39.

Sofou K, Kollberg G, Holmström M, Dávila M, Darin N, Gustafsson CM, Holme E, Oldfors A, Tulinius M, Asin-Cayuela J. 2015. Whole exome sequencing reveals mutations in NARS2 and PARS2, encoding the mitochondrial asparaginyl-tRNA synthetase and prolyl-tRNA synthetase, in patients with Alpers syndrome. Mol. Genet. Genomic Med. 3:59–68.

Stephan W, Song YS, Langley CH. 2006. The Hitchhiking Effect on Linkage Disequilibrium Between Linked Neutral Loci. Genetics 172:2647–2663.

Stewart AJ, Hannenhalli S, Plotkin JB. 2012. Why transcription factor binding sites are ten nucleotides long. Genetics 192:973–985.

Stringer CB, Andrews P. 1988. Genetic and Fossil Evidence for the Origin of Modern Humans. Science 239:1263–1268.

Stubbs L, Sun Y, Caetano-Anolles D. 2011. Function and Evolution of C2H2 Zinc Finger Arrays. In: Hughes RT, editor. A Handbook of Transcription Factors. Dordrecht: Springer Netherlands. p. 75–94. Available from: http://dx.doi.org/10.1007/978-90-481-9069-0_4

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Project 1000 Genomes, et al. 2010. Diversity of Human Copy Number Variation and Multicopy Genes. Science 330:641–646.

Sultan M, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo M-L, Schulz MH, Richard H, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956–960.

Tadepally HD, Burger G, Aubry M. 2008. Evolution of C2H2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains. BMC Evol. Biol. 8:176–176.

Taft RJ, Vanderver A, Leventer RJ, Damiani SA, Simons C, Grimmond SM, Miller D, Schmidt J, Lockhart PJ, Pope K, et al. 2013. Mutations in DARS cause hypomyelination with brain stem and spinal cord involvement and leg spasticity. Am. J. Hum. Genet. 92:774–780.

Tajima F. 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123:585–595.

Teichert M, Eijgelsheim M, Visser LE, Smet PA de, Uitterlinden AG, Hofman A, Buhre PN, Stricker BHC. 2011. Dependency of phenprocoumon dosage on polymorphisms in the VKORC1, CYP2C9, and CYP4F2 genes. Pharmacogenet. Genomics 21:26–34.

The International HapMap Consortium, International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

The UniProt Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43:D204–D212.

Thomas JH, Emerson RO, Shendure J. 2009. Extraordinary molecular evolution in the PRDM9 fertility gene. PloS One 4:e8505.

Tripathi S, Christie KR, Balakrishnan R, Huntley R, Hill DP, Thommesen L, Blake JA, Kuiper M, Lægreid A. 2013. Gene Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large-scale curation effort. Database 2013 Art No Bat062 2013:bat062.

Tryon C, Bailey S. 2013. Testing Models of Modern Human Origins with Archaeology and Anatomy. Nat. Educ. Knowl. 4(3).

Valentine JW, Jablonski D. 2003. Morphological and developmental macroevolution: a paleontological perspective. Int. J. Dev. Biol. 47:517.

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Meth 5:829–834.

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. 2009. A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. 10:252–263.

Vasseur E, Quintana-Murci L. 2013. The impact of natural selection on health and disease: uses of the population genetics approach in humans. Evol. Appl. 6:596–607.

Veeramah KR, Hammer MF. 2014. The impact of whole-genome sequencing on the reconstruction of human population history. Nat Rev Genet 15:149–162.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting Natural Selection in Genomic Data. Annu. Rev. Genet. 47:97–120.

Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. 2004. Structure, function and evolution of multidomain proteins. Curr. Opin. Struct. Biol. 14:208–216.

Voss TC, Hager GL. 2014. Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat Rev Genet 15:69–81.

Wägner AM, Cloos P, Bergholdt R, Boissy P, Andersen TL, Henriksen DB, Christiansen C, Christgau S, Pociot F, Nerup J, et al. 2007. Post-translational protein modifications in type 1 diabetes: a role for the repair enzyme protein-l-isoaspartate (d-aspartate) O-methyltransferase? Diabetologia 50:676–681.

Wägner AM, Cloos P, Bergholdt R, Eising S, Brorsson C, Stalhut M, Christgau S, Nerup J, Pociot F. 2008. Posttranslational Protein Modifications in Type 1 Diabetes - Genetic Studies with PCMT1, the Repair Enzyme Protein Isoaspartate Methyltransferase (PIMT) Encoding Gene. Rev. Diabet. Stud. RDS 5:225–231.

Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. 2005. SVA elements: a hominid-specific retroposon family. J. Mol. Biol. 354:994–1007.

Wang J, Zhuang J, Iyer S, Lin X-Y, Greven MC, Kim B-H, Moore J, Pierce BG, Dong X, Virgil D, et al. 2013. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res. 41:D171.

Warnatz H-J, Vingron M, Lehrach H, Yaspo M-L, Schmidt D, Manke T, Piccini I, Sultan M, Borodina T, Balzereit D, et al. 2011. The BTB and CNC homology 1 (BACH1) target genes are involved in the oxidative stress response and in control of the cell cycle. J. Biol. Chem. 286:23521–23532.

Weirauch M, Hughes TR. 2011. A Catalogue of Eukaryotic Transcription Factor Types, Their Evolutionary Origin, and Species Distribution. In: Hughes TR, editor. A Handbook of Transcription Factors. Vol. 52. Subcellular Biochemistry. Springer Netherlands. p. 25–73. Available from: http://dx.doi.org/10.1007/978-90-481-9069-0_3

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution 38:1358–1370.

Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al. 2014. Mass-spectrometry-based draft of the human proteome. Nature 509:582–587.

Wingender E, Schoeps T, Dönitz J. 2013. TFClass: an expandable hierarchical classification of human transcription factors. Nucleic Acids Res. 41:D165–D170.

Wingender E, Schoeps T, Haubrock M, Donitz J. 2015. TFClass: a classification of human transcription factors and their rodent orthologs. Nucleic Acids Res. 43:D97–D102.

Wingender E, Schoeps T, Haubrock M, Dönitz J. 2015. TFClass: a classification of human transcription factors and their rodent orthologs. Nucleic Acids Res. 43:D97–D102.

Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat. Rev. Genet. 13:59.

Wolfe SA, Nekludova L, Pabo CO. 2000. DNA RECOGNITION BY Cys2His2 ZINC FINGER PROTEINS. Annu. Rev. Biophys. Biomol. Struct. 29:183–212.

Wollstein A, Stephan W. 2015. Inferring positive selection in humans from genomic data. Investig. Genet. 6:5.

Wolpoff MH, Hawks J, Caspari R. 2000. Multiregional, not multiple origins. Am. J. Phys. Anthropol. 112:129–136.

Wolpoff MH, Spuhler JN, Smith FH, Radovcic J, Pope G, Frayer DW, Eckhardt R, Clark G. 1988. Modern human origins. Science 241:772–774.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. 8:206–216.

Wu Z, Fei A, Liu Y, Pan S. 2016. Conditional Tissue-Specific Foxa2 Ablation in Mouse Pancreas Causes Hyperinsulinemic Hypoglycemia. Am. J. Ther.:1.

Xie Y, Larsson O, Skytting B, Nilsson G, Grimer RJ, Mangham CD, Fisher C, Shipley J, Bjerkehagen B, Myklebost O. 2002. The SYT-SSX1 fusion type of synovial sarcoma is associated with increased expression of cyclin A and D1. A link between t(X;18)(p11.2; q11.2) and the cell cycle machinery. Oncogene 21:5791–5796.

Yalley A, Schill D, Hatta M, Johnson N, Cirillo LA. 2016. Loss of Interdependent Binding by the FoxO1 and FoxA1/A2 Forkhead Transcription Factors Culminates in Perturbation of Active Chromatin Marks and Binding of Transcriptional Regulators at Insulin Sensitive Genes. J. Biol. Chem. [Internet]. Available from: http://www.jbc.org/content/early/2016/02/29/jbc.M115.677583.abstract

Yang A, Walker N, Bronson R, Kaghad M, Oosterwegel M, Bonnin J, Vagner C, Bonnet H, Dikkes P, Sharpe A, et al. 2000. p73-deficient mice have neurological, pheromonal and inflammatory defects but lack spontaneous tumours. Nature 404:99–103.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang Z-F, Drumea K, Cormier J, Wang J, Zhu X, Rosmarin AG. 2011. GABP transcription factor is required for myeloid differentiation, in part, through its control of Gfi-1 expression. Blood 118:2243–2253.

Yang Z-F, Drumea K, Mott S, Wang J, Rosmarin AG. 2014. GABP transcription factor (nuclear respiratory factor 2) is required for mitochondrial biogenesis. Mol. Cell. Biol. 34:3194–3201.

Yawata T, Maeda Y, Okiku M, Ishida E, Ikenaka K, Shimizu K. 2011. Identification and functional characterization of glioma-specific promoters and their application in suicide gene therapy. J. Neurooncol. 104:497–507.

Yazdi FT, Clee SM, Meyre D. 2015. Obesity genetics in mouse and human: back and forth, and back again. PeerJ 3:e856.

Yu M, Yang XY, Schmidt T, Chinenov Y, Wang R, Martin ME. 1997. GA-binding protein-dependent transcription initiator elements. Effect of helical spacing between polyomavirus enhancer a factor 3(PEA3)/Ets-binding sites on initiator activity. J. Biol. Chem. 272:29060–29067.

Zhang J, Webb DM, Podlaha O. 2002. Accelerated Protein Evolution and Origins of Human-Specific Features: FOXP2 as an Example. Genetics 162:1825–1835.

Zhang W, Edwards A, Deininger P, Zhang K. 2015. The Duplication and Intragenic Domain Expansion of Human C2H2 Zinc Finger Genes Are Associated with Transposable Elements and Relevant to the Expression-based Clustering. Bioinforma. Comput. Biol. BICOM-2015 ISCA 7th Int. Conf. Bioinforma. Comput. Biol.:7.

Zhang YE, Landback P, Vibranovski MD, Long M. 2011. Accelerated recruitment of new brain development genes into the human genome. PLoS Biol. 9:e1001179.

Zhang YE, Vibranovski MD, Landback P, Marais GAB, Long M. 2010. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. PLoS Biol. 8:e1000494.

Curriculum Vitae *Alvaro Perdomo-Sabogal*

Address

Härtelstrasse 16-18, D-04107
Leipzig, Germany
e-mail: alvaro@bioinf.uni-leipzig.de

Education:

- March 2013 – until present. PhD candidate, Bioinformatics Department, Institute for Computer Science, Universität Leipzig, Germany. Project: Recent evolutionary changes in transcription factor genes (TFs) between primate species and human populations. Supervisor: Dr. Katja Nowick.
- February 2005 – May 2008. Master in Science: Biology-Genetics. National University of Colombia. Project: *Genetic variability of the endangered catfish Pseudoplatystoma magdaleniatum (Siluriformes: Pimelodidae), in Colombia using microsatellite markers*. Supervisor: Dr. Professor. Consuelo Burbano-Montenegro
- February 1998 – December 2002. B.Sc. Biology and Environmental Education. *Honours*. Project: Ecología trófica y reproductiva de A*rgopleura magdalenensis* (Pisces: Characidae), en la cuenca alta de los ríos Cauca y Magdalena. Supervisor: Dr. Professor. Cesar Roman-Valencia

Research interest

Human and primate evolution, population genetics/genomics, genomics, transcriptomics, immunogenetics, conservation biology, evolutionary biology.

Professional experience:

~ ***Research positions:***

- 2013 – until present: Doctoral research. Bioinformatics Department, Institute for Computer Science, Universität Leipzig, Germany. Main tasks and achievements:
  * Built the most complete catalog of Transcription Factor genes in the human genome
  * Population genetics analyses of TF genes in Humans
  * Identification of Human and primate specific binding sites for GA binding protein transcription factor alpha GABPa by using ChIP-Seq data.
  * Comparative analyses of KAP1 gene regulation in Human, Chimpanzee and Orangutan.
- 2011 – 2013. Research Assistantship. Project: Genomic Organization and Evolution of the Major Histocompatibility Complex of Suidae and Tayassuidae (wild pigs and peccaries). Faculty of Veterinary Science, The University of Sydney, Sydney, NSW, Australia. Main tasks and achievements:
  * Analyzed NGS data from the heterologous capture.
  * Built a consensus sequence for the MHC of each one of the 11 species.

    - Predicted gene annotation and performed some phylogenetic analyses.
- 2005 - 2009. Research assistantship at the biology department, Conservation and populations genetics research group from National University of Colombia. Project: Genetic Characterization of six ichthyic species from San Jorge. Main tasks and achievements:
    - Protocols standardization: Samples preservation, DNA extraction, purification, quantification, PCR, genotyping, among other wet lab tasks.
    - Primers design and testing for amplifying and characterizing Microsatellites markers in six species of endangered fish from Colombia.
    - Characterization of the genetic variability of the endangered catfish *Pseudoplatystoma magdaleniatum* in the Magdalena Basin, Colombia.
- 2000 – 2003. Research Assistant. Ichthyology (Taxonomy and Ecology): Main tasks:
    - Samples collection and preservation.
    - Taxonomic classification of aquatic invertebrates that make part of the diet of fish.
    - Estimation of the seasonal reproductive periods of fish.

~ ***Teaching assistantships***

- Teaching assistant: Conservation genetics Lecture. First semester 2012. Faculty of Veterinary Science, The University of Sydney. Main tasks:
    - To assist the students during the wet lab practices: DNA extraction, purification and quantification.
    - To assist the students during the phase they were preparing their final project proposal
    - To grade the exams
- Teaching Assistant: 2nd and 3rd Programming for Evolutionary Biology, 2013, 2014. Leipzig, Germany. Main tasks:
    - To assist the students during the different modules of the course, thus providing help with the code, understanding of it and with efficiently develop their own code.
    - To address particular questions and coordinate the assigned final projects.
- Teaching assistantship. Dept. Biology and Environmental sciences. Limnology course. Universidad del Quindío. May 2002 – December 2002. Main tasks:
    - To prepare the wet lab material.
    - To lead and advise the students during the field trips and during the project proposal design process.

~ ***Other skills***

- Bioinformatics:
    - Next-Generation Sequencing data processing and analyzing:
        - ChIP-Seq data (Advanced)
        - DNA-Seq data (Very good)

        - RNA-Seq data (Good)
    - Microarrays data analyses (good).
    - Programming languages:
        - Perl (Advanced)
        - R (good)
        - BioPerl (good)
    - Other programming languages:
        - Bash (Advanced)
        - AWK (Advanced)
        - Sed (Advanced)
    - Genome browsers and databases:
        - UCSC (Advanced)
        - Ensembl and biomart (Advanced)
        - NCBI (Advanced)
        - OMIM (Advanced)
        - Others
- Wet lab:
    - Animal tissue sample manipulation and preservation
    - DNA extraction, purification and quantification (Phenol-choloform extraction, commercial kits)
    - Polymerase Chain Reactions (PCR amplification), standard, gradient and multiplexed.
    - Gel electrophoresis (Agarose, Polyacrylamid)
    - Use of molecular markers (STR, AFLPs, mtDNA)
    - PCR Primer design
- Languages: Spanish (Native); English (As second language)

~ ***Administrative positions***

- April 2009 – January 2011. Head General Office of curricula. The University of La Sabana (Universidad de La Sabana), Bogotá, Colombia.
    - Defining politics in the context of higher education at institutional level, and mainly addressing the integration of research training at graduate and undergraduate level.
- March 2008 – April 2009. Assistant General Office of curricula, The University of La Salle (Universidad de La Salle). Bogotá, Colombia.
    - Re-structuring the curricula design at graduate and undergraduate programs to make it them coherent and pertinent to the educational context of Colombian higher education.
    - Defining politics to integrate research as part of the training within the curricula at graduate and undergraduate level.

Scholarships, research grants, bursaries and prizes

- 2011 – 2015. PhD scholarship from The Administrative Department of Science, Technology and Innovation, Colombia government.
- Travel bursary. 33rd Conference of the International Society for Animal Genetics, Brisbane.

- Best poster prize. 33rd Conference of the International Society for Animal Genetics, Brisbane.
- 2006 – 2008. Genetic Variability of *Pseudoplatystoma magdaleniatum* in Colombia. Republic Bank of Colombia. USD 20.000.
- 2005 – 2008. Genetic Characterization of six ichthyic species from San Jorge. Basin. National Development Fund. USD 125.000.
- 2006 – 2007. Reproductive ecology of *Poecilia* caucana in three natural streams of Colombia. ECOMAMA. USD 10.000

Publications

~ ***Scientific papers***

- Berto S, **Perdomo-Sabogal A**, Gerighausen D, Qin J, Nowick K. 2016. A consensus network of gene regulatory factors in the human frontal lobe. Front. Genet. [7.
- **Perdomo-Sabogal A**, Nowick K, Piccini I, Sudbrak R, Lehrach H, Yaspo M-L, Warnatz H-J, Querfurth R. 2016. Human lineage-specific transcriptional regulation through GA binding protein transcription factor alpha (GABPa). Mol. Biol. Evol. 10.1093/molbev/msw007
- **Alvaro Perdomo – Sabogal**, Sabina Kanton, Maria Beatriz Walter–Costa, Katja Nowick. 2014. The role of gene regulatory factors in the evolutionary history of humans. Article type: Genetics of Human Origins. Current Opinion in Genetics and Development. Volume 29, December 2014, Pages 60–67.
- Lee, C., Chong, A., **Perdomo-Sabogal, A**., Mach, N., Megens, H., Moroldo, M., Marthey, S., Lecardonnel, J., Wahlberg, P., Estelle, J., Groenen, M., Rogel-Gaillard, C., Gongora, J et al, Heterologous capture of suids' and peccaries' MHC, an source for exploring wild pig's immune evolution. *In preparation*.
- Cuartas – Mendez, D and **Perdomo – Sabogal, A**. 2006. Feeding and reproductive ecology of *Xiphophorus helleri* "Exotic" (Pisces: Poeciliidae) from natural stream, Colombia**.** OFI Journal 51: 13 - 17. The Netherlands.
- Roman – Valencia, C., and **Perdomo – Sabogal A**. 2004. Ecología trófica y reproductiva de A*rgopleura magdalenensis* (Pisces: Characidae), en la cuenca alta de los ríos Cauca y Magdalena, Colombia**.** 2004. Rev. Mus. Argentino. Cien. Nat., ns. 6 (1): 177 – 184.
- Chacon, C., Giraldo, A., and **Perdomo – Sabogal, A**. 2001. Papel del educador ambiental frente a los problemas ambientales – caso Armenia. Universidad del Quindio. Facultad de Educación. Libros pedagogicos interdiciplinarios (3):79–86. 2001

~ ***Talks in conferences***

- **Perdomo-Sabogal, A**., Nowick, K., Piccini, Ilaria., Sudbrak, R., Lehrach, Hans., Yaspo, M., Warnatz, HJ., and Querfurth, R. 2015. Human lineage-specific transcriptional regulation through GA binding protein transcription factor alpha (GABPa). 13th Autum seminar de Bioinformatica, Bioinformatics, Leipzig.

- **Perdomo-Sabogal, A**., Engelken, Johannes., and Katja Nowick. 2013. Signatures of Selection on Transcription Factor Genes in three human populations. 11th Autum seminar de Bioinformatica, Bioinformatics, Leipzig.
- **Perdomo-Sabogal, A**., Nowick, K., Piccini, Ilaria., Sudbrak, R., Lehrach, Hans., Yaspo, M., Warnatz, HJ., and Querfurth, R. Perdomo, A., Mach, N., Megens, H., Moroldo, M., Marthey, S., Lecardonnel, J., Wahlberg, P., Estelle, J., Groenen, M., Rogel-Gaillard, C., Gongora, J. 2012. Evolution and diversity of MHC class I and class II genes in suids and peccaries. 33rd Conference of the International Society for Animal Genetics, Brisbane: University of Queensland Press.
- **Alvaro Perdomo-Sabogal**, Nuria Mach, Hendrik-Jan Megens, Marco Moroldo, Sylvain Marthey, Jerome Lecardonnel, Peer Wahlberg, Jordi Estellé, Martien Groenen, Claire Rogel-Gaillard, Jaime Gongora. 2012. Evolution of MHC class I and class II genes in suids and peccaries. Proceedings postgraduate conferences, Faculty of Veterinary Science, The University of Sydney.
- **Perdomo – Sabogal, A** and Burbano, C. 2009. Molecular markers confirm three species taxonomically described as new for the genus Pseudoplatystoma (Siluriformes: Pimelodidae). Actualidades Biológicas. Vol 31, Pag 113.
- **Perdomo – Sabogal, A** and Burbano, C. 2009. Genetic variability of the endangered catfish Pseudoplatystoma magdaleniatum (Siluriformes: Pimelodidae), in Colombia using microsatellite markers. Actualidades Biológicas. Vol 31, Pag 112.
- Roman – Valencia, C. and **Perdomo – Sabogal, A**. 2002. Ecología trófica y reproductiva de A*rgopleura magadalenensis* en la cuenca del río La Vieja, Alto Cauca, Colombia. Memorias XXXVII Congreso nacional de Ciencias Biológicas.

Poster presentations

- Society for Molecular Biology and Evolution Conference, San Juan - Puerto Rico, 2014. *Evolution of Transcription Factor Genes: a Potential Source for Human Specific Innovations*.
- 33rd Conference of the International Society for Animal Genetics. 2012. *Evolution and diversity of MHC class and class 11 genes in suids and peccaries*. Brisbane: University of Queensland Press.
- XXXVII National congress of biological sciences. 2002. Feeding and reproductive ecology of *Argopleura magdalenensis* in the Cauca and Magdalena Basins, Colombia. 2002.
- First regional symposium. University of Quindío. 2000. Feeding ecology of the Neotroplical fish *Argopleura magdalenensis* in the Cauca basin, Colombia.

Eigenständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 10. Mai 2016

Alvaro Perdomo-Sabogal