Evolutionary history and diversity of human-specific *FAM72A* paralogs: insights from population genetics

by

Ilya Kisselev

A thesis submitted to the Faculty of Graduate Studies in partial fulfillment of the requirements for the Master of Science in Bioscience, Technology and Public Policy

Department of Biology

Master of Bioscience, Technology and Public Policy

The University of Winnipeg

Winnipeg, Manitoba, Canada

Abstract

Gene duplication is a key driver of genetic diversity and adaptation, allowing genomes to develop complexity and redundant sequences that evolve along different trajectories. After gene duplication, selective pressure relaxes, leading to various evolutionary outcomes: neofunctionalization, subfunctionalization, or pseudogenization. In human evolution, gene duplication played an important role: since divergence from the common ancestor with chimpanzees, humans have gained approximately 75 lineage-specific genes, influencing brain development, dietary adaptation, and immune regulation. The *FAM72* gene family, with four paralogs (*FAM72A-D)* that arose after human-chimpanzee divergence, illustrates this process.

The evolutionary history and function of the *FAM72* paralogs remain poorly described. The ancestral FAM72A protein drives early stages of somatic hypermutation in B cells by antagonizing UNG2. However, *FAM72C-D* paralogs have Trp125Arg amino acid substitution that prevents them from interacting with UNG2. This study hypothesizes that after the initial duplication from *FAM72A* to *FAM72B*, *FAM72B* duplicated to *FAM72C* and *FAM72D*. I hypothesize that opposing selective forces operate on *FAM72A-B* and *FAM72C-D* paralogs. Another hypothesis is that population-specific exposure to local environments during human evolution has driven the selection of population-specific adaptive haplotypes of *FAM72A* paralogs.

The study used the 1000 Genomes dataset, testing selection through neutrality metrics and haplotype-based scores, and investigated functional divergence by comparing conserved amino acid sites and gene-wide LD patterns across human populations. Bayesian divergence time estimation between *FAM72* paralogs was performed using the

most common haplotypes in humans and chimpanzees. The hypothesized sequence of duplication events was supported by the phylogenetic analysis. The neutrality metrics identified *FAM72C* as recovering from a selective sweep, with other paralogs not showing signals of positive selection. Integrated haplotype scores of *FAM72D* suggested a recent selective sweep in African populations, and *FAM72A-B* showed high conservation. Linkage disequilibrium analysis highlighted functional regions, with *FAM72A* and *FAM72B* sharing active LD-enriched promoters, while *FAM72C* contained an active enhancer linked to immune cell function. Finally, multiple signatures of balancing selection were observed in an intronic region of *FAM72C*.

The results suggest neutral or relaxed selection for *FAM72A-B*, but purifying selection following a selective sweep for *FAM72C-D*. The divergence of paralog pairs is evident in regulatory and functional shifts, notably with *FAM72C*'s unique immune cell associations. No clear signs of population-specific adaptation were identified, but *FAM72B* shows distinct haplotypes between East Asian and South Asian populations, hinting at either population bottlenecks or adaptive evolution. The findings show how gene duplication within the *FAM72* gene family has contributed to genetic diversity and potential adaptability, with some members potentially shaping the evolutionary trajectory of immune function in human populations.

Acknowledgments

This thesis could not have been completed without the support and encouragement from many wonderful people in my life.

Foremost, I would like to thank Dr. Sara Good. Your patience and guidance have been invaluable throughout this journey. In our research discussions, your openness to new ideas always stood out. You encouraged me to think independently, explore different angles, and bring my own insights to the table. This freedom to experiment and think creatively has been incredibly empowering and has greatly improved my research skills.. Your support, especially when things got tough, made a big difference. You are the best mentor I could have hoped for, and I am incredibly grateful for your help over these years.

To my committee members, Dr. Xiaoqing Liu and Dr. Alberto Civetta, thank you for your time and the insightful feedback that improved this thesis a lot.

A big shoutout to my friends – you know who you are. Thanks for being there for me, for the laughs, the late-night chats, and for just being a call away when things got tough.

Lastly, to my family, I owe you more than words can say. Thank you for your endless support and for believing in me, even when what I do might seem a bit out there. Your faith in me has been a driving force behind everything I have achieved.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Gene duplication and evolution

The diversity of life on earth and the complexity of biological systems are largely attributed to the molecular mechanisms that govern genetic variation, including nucleotide substitution and structural genomic changes. Gene duplication provides raw material for the evolution of new genetic functions and contributes to the innovation of phenotypes (Ohno, 1970).

Structurally complex regions of the human genome serve as the birthplace for many unique human genetic variants. These areas are prone to structural changes, which thereby enhance the emergence of new structural variants and gene copy number changes. These changes are crucial for evolution, providing the raw materials from which genes unique to humans have emerged, thereby contributing to significant phenotypic changes in a short period of evolutionary time and potentially making an important contribution to human-specific traits.

In this introduction, I aim to describe the mechanisms of gene duplications, how duplicates become fixed within populations and species, their evolutionary fates, and the role of segmental duplications in the evolution of humans and primates. I will also synthesize the current research on the function of human-specific *FAM72A* paralogs and their potential adaptive benefits in human evolution.

1.1.1 Mechanisms of gene duplication

The evolutionary paths and outcomes of duplicated genes are diverse and shaped by the conditions under which the duplicates emerge (Davis and Petrov 2005; Guan, Dunham, and Troyanskaya 2007; Makino and McLysaght 2010). Gene duplicates can

arise from several different genomic processes: these processes leave idiosyncratic molecular signatures informing us as to the mechanisms that generated the duplicates. The mechanisms driving gene duplication also occur at different frequencies and have different selective forces shaping their evolutionary fate. There are several strategies to determine the age of a new gene, and depending on the gene's classification, it may be either straightforward or challenging to discern which copy is the ancestral gene and which is the new or derived gene.

Gene duplications occur on two different scales: whole-genome duplication (WGD) and small scale duplications (SSD) that involve duplication of individual genes or chromosomal segments. WGD, often the result of errors in meiotic or mitotic cell division, is a source of gene duplicates that typically leads to significant functional and morphological innovations throughout evolution. The fixation rate of WGD events is low, which is reflected in the small number of extant polyploid species (Van de Peer, Maere, and Meyer 2009; Arrigo and Barker 2012; Van de Peer, Mizrachi, and Marchal 2017). Despite their rarity, the lineages that have managed to resolve the complex genomic issues such as genome instability, higher mutation load and doubled energy and resources costs following a WGD, have often become extremely successful, as evidenced in both vertebrates (Dehal and Boore 2005) and flowering plants (Tang et al. 2008).

Duplicated genes that are generated either via WGD or SSD are known as paralogs: the subset of paralogs that are specifically derived from a WGD are referred to as ohnologs. The generation of ohnologs represents a significant opportunity for genetic novelty since the redundancy of having entire sets of genes allows for the entire network to evolve new functions or become specialized while retaining the function of the

ancestral set of genes or network. The opportunity for the generation of specialized or novel pathways following WGD occurs because duplication of the entire genome preserves the stoichiometry of protein-protein interactions and, therefore, prevents dosage imbalance. Further, duplication via WGD involves duplication of regulatory elements that sometimes exist at considerable distances from the protein-coding region of a gene: this further facilitates the creation of complex and redundant gene regulatory networks that contribute to increased genomic complexity and robustness. However, computer simulations have shown that post-WGD, these redundant gene regulatory networks can decrease the fitness of organisms in stable environmental conditions due to the accumulation of random mutations. On the other hand, in fluctuating environments or near-extinction events, these duplicated networks have a greater likelihood of being retained because they can improve the evolvability and adaptive capacity of the genomes, allowing species to survive and adapt to changing conditions (Yao, Carretero-Paulet, and Van de Peer 2019). Over time, when the environment stabilizes, polyploid genomes undergo a process known as diploidization, whereby there is massive gene loss and genomic reorganization with only a few adapted ohnologs being retained, and chromosomes segregate in dyads (Wolfe 2001). Some studies have argued that the loss of different sets of ohnologs between populations may lead to rapid divergence between them, leading to radial speciation (Paterson, Bowers, and Chapman 2004; Hoegg et al. 2004). However, the long-term role of WGD in increasing diversification rates of lineages, particularly in flowering plants, remains an area of debate.

While whole-genome duplication can lead to evolutionary novelties, the likelihood of the successful establishment of polyploid lineages is low, making its

contribution to adaptive evolution less frequent. Paralogs are more commonly produced through SSD of chromosomal segments known as segmental duplications, which are influenced by the genomic environment of the duplicated region. These duplications often occur due to errors in DNA repair mechanisms or during DNA replication (Bailey et al. 2001).

Double-strand DNA breaks can be repaired either through non-allelic homologous recombination (NAHR) or non-homologous end joining (NHEJ). NAHR operates by aligning homologous sequences with high similarity, which can mispair and crossover during cell division, leading to deletions, duplications, or inversions depending on how the sequences align. The involvement of NAHR in recent gene duplications is usually indicated by identical sequences flanking the duplicated area. Conversely, in the absence of such sequences, NHEJ or DNA replication errors are typically considered responsible.

Notably, ohnologs are rarely duplicated further by segmental duplication in humans (Maere et al. 2005; Makino and McLysaght 2010). This observation may be attributed to more stringent dosage balance constraints, supported by a higher probability of haploinsufficiency in ohnologs (Vance and McLysaght 2023).

New gene copies may also arise via retrotransposition, which is an RNA-mediated duplication mechanism distinct from DNA-mediated processes. In mammals, the primary facilitators of this mechanism are the long interspersed nuclear elements (LINEs), which are capable of reverse transcription due to their encoded reverse transcriptase enzyme (Q. Feng et al. 1996). Most retrocopies generated through retroposition initially lack promoters, rendering them inactive; however, they may acquire promoters secondarily *de novo*. Conversely, some retrogenes may already possess promoters if they were

transcribed from regions with distributed alternative start sites, resulting in transcriptionally active retrogenes upon insertion (Okamura and Nakai 2008). Additionally, retrotransposed genes can integrate into exons of other genes, creating novel fusion genes with altered protein domain architectures (Akiva et al. 2006). These retrogenes are characterized by a faster resolution of evolutionary trajectory, whether they are rapidly purged due to deleterious effects or fixed in the genome due to adaptive fitness effects, compared to retropseudogenes and products of segmental duplications (Carelli et al. 2016).

1.1.2 Evolutionary fates of duplicated genes

Following a segmental duplication that results in a paralog formation, the new gene enters a polymorphic phase. During this period, it is characterized by instability, as it could either become fixed within the population or be completely lost. Under an assumption of neutrality, the probability of the segmental duplication achieving fixation by genetic drift in a diploid species with a census population size N_C is $(2N_C)^{-1}$ (Kimura and Ohta 1969). However, this probability shifts when considering duplications that have fitness effects. In large populations, duplicates that have beneficial effects are more likely to become fixed, while deleterious duplicates are swiftly purged from large populations (Crow and Kimura 1970). Fundamentally, while deleterious gene duplicates are typically eliminated shortly after their formation, the probability that a neutral duplicate becomes fixed is equal to its frequency in the population, and the probability that a beneficial gene duplicate becomes fixed depends on the selective advantage of the gene and the population size (Otto and Yong 2002).

While gene duplication is predominantly associated with deleterious phenotypic consequences (Conrad and Antonarakis 2007; Makino and McLysaght 2010; Mahjani et al. 2022), duplicated genes can provide direct adaptive benefits (Rapoport 1940; Ohno 1970). One beneficial scenario is gene dosage amplification, particularly advantageous when the ancestral gene does not produce enough of its protein to meet the organism's needs under specific environmental conditions. Under such circumstances, the expression of a duplicated gene can effectively double the gene dosage, potentially providing an adaptive advantage that could lead to the rapid fixation of the duplicate within the population. This process is exemplified by the widespread duplication of ribosomal and transfer RNA genes across various species (Ohno 1970) and the duplication of amylase genes (Perry et al. 2007). Notably, this mechanism would lead to a distinct pattern of duplicated gene maintenance: the newly duplicated gene would provide an immediate adaptive advantage, prompting a strong selective sweep that rapidly fixes the duplicate in the population. Following fixation, the fixed locus would be almost identical to the ancestral gene (Hahn 2009).

Segregation avoidance is another mechanism that can direct fixation of a duplicated gene (Spofford 1969; Ohno 1970). According to this model, heterozygous individuals show higher fitness than homozygous individuals, which carry identical alleles. However, due to the principles of Mendelian inheritance, these heterozygous individuals would produce offspring, half of which would be homozygous and have lower fitness. To avoid segregation load, a single-copy gene may be duplicated in the genome. This duplication allows one copy to become fixed with one allele, and the other to become fixed with the alternate allele, thereby maintaining persistent heterozygosity

within the population. The mechanism ensures that all offspring will inherit the fitness advantages of heterozygosity. Segregation avoidance was observed in the mosquito species *Culex pipiens*, where duplication of the gene coding for acetylcholinesterase led to the presence of individuals carrying both an insecticide-resistant allele, which had reduced activity, and a sensitive allele with normal enzymatic activity. Heterozygous individuals were not affected by pesticides but also did not suffer from fitness costs associated with the reduced enzyme activity (Lenormand et al. 1998).

After achieving fixation, the long-term fate of a duplicated gene is typically categorized into one of three evolutionary trajectories, as illustrated in Figure 1. One of the duplicate genes often experiences relaxed selection, meaning it no longer faces the strong evolutionary pressure to retain its function. This happens because the other duplicate can maintain the necessary function, freeing its counterpart from the evolutionary constraint. Over time, the gene under relaxed selection can accumulate mutations. Some of these mutations, if they disrupt the gene's function, can lead to pseudogenization, converting the gene into a non-functional pseudogene. However, during this critical period, the duplicated gene also has a limited window to acquire gain-of-function mutations. Such mutations could result in novel, selectable traits, providing new opportunities for different selective forces to act upon the gene, and preventing it from pseudogenization (Walsh 1995). In addition to the high probability of pseudogenization following duplication, some genes that are successfully neo-functionalized can, at a later time point, undergo pseudogenization due to changes in the environment, as illustrated by the rapid loss of functional olfactory receptors in the Homininae clade (Hughes, Teeling, and Higgins 2014).

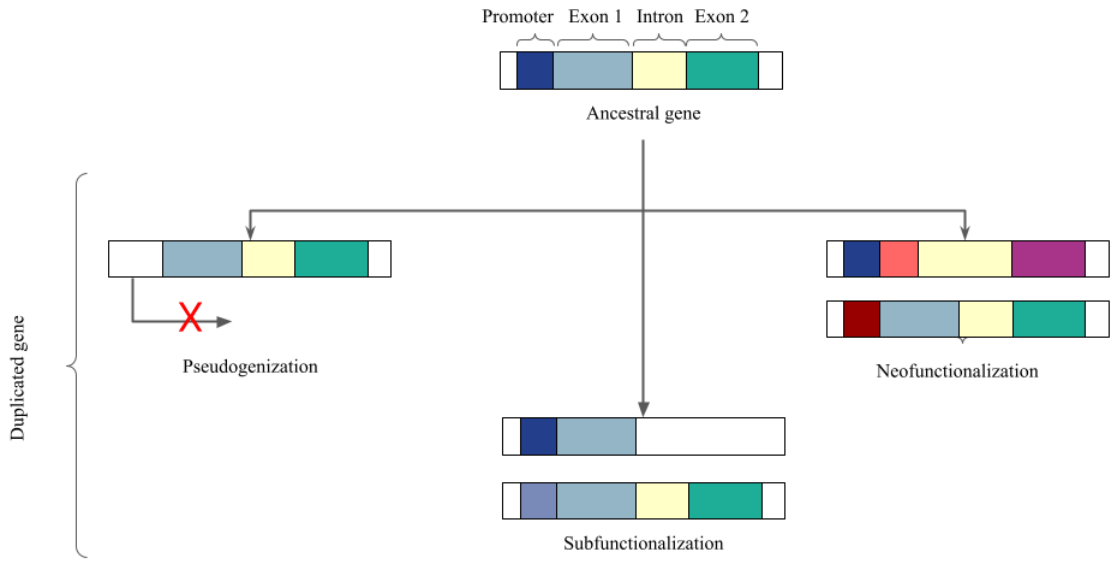

Figure 1. Evolutionary fates of duplicated genes. In pseudogenization, the duplicated gene accrues detrimental mutations, such as start-loss mutations, leading to a loss of protein expression. Subfunctionalization is marked by the complementary partitioning of the ancestral gene's functions and/or expression profiles between the derived genes, effectively maintaining the ancestral state across the duplicates. Coding neofunctionalization involves the acquisition of novel protein functions through the accumulation of coding mutations, resulting in functions not present in the ancestral gene. Regulatory neofunctionalization, on the other hand, is defined by the emergence of new expression patterns not present in the ancestral gene, contributing to the diversification of gene function post-duplication.

Prior to the fixation of loss-of-function mutations in one of the duplicate genes, the redundancy provided by the duplicated gene creates an opportunity for the exploration of mutations that were previously unfavourable. These mutations might enable the gene to overcome local fitness peaks in search of a global fitness maximum, or they might lead to the development of novel protein functions. This process, where duplicated genes evolve new functions due to relaxed purifying selection followed by positive selection of adaptive mutations, is termed neofunctionalization by Ohno (1970).

Two types of neofunctionalization are distinguished: regulatory neofunctionalization (Figure 1, lower gene diagram), which results from mutations in regulatory elements leading to new spatial and/or temporal patterns of gene expression

not observed in the ancestral gene; and coding neofunctionalization (Figure 1, upper gene diagram), which results from gain-of-function mutations in the gene's open reading frame (Moore and Purugganan 2005). Both types have been observed recurrently in the evolution of opsin genes across multiple vertebrate lineages (Yokoyama 2008; Cortesi et al. 2015).

The probability of a random mutation being beneficial is very low, making Ohno's adaptation model of neofunctionalization as the sole mechanism for the retention of functional duplicates implausible. According to the Dykhuizen–Hartl model, initially proposed by Kimura (1983), mutations in a duplicated gene do not become fixed immediately after their emergence. Rather, in this model, neutral variants segregating in a population may become subject to positive selection in the presence of new environmental challenges. On the other hand, the adaptation model suggests that neofunctionalization occurs through the ongoing adaptive fixation of mutations at one of the duplicated loci.

Näsvall et al. (2012) proposed the "innovation–amplification–divergence" model of neofunctionalization. In this model, a gene with a minor secondary function, which is not initially essential for survival, can gain selective advantage if environmental conditions change. Duplication of this gene may result in a higher dosage of an adaptive protein with immediate fitness effects. Subsequently, the duplicated genes can specialize, each developing distinct functions. This model of paralog evolution was observed in *Salmonella enterica*, whereby the ancestral gene was originally involved in histidine biosynthesis and had a nonessential tryptophan biosynthesis activity. Three thousand

generations following the gene duplication event, the paralogs specialized, one performing histidine and the other tryptophan biosynthesis.

Reliance of neofunctionalization models on beneficial adaptive consequences of gene duplications and relaxed selective pressures acting on them makes these models unlikely in many ecological contexts. An alternative model of duplicate retention by subfunctionalization was proposed and has become more widely accepted (Force et al. 1999; Lynch and Conery 2000). In this model, the duplicated genes go through a two-phase process known as duplication-degeneration-complementation (DDC). In the early stages after gene duplication, the fate of the gene copies can diverge in several ways: they might become nonfunctional and turn into pseudogenes, evolve new roles via neofunctionalization, or accumulate mutations. These mutations might limit the range of activities that the gene copies can perform or reduce their expression under relaxed selection. However, the mutations might allow each duplicate to complement the other, effectively restoring the functional capacity of the ancestral gene. This process of complementation serves to retain both genes in the population, since neither can perform the full functions of the parent gene on its own. It will also limit the genes' potential for undergoing neofunctionalization or turn into pseudogenes. If such changes occurred, it would render the second duplicate incapable of performing the ancestral gene's functions, thereby solidifying the need for both duplicates to coexist and cooperate.

While the DDC model provides a mechanism for the long-term maintenance of duplicated genes in a genome, it falls short in accounting for the rapid fixation of a duplicated gene, attributing this phenomenon mainly to stochastic genetic drift. In contrast, the "Escape from Adaptive Conflict" (EAC) model assumes that adaptive

mutations drive the fixation and retention of duplicated genes. This model presupposes that the ancestral gene, prior to duplication, carried out two distinct functions. However, selection for the performance of one function inadvertently led to a compromise in the other. The duplication event thus provides an opportunity to distribute the diverging functions of the ancestral gene across the two resulting duplicates. This functional split allows each duplicate to undergo function-specific adaptive mutations, leading to their rapid fixation and preservation. Over time, this process results in a rapid functional and structural divergence between the two duplicates, enabling them to collectively outperform the ancestral gene in carrying out its original functions (Sikosek, Chan, and Bornberg-Bauer 2012; Barkman and Zhang 2009).

The various models proposed to explain the retention of duplicated genes in a genome do not operate in isolation; rather, they are interconnected and can influence each other in numerous ways. It is plausible to suggest that gene families have undergone multiple cycles of subfunctionalization, serving to distribute various functions across different genes that originated from a single ancestral gene. This process of spreading out functions can reduce the constraints on these new genes, creating favorable conditions for subsequent neofunctionalization. As a result, the reduced constraints on these new genes can enhance neofunctionalization (Rastogi and Liberles 2005; He and Zhang 2005; Jouffrey, Leonard, and Ahnert 2021). Similarly, ongoing cycles of neofunctionalization can lead to a single gene accumulating a variety of functions over time. When such a gene undergoes duplication, followed by subfunctionalization, these accumulated functions can be distributed across multiple genes. This distribution not only preserves

the functions but also provides additional opportunities for the evolution of novel functions.

1.1.3 Segmental duplications drive adaptation in great apes and humans

Comparative analyses of high-quality genome assemblies from various mammals, including primates, have revealed that a significant portion of their genomes is made up of segmental duplications (SDs). Segmental duplications are highly identical blocks of genomic DNA either scattered across and between chromosomes or serially (tandemly) organized in close proximity to each other (Bailey et al. 2001), These regions are enriched for repetitive DNA and can contain genes with preserved intron-exon structure. Notably, primates, and particularly humans, show a higher proportion of these SDs compared to other mammals (She et al. 2008; Liu et al. 2009; Nicholas et al. 2009; Vollger et al. 2022; Shao et al. 2023). In contrast to the 70-90% of SDs in mice (She et al. 2008) and rat (Gibbs et al. 2004) that occur as tandem duplications, the majority of SDs in humans are dispersed throughout the genome, indicating a shift from tandem to interspersed duplications in the primate lineage, which is most evident in humans and great apes.

A study of 50 primate genomes (Shao et al. 2023) confirmed a significant increase in segmental duplications in the ancestor of the great apes. This research also highlighted numerous lineage-specific duplications in primates that are thought to contribute to their unique evolutionary adaptations. Sudmant et al. (2013) reported that in the lineage of African great apes, segmental duplications have occurred at a rate 2.8 times higher than deletions. Moreover, the contribution of copy number changes to the genetic

differentiation between species was found to be 1.4 times greater than that of single nucleotide substitutions in the period from 8 to 12 million years ago.

The pattern of these duplications is not random, with SDs in the hominid lineage forming large, fractally organized clusters (Jiang et al. 2007). Using a repeat-graph approach, researchers identified 24 core duplicons — segments each about 15 kilobases long — that are overrepresented within 437 duplication blocks in the human genome. These core segments serve as starting points for further accumulations of SDs, leading to larger blocks of duplicated DNA over 250 kilobases in size. These larger blocks are made up of smaller, newer segments that are positioned increasingly further from the core duplicon. Intriguingly, these central duplicons have been independently expanded in different primate lineages (Cantsilieris et al. 2020).

In the human genome, SDs that are unique to our lineage are known as human-specific duplicates (HSDs). Dennis et al. (2017) identified 30 gene families with more than 80 genes that are present in > 90% of humans. Expression divergence analysis between these human-specific paralogs and corresponding single-copy orthologs failed to identify the unifying factor driving this process (Shew et al. 2021). However, the authors showed that expression of HSDs is asymmetrically conserved, with the derived paralogs having reduced and more divergent expression across tissues compared to the ancestral gene.

Many HSDs are associated with traits that are characteristic of human neural development, and are actively expressed in the brain (Dougherty et al. 2018). For instance, the gene *SRGAP2C* is involved in human-specific synaptogenesis(Schmidt et al. 2019); *ARHGAP11B* plays a role in the amplification of neocortical basal progenitor cells

(Heide et al. 2020; Fischer et al. 2022); and *NOTCH2NLB* is implicated in cortical neurogenesis (Fiddes et al. 2018). Moreover, HSDs were also implicated in more recent non-neuronal adaptations. For example, the adaptation to a starch-rich diet following the Neolithic revolution is associated with the segmental duplication and copy number variation in the *AMY1* gene, which codes for the salivary amylase enzyme (Perry et al. 2007; 2015). Shew et al. 2021 proposed that paralogs of the *NCF1* gene in humans may offer protection against autoimmune diseases. This hypothesis is supported by evidence that mice lacking the equivalent gene exhibit a higher incidence of arthritis and autoimmune encephalomyelitis (Hultqvist et al. 2004). Furthermore, having additional copies of the *NCF1* gene correlates with a lower risk of systemic lupus erythematosus in humans (Jiang et al. 2007; Zhang et al. 2022). Although the paralogs *NCF1B* and *NCF1C* are generally nonfunctional pseudogenes, gene conversion events can occasionally revert them to their functional ancestral state, resulting in an increased gene dosage in some individuals (Heyworth, Noack, and Cross 2002).

1.2 *FAM72A* gene family

Evolutionary and population genetic studies provided evidence of varying strength for the contribution of human-specific genes to the traits that distinguish humans from other ape species (Dennis et al 2017, Hsieh et al. 2021). Feng et al. (2021) showed the critical role of ancestral FAM72A in adaptive immune response. This gene, unique to humans due to three segmental duplication events, has been identified as a key factor in somatic hypermutation and class switch recombination in B cells, primarily through its interaction with UNG2. However, the human-specific paralogs *FAM72C* and *FAM72D* have a critical Trp125Arg amino acid substitution, which inhibits their ability to bind to

UNG2. Despite this divergence from the ancestral function, these genes exhibit a high degree of conservation, suggesting alternative functional roles or evolutionary pressures maintaining their presence in the genome. To further explore this phenomenon, my research aims to identify the evolutionary drivers behind the conservation of these genes and generate hypotheses regarding the novel functions of the derived *FAM72A* paralogs.

1.2.1 *FAM72A* evolution

The evolutionary history of the Family with sequence similarity 72 (FAM72) is not well annotated or understood. FAM72A is an archaic group of proteins hypothesized to have originated during the emergence of Opisthokonta, but subsequently lost within the Protostomia lineage. The domain architecture diverges significantly between fungal and metazoan FAM72 proteins. Specifically, the latter demonstrates a notable degree of protein conservation characterized by the presence of a single functional domain. In contrast, the former taxonomic cluster is characterized by the presence of several functional domain groups along with the FAM72 domain. These structural differences potentially mirror the functional divergence of these proteins over the course of their evolution. Following two successive rounds of WGDs that occurred prior to the emergence of jawed vertebrates, the duplicates of the *FAM72A* gene rapidly underwent subsequent losses. The presence of multiple gene copies of *FAM72A* within the Cyprinidae and Salmonidae families can be traced back to lineage-specific WGD events. In mammalian species such as humans, blue whales, and degus, several paralogs of the *FAM72A* gene can be identified. In degus and blue whales, the duplication of *FAM72A* took place through retrotransposition, resulting in a duplicated gene characterized by its

CDS being encoded within a single exon. In contrast, the duplicated *FAM72A* genes in humans have preserved the structural configuration of the ancestral gene.

Within the human genome, the *FAM72* gene family comprises four paralogous genes located on chromosome 1. These genes originated from three lineage-specific segmental duplication events. The paralogs are colocalized with the *SRGAP2* paralogs and underwent sequential duplication together. By analyzing shared genomic regions among *SRGAP2* paralogs, Dennis et al. (2012) proposed that approximately 3.4 million years ago (Mya), a duplication in the *FAM72A-SRGAP2* region led to the emergence of *FAM72D-SRGAP2B*. Subsequently, around 2.4 Mya, another duplication event gave rise to *FAM72B-SRGAP2C*, originating from the pre-existing *FAM72D-SRGAP2B* locus. Finally, approximately 1 million years ago, the *FAM72C-SRGAP2D* locus emerged from the same genomic region (Figure 2). This proposed model establishes a connection between the formation of the *FAM72D-SRGAP2C* and *FAM72C-SRGAP2D* loci with the evolutionary emergence of the *Homo* genus.

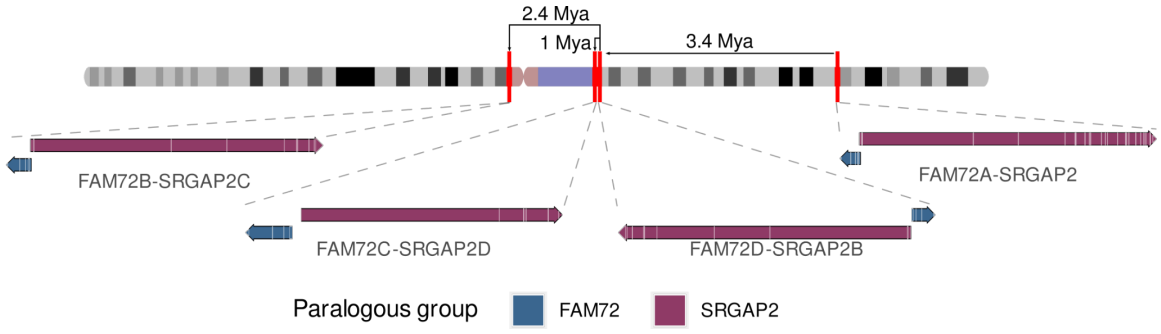

Figure 2. Chromosomal localization of segmentally duplicated *FAM72-SRGAP2* loci. Each gene is represented by an arrow, with the direction indicating its orientation; genes on the upper part are on the positive strand, while those on the lower part are on the negative strand. The white segments within the arrows denote the exonic regions. Additionally, black arrows mark the proposed sequence of locus duplication events based on Dennis et al. (2012). The pink-shaded region on the chromosome corresponds to the centromeric region, while the blue area represents the pericentromeric region.

1.2.2 FAM72A gene structure

FAM72A paralogs are located on chromosome 1, with the ancestral gene residing on 1q32.1 distantly from the derived genes. Three derived genes are positioned in euchromatin around the pericentromeric and centromeric regions, with *FAM72C* and *FAM72D* located on the q arm, and *FAM72B* on the p arm (Figure 2). Importantly, the *FAM72A-C* genes are oriented on the negative strand, while *FAM72D* is located on the positive strand. Each *FAM72A* paralog is found in close proximity to one of the four human-specific *SRGAP2* paralogs on the opposite strand. The first exons of the *FAM72A-SRGAP2* and *FAM72B-SRGAP2C* gene pairs share overlapping regions spanning 1068 and 629 bp, respectively. Conversely, the first exons of the *FAM72C-SRGAP2D* and *FAM2D-SRGAP2B* gene pairs are separated by 3010 and 446 bp (Figure 2) (Kutzner et al. 2015).

In contrast to divergent co-duplicated *SRGAP2* paralogs, human *FAM72A* paralogs display a remarkably high degree of similarity in both gene sequence and structure. The derived *FAM72A* paralogs originated from complete gene duplication that retained the entirety of the ancestral ORF, whereas two duplicated *SRGAP2* genes encode truncated ancestral proteins, and the third derived gene is pseudogenized (Dougherty et al. 2018; Sporny et al. 2017). The Ensembl 110 (Cunningham et al. 2022) canonical gene models of *FAM72A* paralogs consist of four exons (Figure 3, Figure 4A), the length of the second and third is constant across all paralogs, while the lengths of the first and the fourth exons vary due to less conserved 5'- and 3'-UTRs (Figure S1). The three introns of *FAM72A*, *FAM72B*, and *FAM72C* have the same intron lengths (~ 2kb, 3.9kb and 8.4 kb), while *FAM72D* possesses an approximately 200 bp longer first intron, and an

approximately 1 kb longer second and third introns compared to the other paralogs' (Kutzner et al. 2015).

Segmentally duplicated *FAM72* genes are characterized by high sequence similarity. The average pairwise sequence identity across the whole gene among human *FAM72* paralogs is 99.15%. The lowest sequence identity is observed between *FAM72A* and *FAM72D* at 98.91%, while the highest sequence identity of 99.39% is found between *FAM72C* and *FAM72D*. When comparing these derived paralogs to their ancestral gene, a non-random pattern of sequence divergence becomes apparent, as illustrated in Figure 3. Specifically, the 5'-UTR region, as well as several sections in the second and third introns, emerge as the primary drivers of sequence dissimilarity. Comparative analysis with Neanderthal and Denisovan sequences indicated that the second and third introns of *FAM72A* demonstrate the most variability (Kutzner et al. 2015). Although the 5'-UTR region showed the greatest dissimilarity among the paralogs, their divergence is primarily attributed to variation in the lengths of the beginning of the region that differs by up to 60 nt between paralogs (Figure S1), whereas the aligned portion of the region differs only at five nucleotide positions (Kutzner et al. 2015). On the contrary, the 3'-UTR region is highly identical among paralogs with no observed truncations. In the 3'-UTR, the paralogs differ by the presence/absence of seven indels, each ranging in size from 1-3 nt, and by seven substitutions in potential AU-rich areas known to affect translation (Kutzner et al. 2015; Otsuka et al. 2019).

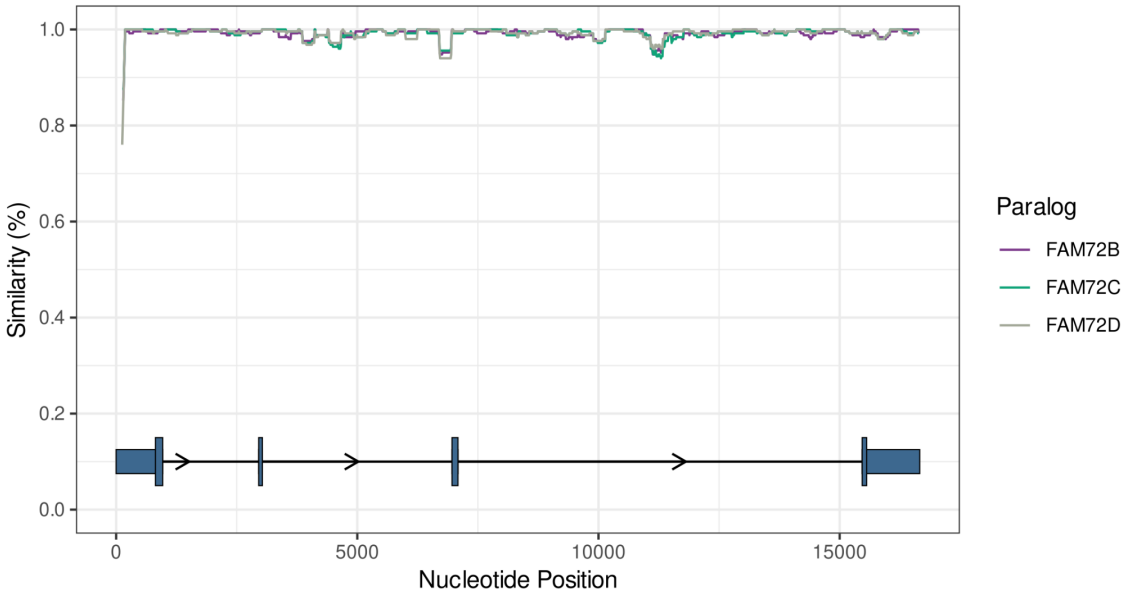

Figure 3. Comparative sequence similarity of human *FAM72* paralogs. This figure represents the sequence similarity between the derived *FAM72B-D* genes and the ancestral *FAM72* gene, assessed using overlapping windows of 250 nucleotides (nt) with a step size of 10 nt. At the bottom of the plot, the structure of the canonical *FAM72A* transcript is plotted.

The *FAM72* genes in humans produce multiple transcripts, each distinct in exon content and amino acid composition (Figure 4A). Canonical transcripts from the Ensembl database for the four paralogs encode a consistent 149 amino acid peptide, a pattern also mirrored in other mammalian species. Intriguingly, this highly conserved isoform has only five variable amino acid positions among the four human paralogs (Figure 4B). In addition to this canonical transcript, *FAM72A-B* paralogs generate a shorter isoform of 109 amino acids. This isoform begins with the same start codon as the canonical version, but its first exon is truncated due to an alternative upstream donor site. Similarly, *FAM72B-C* paralogs produce a 94 amino acid isoform by skipping the third exon, a variation also identified in multiple other mammalian species. Besides well-supported isoforms, several novel *FAM72A-B* isoforms with additional upstream exons and alternative start codons were computationally predicted (Cunningham et al. 2022).

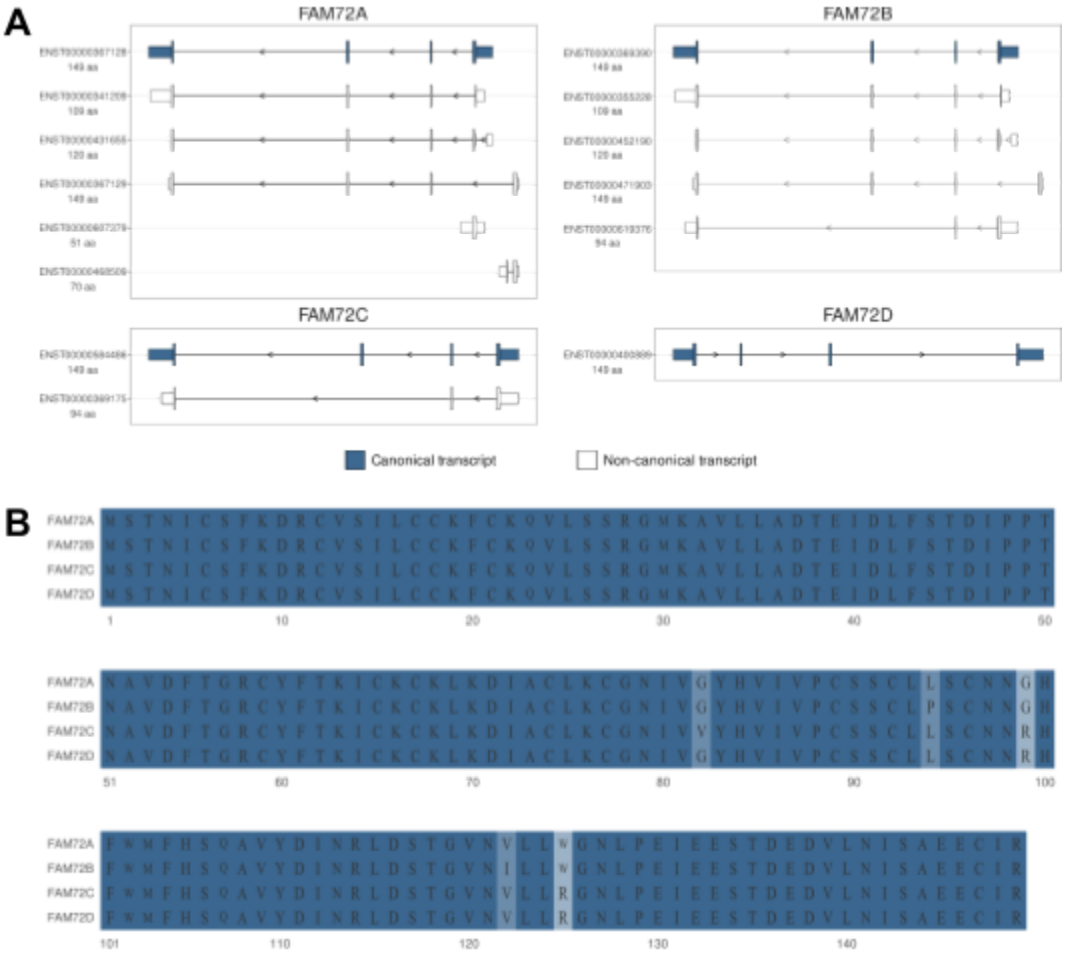

Figure 4. Isoform and protein diversity of human *FAM72* paralogs. A. Ensembl gene models of human *FAM72* paralogs. The canonical gene models are shown in blue color; gene models, for which CDS is not defined, are not provided. **B.** Multiple sequence alignment of proteins encoded by canonical gene models. Amino acid positions are color-coded based on their conservation level, with lighter shades indicating lesser conservation.

1.2.3 *FAM72A* expression

As described earlier, divergence in gene expression is one of the mechanisms by which paralogous genes are maintained in the genome. This process, involving tissue- and developmental-stage-specific gene expression, fosters purifying selection which protects genes from accruing loss-of-function mutations and subsequently becoming pseudogenes. Considering the *FAM72A* paralogs, which show low divergence in protein

sequence but pronounced variations in the 5'-UTR region (Figure 3), expression divergence could be a major factor in the retention of these derived genes.

Supported by harmonized expression data from the Human Protein Atlas (Human Protein Atlas 2023; Uhlén et al. 2015) widespread expression of the *FAM72A* gene was observed, while its derived counterparts demonstrate tissue-specific variation in expression magnitude (Figure 5). Notably, *FAM72C* is not expressed in the central nervous system, in contrast to the other genes. Additionally, when comparing expression levels, *FAM72A* stands out as the major contributor to the total expression in tissues, with the other genes contributing less than half of all *FAM72* transcripts. However, in immune tissues, this pattern shifts: there is an overall higher expression of paralog genes, and *FAM72A's* contribution drops to around a quarter of the combined expression of paralogs.

Examining the expression profile of *SRGAP2* paralogs, co-duplicated counterparts to *FAM72* with the ancestral gene still dominating the expression profiles, reveals another layer of contrast. Firstly, these genes exhibit a nearly six-fold increase in expression levels compared to *FAM72*. Secondly, there is an observed tissue specificity in the magnitude of expression: *FAM72* genes predominantly express in lymphoid organs, while *SRGAP2* genes mainly express in central nervous tissues. Ho, Kutzner, and Heese (2019) corroborate this, demonstrating that activation of the Fam72-Srgap2 intergenic region promoter in response to contrary growth factors (Ngf and Egf) results in the selective overexpression of Srgap2 or Fam72a, respectively.

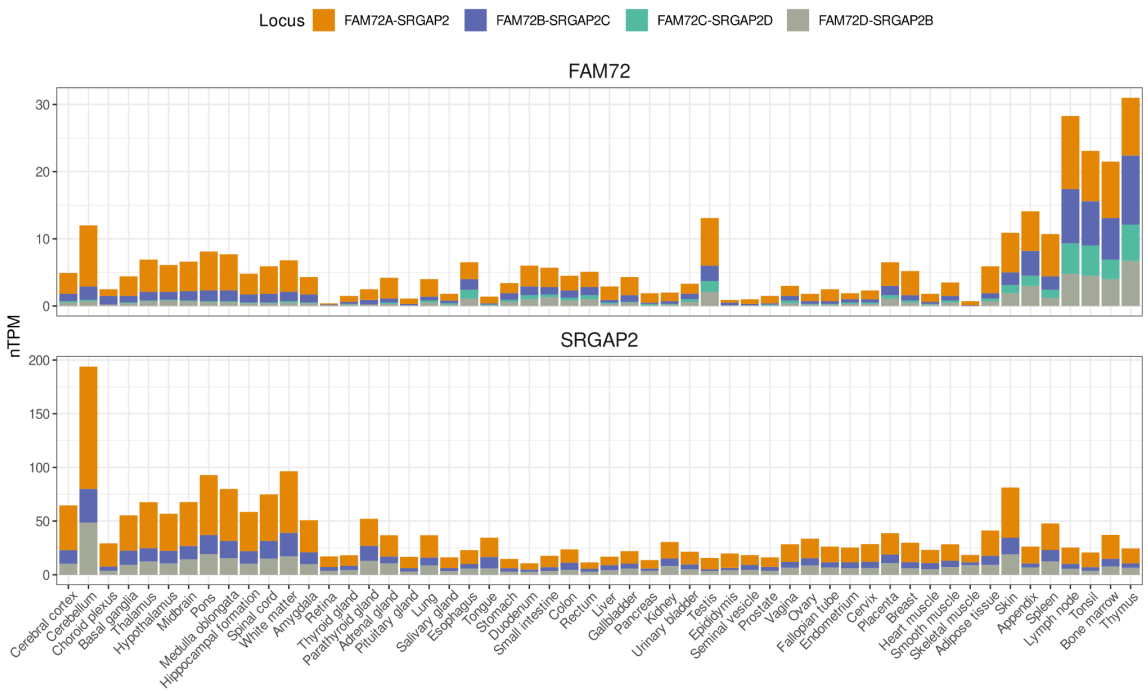

Figure 5. Expression profile of *FAM72A* and *SRGAP2* paralogs across tissues. The data is represented as normalized transcripts per kilobase million from the Human Protein Atlas (HPA) and Genotype-Tissue Expression (GTEx) project. *FAM72D* TPM values were filtered out.

As discussed in Section 1.3.2, *FAM72A-B* genes are characterized by the presence of several alternative start codons and splice sites leading to the existence of distinct isoforms. Unfortunately, short-read RNA-sequencing (RNA-seq) technologies usually fail to accurately capture and quantify the transcript isoforms of genes that are highly identical and segmentally duplicated (Hardwick et al. 2019). In an effort to overcome these limitations, Glinos et al. (2022) used Oxford Nanopore Technologies to generate a log-read RNA-seq dataset from GTEx tissues and cell lines, where they detected over 70,000 novel transcripts that were not previously annotated. Being compatible with the short-read combined HPA-GTEx data, the overall expression of *FAM72* genes remained low across all tissues. *FAM72A* stood out as the primary contributor to gene expression in these tissues (Figure 6).

At the transcript level, only two transcripts of *FAM72A* were detected. Intriguingly, neither of them matched the canonical transcript listed in Ensembl. The transcript with the highest expression was a previously unannotated and novel one. The second transcript, while also detected, was characterized by low support evidence in Ensembl. On the other hand, the derived *FAM72A* paralogs were found to have the highest expression of the canonical transcripts. Additionally, a low-support *FAM72B* transcript was identified, which was similar to the one detected in *FAM72A*. The absence of ancestral primate *fam72a* canonical transcript produced by *FAM72A* in humans may indicate its neofunctionalization after gene duplication.

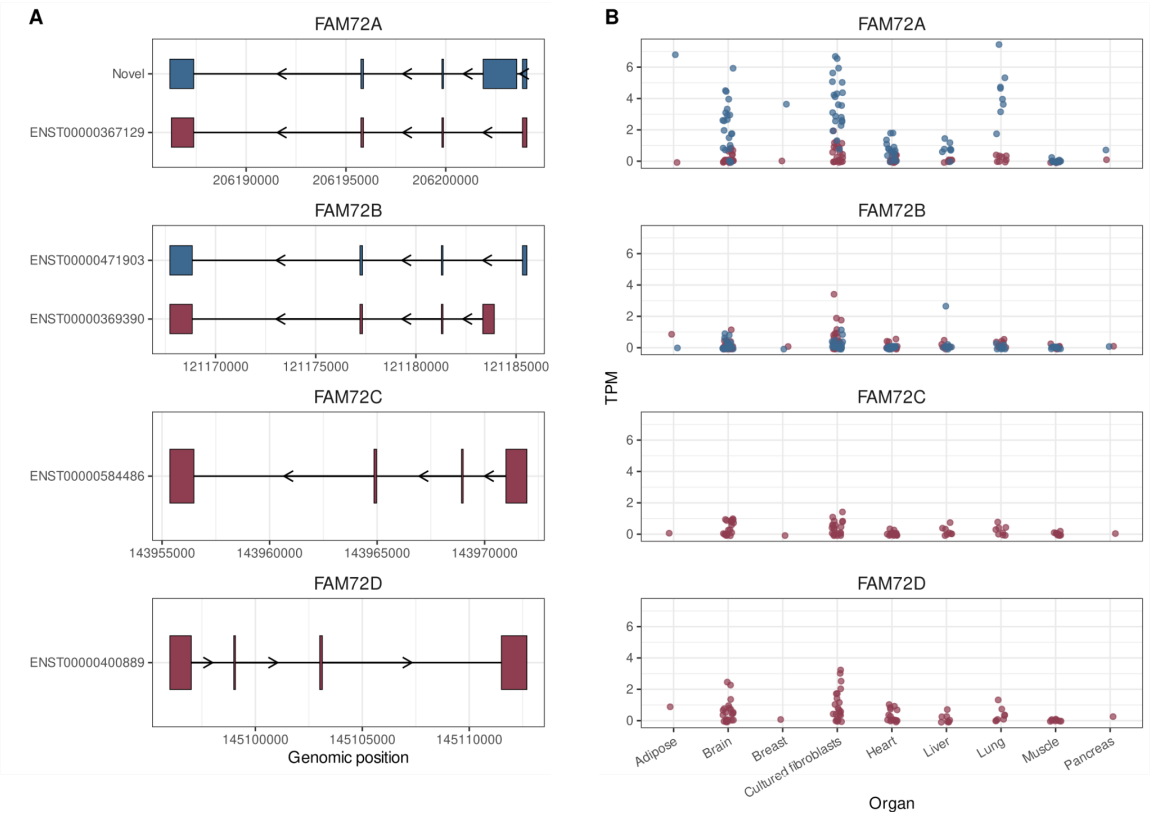

Figure 6. Transcript-level expression profile of *FAM72A-D* genes across organs based on the GTEx long-read RNA-seq dataset. A. *FAM72A-D* transcripts that were identified in the dataset. **B.** Expression profiles of *FAM72A-D* genes, TPM values color-coded to match the associated transcript from Panel A. Data from distinct regions within the same organ have been aggregated.

1.2.4 FAM72 and immunity

Somatic hypermutation (SHM) plays a vital role in diversifying antibodies, a key aspect of the adaptive immune response. This process refines and enhances antibody function after the initial V(D)J recombination event in activated B cells. SHM primarily occurs in the germinal centers located in secondary lymphoid organs. During SHM, the enzyme activation-induced cytidine deaminase (AID) introduces mutations at a high rate specifically in the variable regions of immunoglobulin genes by converting C:G base pairs into U:G mismatches (Neuberger et al. 2003).

Typically, the appearance of U:G mismatches would activate the base excision repair (BER) machinery. The first step of this repair process involves Uracil DNA-Glycosylase 2 (UNG2), which removes the uracil base, leaving behind an abasic site. The BER machinery is then supposed to resolve this site to maintain genomic integrity. However, the scenario is different in germinal B cells. These cells express the FAM72A protein, which targets and degrades the UNG2 protein (Y. Feng et al. 2021; Rogier et al. 2021). As a result, the U:G mismatches introduced by AID persist through the cell cycle until the S-phase. At this stage, the error-prone mismatch repair (MMR) system excises mismatched nucleotides along with surrounding nucleotides, producing single-stranded DNA, which is then repaired by low-fidelity, error-prone polymerases. This process, while seemingly counterintuitive, is essential for creating the diverse range of antibodies required for a robust adaptive immune response (Chahwan et al. 2012).

Despite existing evidence that demonstrates FAM72A's role in UNG2 degradation, the exact mechanism behind this process remains unclear. Previous research using murine models showed that FAM72A can moderately inhibit UNG2 upon binding

(Y. Feng et al. 2021). However, this interaction appears to be species-specific, as the human paralogs of these proteins do not exhibit the same behavior (Guo et al. 2008). Guo et al. (2008) utilized co-immunoprecipitation of FAM72A with various deletion constructs of UNG2, followed by Western blot analysis, to identify the protein regions critical for their interaction. Their findings pinpointed the necessity of the first 25 amino acids at the N-terminus of the UNG2 protein for effective protein binding. Furthermore, they identified a crucial tryptophan residue at position 125 of the FAM72A protein, which is required to facilitate the UNG2-FAM72A binding. Interestingly, the *FAM72C-D* genes encode an arginine instead of tryptophan at the equivalent position, rendering their protein products incapable of binding to UNG2. Molecular modeling studies further contributed to our understanding of this interaction. (Renganathan et al. 2021). demonstrated that, in addition to the W125 residue, the F61, F104, and T131 residues are essential for maintaining FAM72A stability and ensuring effective binding to UNG2. However, experimental data, particularly from studies conducted by Stewart and Bhagwat (2022) suggest that only mutations in F104 and W125 have a more pronounced impact, significantly disrupting the expression and stability of murine FAM72A.

UNG is also crucial for the innate immune response, working downstream of the APOBEC (Apolipoprotein B mRNA Editing Catalytic Polypeptide-Like) enzymes. APOBEC enzymes deaminate cytosine bases in DNA, transforming them into uracil, particularly when pathogen DNA is detected within the cytoplasm. However, herpesviruses have evolved strategies to counteract the APOBEC3 enzyme, helping them to escape the inhibitory effects of UNG. By doing so, these viruses can take advantage of UNG's DNA repair functions while safeguarding their genomic stability from its

potentially damaging effects. Additionally, herpesviruses encode their version of UNG, the function of which is unknown (Savva 2020). The human *FAM72A* paralogs, which are unable to degrade human UNG2, might have acquired new functions or specialized roles to address the challenge posed by viral Ung proteins. Indeed, indirect evidence of *FAM72A* paralogs involvement in innate immune response exists. For example, in a study on the embryonic mouse neocortex, when human *FAM72D* was ectopically expressed, three out of the six genes that showed increased expression levels were linked to the immune response (Andrä 2020). In a separate study, (Y. Li et al. 2019) employed a Bayesian gene co-expression network analysis to identify central hub and driver genes implicated in the development of systemic lupus erythematosus, an autoimmune disease. Within this network, *FAM72C* emerged as one of the top 12 hub genes. Notably, 10 out of these 12 genes demonstrated a strong association with interferon (IFN) levels. A hub gene is defined as a gene that exhibits a statistically significant higher number of expression correlations with other genes than what is expected on average.

1.2.5 FAM72 and neurogenesis

Given the role of *SRGAP2* in cortical migration and development, emergence of human-like characteristics in pyramidal neurons (Charrier et al. 2012; Fossati et al. 2016), and high expression of *FAM72* in proliferating neural stem cells in hippocampus (Benayoun et al. 2014), Ho et al. proposed that co-duplicated *SRGAP2* and *FAM72* paralogs operate together to enhance higher cognitive functions of the brain (Ho, Kutzner, and Heese 2019; Ho et al. 2021). Specifically, their model suggests that while *SRGAP2* triggers the differentiation of neural stem cells, *FAM72* works to preserve a pool of undifferentiated cells.

Indirect evidence in support of *FAM72* as a regulator of neural progenitor cell (NPC) proliferation comes from Benayoun et al. (2014), who isolated mouse NPCs and assessed regulatory landscape using ChIP-Seq targeting H3K4me3. They found *Fam72a* among the genes with the top 5% broadest domains. Further experimental evidence showed that *Fam72a* knock-off in these cells decreased their proliferation and led to increased neuronal differentiation. Additionally, cancer studies associated *FAM72A* overexpression with increased proliferation of neoplastic cells (Guo et al. 2008; Rahane, Kutzner, and Heese 2019; Ho et al. 2021).

1.3 Rational for the study

Segmental duplications have been pivotal in the evolutionary divergence and adaptation of African great apes including humans. In humans, the *FAM72A-SRGAP2* locus has undergone sequential duplications over two million years, resulting in the retention of seven non-pseudogenized paralogs from two gene families, which suggests a process of either neofunctionalization, where new genes acquire novel functions, or subfunctionalization, where gene copies divide the work of the ancestral gene. Indeed, *SRGAP2* and its paralog *SRGAP2C* are recognized as key players in the development of human-specific patterns of neurogenesis. However, the remaining *SRGAP2* paralogs do not seem to be under the same evolutionary pressure to maintain function.

The hypothesis that *SRGAP2* alone or the combined action of *FAM72-SRGAP2* as master locus are responsible for conserving all four loci does not fully align with the observed discrepant conservation patterns of these paralogs. Here, I address recent evolutionary history and population genetics of human-specific *FAM72A* paralogs

attempting to untangle evolutionary mechanisms responsible for the maintenance of these genes.

1.4 Objectives

The primary objectives of my thesis are to determine whether different selection forces act on *FAM72A* paralogs and whether duplication events resulted in functional divergence of the genes.

My secondary objective is to refine the timeline of duplication events leading to the emergence of *FAM72A* paralogs.

1.5 Hypotheses

H1: Ancestral *FAM72A* gene was duplicated to *FAM72B*, which gave rise to *FAM72C* and *FAM72D*.

H2: Opposing selective forces operate on *FAM72A* paralogs: *FAM72A* and *FAM72B* show signatures of neutral evolution or balancing selection, *FAM72C* and *FAM72D* demonstrate signatures of positive selection.

H3: Given the involvement of FAM72A protein in immune response, population-specific locally beneficial haplotypes exist.

Chapter 2: Methods

2.1 Data acquisition and preprocessing

2.1.1 Data sources and code availability

Links to publicly available datasets used in this work are provided in Table S1. Code that was used for data processing and the analysis is available on GitHub at https://github.com/iykisselev/fam72_evo.

2.1.2 1000 Genomes data preprocessing

PCR-free high-coverage whole-genome sequencing data provided by the 1000 Genomes Project (Byrska-Bishop et al. 2022) was used as a main dataset in the present work. The dataset consists of 3,202 individuals, including 602 trios, from 26 populations across the globe (Table S2). The key steps of data processing performed by data providers are summarized in this section.

The Illumina NovaSeq 6000 sequenced reads were aligned to the 1000 Genomes version of GRCh38 that includes additional decoy sequences and alternate versions of the HLA locus using a pipeline from the Centers for Common Disease Genomics (Regier et al. 2018). Variant discovery was performed using the HaplotypeCaller tool in GVCF mode, incorporating sex-specific ploidy configurations for chromosomes X and Y. Before phasing, only SNVs that passed variant quality score recalibration (VQSR), had genotype missingness rate <5%, Hardy-Weinberg equilibrium (HWE) exact test p-value > 10^{-10} in at least one superpopulation, mendelian error rate (MER) < 5%, and minor allele count (MAC) > 2 were retained. Phasing was conducted chromosome-wise using the SHAPEIT-duohmm software (O'Connell et al. 2014).

2.1.3 Chimpanzee FAM72A haplotypes

Whole-genome sequencing FASTQ files of three chimpanzee subspecies (*P. t. troglodytes* (n = 14), *P. t. schweinfurthii* (n = 12), *P. t. verus* (n = 5)), sequenced by de Manuel et al. (2016) were downloaded from the European Nucleotide Archive (ENA) at EMBL-EBI with accession number PRJEB15086. Low-quality base tetramers with an average base quality below 30 were removed using trimmomatic v0.39 (Bolger, Lohse, and Usadel 2014). Subsequently, the trimmed FASTQ files were divided into 20 chunks with Seqkit v2.3.0 (Shen et al. 2016), facilitating parallel mapping to the chimpanzee NHGRI_mPanTro3-v1.1 assembly using the BWA-MEM algorithm of bwa v0.7.17 (H. Li and Durbin 2009) with the specified options "-Y -K 100000000".

After mapping, the individual SAM files were merged, and using samtools fixmate (v1.17, Danecek et al. 2021) mate coordinates, insert size, and related flags were refined. PCR duplicates were then removed with samtools rmdup. The deduplicated bam files were filtered to include reads that mapped to chromosome 1, had a mapping quality of 30 or higher, and were part of a properly aligned pair. In contrast, the reads that were unmapped, had an unmapped mate, were duplicates, or were supplementary alignments were excluded. This specific filtering was executed using the samtools view command with the arguments "-q 30 -f 3 -F 1796".

Freebayes v1.3.6 (Garrison and Marth 2012), was used to jointly call genotypes for individuals from the three chimpanzee subspecies separately. The resulting VCF files were filtered to retain only those variants with specific criteria: a quality score exceeding 19, an average quality per alternate read above 10, support from both forward and reverse strands for the alternate allele, and more than one read at each tail distance, ensuring the

minimization of positional bias for the reference allele. This was defined by the filter: "QUAL > 19 && QUAL / INFO/AO > 10 && SAF > 0 && SAR > 0 && RPR > 1 && RPL > 1". Using bcftools norm and the argument "-m -any", multiallelic sites in these VCF files were converted to biallelic, followed by the removal of any variants with absent genotypes. A custom R script was then used to shift the positions of multiallelic sites, preventing position overlap.

Phasing of haplotypes from the *FAM72* gene, along with its ±5 kb flanking region, was carried out using SHAPEIT v2.r904 (O'Connell et al. 2014). A 0.5 Mb window was used, and the recombination rate parameter (rho) and effective population size (Ne) were set to 0.00119 and 32492 for *Pan troglodytes schweinfurthii*, and 0.00119 and 44000 for *Pan troglodytes troglodytes*, as inferred by Fontsere et al. (2022). *Pan trogodytes verus* (n = 5) samples were excluded from the analysis due to insufficient sample size.

Post-phasing, the SHAPEIT output was converted back to VCF files. The genomic positions were shifted to their original locations, and positions for multiallelic sites were merged using "bcftools norm -m +any". Finally, FASTA consensus sequences for the chimpanzee FAM72A haplotypes were generated using bcftools consensus.

2.2 Phylogenetic analysis

To infer phylogenetic relationships and the order of duplication events of *FAM72* genes in the human lineage, a Bayesian phylogenetic-based analysis was performed using BEAST (Bouckaert et al. 2019). First, the *FAM72* paralogs' haplotypes, with frequencies > 100 across human populations, were aligned alongside the predominant *FAM72A* haplotypes from two chimpanzee subspecies and a gorilla *FAM72A* gene sequence sourced from the NCBI reference genome (NHGRI_mGorGor1-v1.1-0.2). This multiple

sequence alignment was performed using the L-INS-i iterative refinement method implemented in MAFFT v7.490 (Katoh and Standley 2013).

IQ-TREE v2.1.2 (Minh et al. 2020) was used for substitution model selection (Kalyaanamoorthy et al. 2017). The BEAST substitution settings, based on the IQ-TREE's results, included the Gamma site model (Yang 1994) with Category Count set to 1, a proportion of invariant sites of 0.7582, and the HKY substitution model (Hasegawa, Kishino, and Yano 1985) with transition/transversion ratio of 5.1714. The Calibrated Yule phylogenetic branching model was chosen, setting the birthRate parameter's distribution to Gamma (0.001, 1000) for both the Strict and Optimized Relaxed Clock models. Specifically for the Strict Clock model, the clockRate was set to Gamma (0.001, 1000). The average of absolute mutation rate estimates for humans, chimpanzees, and gorillas, taken from Besenbacher et al. (2019), was set at 0.00054 substitutions per site per million years (Myr). For calibrations based on fossil records, the divergence distribution for the human-chimpanzee split was set as uniform [4.631, 15 Myr] (Vries and Beck 2023), the Hominini-Gorillini split was set to normal (13.0, 1.8 Myr) based on estimates by Langergraber et al. (2012).

Both the Strict Clock and Optimized Relaxed Clock models were applied to the dataset to determine which provided a better fit to the data. The performances of the different clock models were evaluated using nested sampling (Russel et al. 2019). This calculation involved using 4 particles, a ChainLength of 10^5, and a SubChainLength of 2×10^4. The preferred clock model was identified by contrasting marginal log-likelihoods to derive the Bayes factor and used to generate estimates for the Time to the Most Recent Common Ancestor (TMRCA) for internal nodes on the tree.

Subsequent analysis was conducted using the top-performing clock model across three independent runs, each with a ChainLength of 10^8 and a 10% burn-in. The most suitable run for further analysis was selected based on the criteria of highest effective sample size (ESS) and likelihood, as assessed through Tracer v1.7.2 (Rambaut et al. 2018). The maximum clade credibility tree was then extracted using TreeAnnotator v2.7.5, discarding a 10% burn-in. This tree was visualized using ggtree v3.8.0 (S. Xu et al. 2022).

2.3 Diversity of FAM72A paralogs

2.3.1 Genetic differentiation

The fixation index, Fst, measures the proportion of genetic diversity between populations relative to the overall genetic variance. Both selective pressures and demographic events influence the extent to which populations genetically differ from each other. A higher Fst value in certain genomic regions may indicate directional selection in one population if an allele becomes favoured and increases in frequency in one population but not in another, leading to marked differentiation. On the other hand, lower Fst values indicate a lack of time to diverge by drift, purifying selection that removes new variants or balancing selection, in which multiple alleles are maintained at intermediate frequencies across populations. While Fst values are influenced by both selection and demographic factors, the effects of selection are specific to individual loci, whereas demographic history and population structure exert their influence across the entire genome.

Pairwise genetic differentiation between 26 human populations from the 1000 Genomes Project (Table S2) at *FAM72* loci was calculated using the method of Weir and

Cockerham (1984). This analysis was performed using the hierfstat R package v0.5-11 (Goudet 2005), using multiallelic indel-free VCF files that were imported in R and converted to genind objects. Then, nonmetric multidimensional scaling (vegan v2.6 (Dixon 2003)) was used to inspect patterns of genetic differentiation patterns among the populations.

To determine which model — either the Out-of-Africa (OoA) or biogeographical ancestries/Superpopulations (population groups based on ancestral continental origins) — better explained the observed variation in genetic differentiation, the Permutational Multivariate Analysis of Variance (PERMANOVA) introduced by Anderson (2001) was used. This analysis was conducted using the vegan v2.6 R package (Dixon 2003), with 10^6 permutations for each model. In cases where both models yielded a p-value < 0.05 (indicating the proportion of permutations with a pseudo-F value equal to or exceeding the actual data), the model with the higher partial R^2 was selected as the best fit.

2.3.2 Haplotype networks

While genetic differentiation may indicate presence of selective forces, their targets within the gene cannot be inferred using Fst values alone. Analyzing haplotype frequencies across various populations can shed light on their genetic diversity and evolutionary trajectories. Isolation between populations often results in the divergence of allele frequencies between populations due to random genetic drift, the accumulation of population-specific genetic variants and can change the frequency of alleles or haplotypes between populations if different selective pressures are operating in different populations; each of these evolutionary forces can lead to distinct signatures in the distribution and nature of haplotypes between populations.

FAM72 genetic variants were extracted and saved in separate VCF files from the phased indel-free 1000 Genomes chromosome 1 VCF file. Then, each file was read into R using the vcfr v1.14.0 package (Knaus and Grünwald 2017) and transformed into a DNAbin object. The frequency of each distinct haplotype was calculated, the haplotype networks were inferred and plotted with pegas v1.2 R package (Paradis 2010) using haplotypes with a minimal frequency of 15.

2.3.3 Genetic variation profile of *FAM72A* paralogs

To assess the conservation of human-specific *FAM72A* paralogs at the nucleotide and protein level, the chr1 VCF file was sourced from the gnomAD v3.1.2 dataset (Chen et al. 2022). Variants were filtered to exclude those with a zero allele count, low-confidence genotypes (GQ < 20; DP < 10; and AB < 0.2 for het calls), those that did not meet the VQSR filtering thresholds of -2.7739 for SNPs and -1.0606 for indels, and those with an inbreeding coefficient of less than -0.3. The filtered dataset was then subset to include only variants of *FAM72A* paralogs. These selected genetic variants were then processed in R to classify and quantify their mutation types, utilizing the VariantAnnotation v1.46 package (Obenchain et al. 2014).

3.3 Natural selection of *FAM72A* paralogs

3.3.1 Phylogenetic protein conservation

Highly conserved regions across different species often indicate critical functional or structural roles of the protein. High conservation suggests that these regions are essential for the protein's function and have been maintained through evolutionary pressures, whereas variation in conserved regions might indicate relaxed selection or adaptive evolution in response to different environmental pressures.

Multiple sequence alignment (MSA) of vertebrate FAM72A protein sequences from EggNOG v5.0 (Huerta-Cepas et al. 2019) was performed using L-INS-i iterative refinement method implemented in MAFFT v7.490 (Katoh and Standley 2013). Before alignment, protein sequences were reviewed, and non-canonical discrepant isoforms and potential assembly errors were discarded. Using this alignment, two conservation scores for each position were calculated. The first score, based on Shannon entropy, offers a straightforward conservation measure that correlates with consensus amino acid frequencies, but does not account for amino acid substitution probabilities. It was calculated using BALCONY v0.2.10 R package (Płuciennik et al. 2018). Conversely, the score by Bodenhofer et al. (2015) in msa v3.17 R package that takes these substitutions into account, yielding a more moderate conservation estimate, was used with BLOSUM62 substitution matrix.

3.3.2 Neutrality tests

To test the departures of *FAM72A* paralogs from neutral evolution, I computed several metrics: Tajima's D, Fay and Wu's D, Zeng's H, nucleotide diversity, and haplotype diversity. These tests differ by their ability to detect selective sweeps at different time scales. While Tajima's D and Zeng's H detect loci that restore depleted variation after a selective sweep, Fay and Wu's D detects signatures of ongoing selective sweep. Given that these metrics can be influenced by both selective forces and demographic processes, the method outlined by Atkinson et al. (2018) was adopted. Specifically, to account for unique population histories characterizing specific biogeographical ancestry groups, I derived background distributions of the neutrality statistics individually for each superpopulation. Superpopulation-specific neutrality

metrics were considered extreme, if they were within the top or bottom 5% of the empirical distribution for that metric.

The neutrality statistics were calculated in adjacent non-overlapping segments equivalent to the average length of *FAM72A* paralogs (17,387 bp) on chromosome 1. To speed up and parallelize computation, the multiallelic, indel-free, filtered chromosome 1 VCF file was segmented into approximately 250 subsets using bcftools (Danecek et al. 2021). Using the PopGenome v2.7.5 R package (Pfeifer et al. 2014), these segments were imported into R and the neutrality statistic values in non-overlapping sliding windows for each superpopulation were calculated. Regions containing fewer than 5 polymorphic sites were omitted to prevent inaccurate results caused by low variation.

Using the same dataset, a sliding window approach — with a window size of 5 kb and a shift of 500 bp — was applied to determine Tajima's D values throughout the *FAM72-SRGAP2* loci, including its adjacent ±100 kb regions. This strategy aimed to produce a detailed Tajima's D map, highlighting potential regions of interest within the gene bodies. In addition, neutrality statistics for *FAM72* and *SRGAP2* paralogs, along with the intergenic regions, were computed separately for each human population.

3.3.3 Integrated Haplotype Score (iHS)

While classical neutrality statistics can detect departures from neutral evolution within a locus of interest, they cannot identify core SNPs that drive selective sweep of the haplotype. In contrast, integrated haplotype scores identify regions of extended haplotype diversity at the center of which selected SNPs are located.

The variation VCF file for *Homo sapiens* from Ensembl 110 was filtered to retain only SNPs located on chromosome 1. Biallelic sites in this file were merged, after which

the chromosomal position, and ancestral allele information for every variant were extracted. In this file, the ancestral state was inferred from the EPO multiple alignments using Ortheus by the Ensembl team (Paten et al. 2008).

In the phased VCF file for chromosome 1 from the 1000 Genomes project, parents of trios were omitted, indels were removed and biallelic variants were merged into multiallelic format. This VCF file was then annotated with ancestral allelic state data from the filtered Ensembl file. Any variants lacking ancestral data were discarded. The multi-sample VCF file was then split into 26 population-specific VCF files, which were then processed concurrently.

Each of these VCF files was loaded into R and converted into a haplohh object using the rehh v3.2.2 R package (Gautier, Klassmann, and Vitalis 2017). During this process, alleles were polarized using ancestral allele information provided in the INFO/AA tag. The integrated extended haplotype homozygosity (EHH) was computed for haplotypes observed a minimum of two times. Parameters set during this phase included an EHH termination threshold at 0.01, a maximum permissible gap between two markers of 20 kb, and the deactivation of EHH curve interpolation. The standardized ratio of iHH values for two alleles was determined using variants that had a minor allele frequency (MAF) of at least 0.01. The standardization frequency bin was also fixed at 0.01. Adjustments to the p-values for multiple comparisons were made using the Benjamini and Hochberg (1995) method.

3.3.4 Balancing selection statistic $\beta^{(1)}$

The $\beta^{(1)}$ score, introduced by (Siewert and Voight 2017), serves as a quantitative tool to identify genomic positions under the effect of long-term balancing selection. This

metric is based on an expectation that alleles near a site undergoing LTBS should exhibit a correlated frequency pattern. To test this hypothesis, the $\beta^{(1)}$ score contrasts two distinct measures of population mutation rate: Watterson's theta (θ_W) and θ_β, the latter of which is calculated as an average of variant counts weighted for similarity in allele frequency to the core variant of interest.

To detect signatures of balancing selection, 26 population-specific VCF files containing multiallelic SNPs with identified ancestral allelic states were used. Ancestral alleles were extracted from the INFO/AA flag and transformed into homozygous diploid genotypes and added as an extra "dummy" individual to the VCF files. Subsequently, each VCF file was transitioned into an ACF file using glactools (Renaud 2018) and the vcfm2acf command with the argument "--onlyGT". Then, the dummy individual with homozygous ancestral genotypes was set as the ancestor and used as the root populations using the usepopsrootanc command followed by conversion into betascan format with acf2betascan command using "--useanc" parameter to obtain unfolded site frequency spectrum.

To obtain standardized $\beta^{(1)}$, a chromosomal map of Watterson's θ is required. The population-specific Watterson's θ were calculated for 0.5 Mb non-overlapping windows on chromosome 1, adopting an approach analogous to what is described in the Neutrality tests section. Population-specific unfolded site frequency spectrum files and Watterson theta maps were then used to run BetaScan (Siewert and Voight 2017) on 2kb windows. A $\beta^{(1)}$ statistic that fell within the topmost 2% was considered extreme.

3.3.5 Linkage disequilibrium

To explore the potential evolutionary and functional implications of haplotype stratification, I analyzed the linkage disequilibrium (LD) correlation coefficients of the *FAM72-SRGAP2* loci. The phased chr1 VCF file was subset to include only biallelic variants at *FAM72-SRGAP2* loci using bcftools (Danecek et al. 2021). This VCF file was subsequently transformed into a DNAbin object, and in the process, indels were excluded. Pairwise LD correlation coefficients for all potential SNP pairs within each locus and population were then calculated using the pegas v1.2 package (Paradis 2010).

To identify regions that are enriched in LD, I employed a two-step approach. First, SNPs falling within the uppermost 2% of LDs in every population were identified. From this subset, the first genomic position of each SNP pair was extracted. We then counted the number of populations in which a specific SNP was associated with the top 2% LD. This strategy identifies both: genomic positions exhibiting robust LD across various populations, and positions displaying strong LD with several regions concurrently. Based on the frequency of occurrence gene-wise, the top 10% of these positions were binned into 100 bp genomic intervals. These intervals were annotated with overlapping regulatory elements using the Ensembl Regulatory Build (release 110) (Zerbino et al. 2015).

Chapter 3: Results

3.1 Phylogenetics and emergence of human-specific *FAM72A* paralogs

The duplication model of the *FAM72-SRGAP2* locus presented by Dennis et. al (2012), which primarily relied on *SRGAP2* sequence divergence, convincingly described the structural evolution of the locus. The analysis was based on *SRGAP2* gene sequences from humans, chimpanzees, and orangutans that were used for the construction of a neighbor-joining gene tree. Tajima's relative rate test was used to equalize the substitution rates across all branches. The corrected phylogenetic tree was used to estimate the timing of the duplication events. This involved using the standardized substitution rate and the chimpanzee-human divergence time (6 million years) to calculate the approximate timeframes for the initial and subsequent duplication events. However, the inferences from their study were not tested for concordance with variation across *FAM72A* paralogs. To properly resolve the chronology of *FAM72* gene duplication events, I used seven common human haplotypes from *FAM72A*, six from *FAM72B*, three from *FAM72C*, and five from *FAM72D*. Additionally, three haplotypes from two chimpanzee subspecies were included to determine the time to the most recent common ancestor (TMRCA) using Bayesian molecular sequence analysis.

Using IQ-TREE's ModelFinder (Kalyaanamoorthy et al. 2017), the HKY+F+I model was identified as the most suitable for the dataset, based on the Bayesian Information Criterion (BIC) with a score of 61607.012 (Table S3). This model combines the Hasegawa-Kishino-Yano (HKY) model for nucleotide substitution, variable base frequencies (F), and a consideration for invariant sites (I). The rate parameters indicated higher transitions rates between A-G and C-T (5.1714) compared to other nucleotide

substitutions. The rate heterogeneity model revealed a significant proportion of invariant sites (0.7582), indicating notable conservation within the dataset.

When contrasting two clock models, the optimized relaxed model markedly outperformed the strict clock model (log(BF) = 13587.7), and it was used for phylogenetic dating. The three runs obtained using this model did not show significant disparities, and the most suitable run with the highest log likelihood and ESS of 1637.67 was selected. The summary of the estimated parameters of the model, including the estimated TMRCA of internal nodes, is provided in Table S4.

The resulting phylogenetic tree from this model demonstrates the distinct haplotypes of the *FAM72A* paralogs converging into well-supported monophyletic clades (Figure 7). The inferred timeline for the duplication events of *FAM72A-B* was approximately 3.16 million years ago (Mya) (95% High Posterior Density (HPD): 1.25-5.36 million years (Myr)), for *FAM72B-CD* was around 2.38 Mya (95% HPD: 0.89-4.12 Myr), and the *FAM72C-D* divergence was estimated to be 1.57 Mya (95% HPD: 0.51-2.83 Myr). Another important observation is the relatively recent divergence timelines of the *FAM72* paralogs' haplotypes. Specifically, the divergence between *FAM72A* and *FAM72B* haplotypes is estimated at 0.48 Mya (95% HPD: 0.10-0.83 Myr) and 0.42 Mya (95% HPD: 0.10-0.83 Myr), respectively, while those for *FAM72C* and *FAM72D*, are estimated at 0.17 Mya (95% HPD: 0.01-0.38 Myr) and 0.2 Mya (95% HPD: 0.03-0.42 Myr), respectively.

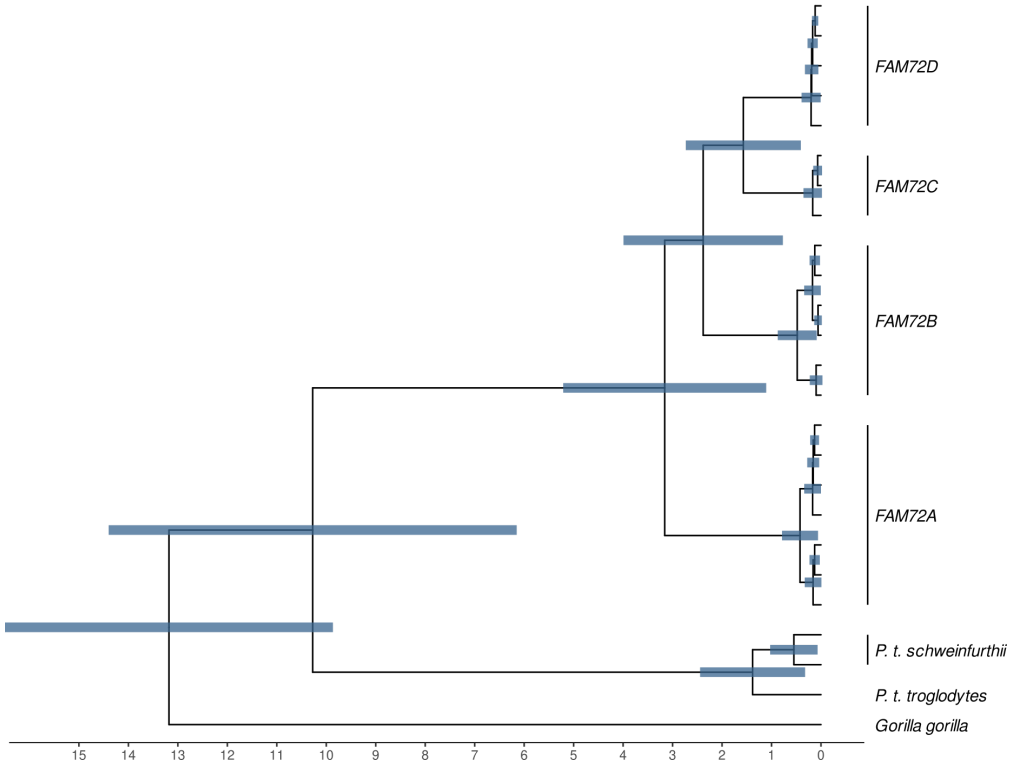

Figure 7. Maximum clade credibility phylogenetic tree illustrating the evolutionary relationships among human *FAM72A-D* haplotypes. The x-axis indicates divergence time in million years before the present, nodes represent the most recent common ancestors of each branch, and blue bars indicate the 95% high posterior probability densities of TMRCA point estimates.

3.2 Diversity of *FAM72A* paralogs

3.2.1 Genetic differentiation

For this study, genetic differentiation in *FAM72* loci was assessed between 26 populations from the 1000 Genomes dataset using unbiased Weir and Cockerham's estimator of Fst (Weir and Cockerham 1984). The findings indicate a generally low level of genetic differentiation across the populations and loci examined, with an average pairwise Fst of 0.061 for *FAM72A*, 0.076 for *FAM72B*, 0.045 for *FAM72C*, and 0.052 for *FAM72D*. A deeper look into superpopulation data, as illustrated in Figure 8A, shows that African populations consistently clustered across all genes. Notably, other

superpopulations form distinct clusters only in the *FAM72A* and *FAM72B* genes. For *FAM72C* and *FAM72D* genes, non-African populations formed an Out-of-Africa (OoA) cluster (Figure 8A).

Varying patterns of population differentiation between *FAM72A-B* and *FAM72C-D* genes may indicate either demographic shifts or distinct evolutionary processes. To determine the statistical significance of the observed clusters, I used permutational analysis of variance (Anderson 2001) that compared two simplified models of population grouping. The first model organized the 26 populations based on their biogeographical ancestries into five superpopulations based on prehistoric ancestral continental origins. In contrast, the second model grouped all non-African populations, contrasting them from the African superpopulation.

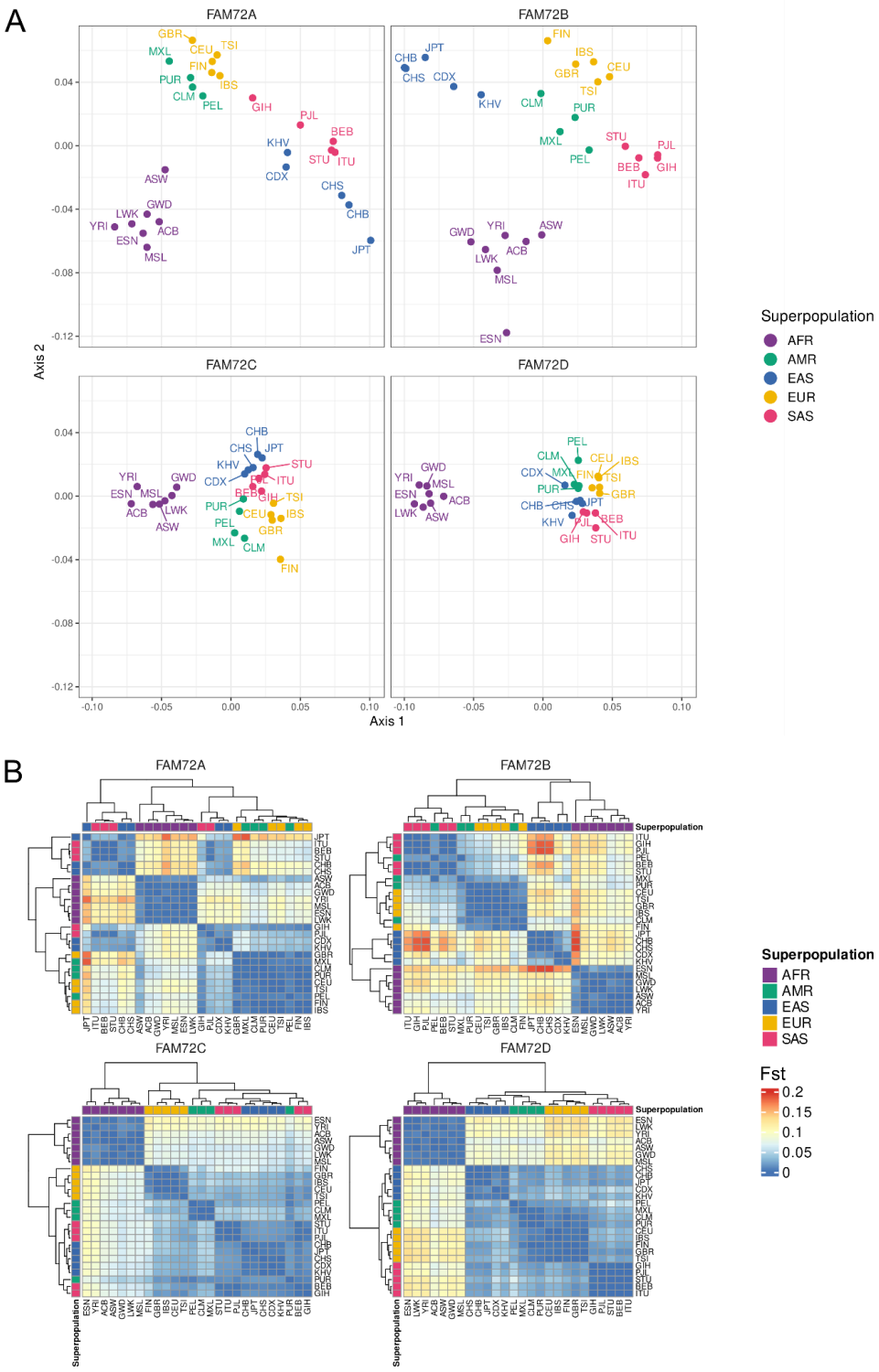

Figure 8. Population differentiation of *FAM72* paralogs across 26 populations from the 1000 Genomes project. A. Non-metric Multidimensional scaling of pairwise Fst values. B. A heatmap depicting pairwise Fst values between populations. Hierarchical clustering of populations was performed using Ward's method.

Both models showed statistically significant cluster assignments. However, the partial R^2, which quantifies the variance explained by reduced models in relation to the full model, showed a clear pattern across the loci. The biogeographical ancestry model had consistent performance across all four genes, whereas the Out-of-Africa (OoA) model performed well only for the *FAM72D* gene (Table 1).

Table 1. PERMANOVA results for out-of-Africa and biogeographical ancestry models.

	Model	*FAM72A*	*FAM72B*	*FAM72C*	*FAM72D*
P-value	Biogeographical ancestry	9.9×10^{-5}	9.9×10^{-5}	9.9×10^{-5}	9.9×10^{-5}
	Out-of-Africa	9.9×10^{-5}	9.9×10^{-5}	9.9×10^{-5}	9.9×10^{-5}
Partial R^2	Biogeographical ancestry	0.973	0.981	0.986	0.994
	Out-of-Africa	0.508	0.432	0.769	0.892

When examining pairwise Fst values between populations more closely (as shown in Figure 8B), distinct patterns of genetic differentiation were observed. *FAM72C* exhibited the least differentiation among populations, followed closely by *FAM72D*. On the other hand, *FAM72B* was the most differentiated gene. Interestingly, while African populations consistently showed differentiation from other superpopulations across all genes, the *FAM72B* gene is also characterized by exceptionally high Fst values between East Asian and South Asian populations.

3.2.2 Haplotype networks

To identify potential adaptive haplotypes that could drive genetic differentiation across populations, I analyzed haplotype frequencies and networks of the four *FAM72A*

paralogs across five superpopulations. When examining haplotype frequencies (Figure 9A), population-specific haplotypes were observed. *FAM72C* and *FAM72D* genes displayed a few dominant haplotypes, uniformly spread across the superpopulations. In contrast, *FAM72A* showed two primary haplotypes that had very low frequency in African groups. The haplotype diversity pattern of *FAM72B* diverged from the other paralogs, demonstrating six prevalent haplotypes, each bearing varied ancestry compositions.

To further examine gene- and ancestry-specific haplotype diversity, median-joining haplotype networks for each gene were constructed (Figure 9B). These networks visually represent evolutionary ties between haplotypes, positioning each based on its genetic relationship to others, based on location and frequency of mutations. Closely related haplotypes cluster together, while more genetically distant ones, and those harbouring rare alleles, remain apart. Notably, the haplotype network for *FAM72A-B* genes are similar, as well as those for *FAM72C* and *FAM72D* are more similar to each other. The *FAM72C* network forms a star-like cluster, marked by a central prevalent haplotype surrounded by minor, derived ones. The *FAM72D* gene, although less prominent, mirrors this pattern with the emergence of two derived haplotypes. Conversely, *FAM72A* and *FAM72B* genes have no overrepresented haplotype, instead presenting a chain of derived haplotypes. In *FAM72B*, the haplotypes form into two separate clusters, seemingly correlated with ancestry.

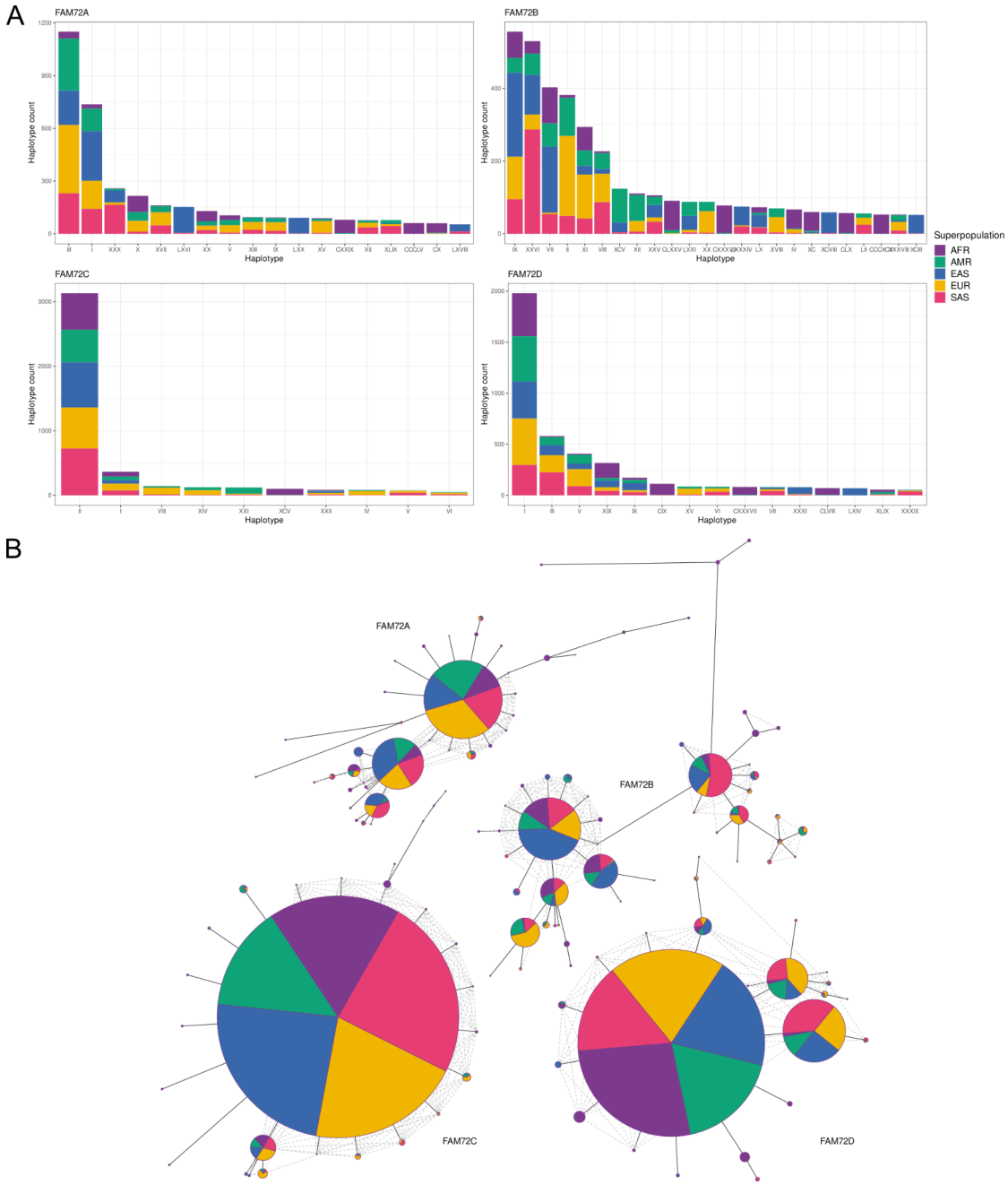

Figure 9. Haplotype composition of FAM72 genes. A. Haplotype frequencies of four *FAM72A* paralogs. Only haplotypes with > 50 observations were plotted. B. Median-joining haplotype networks of FAM72 genes. Sector areas are proportional to the number of observations, the colors represent different Superpopulations.

3.2.3 Genetic variation profile of *FAM72A* paralogs

Genetic variants across different functional genomic regions may provide insights into the evolutionary pressures acting on a gene. For instance, intronic variants can

modulate gene expression or alternative splicing events, whereas variants in the untranslated regions (5'UTRs and 3'UTRs) can disturb post-transcriptional regulation via mechanisms such as altered mRNA stability or microRNA interaction sites. In promoter regions, variants may impact transcriptional initiation kinetics and thus, the gene's overall expression profile. By systematically assessing the distribution and nature of variants across genomic compartments, it is possible to delineate the selective landscapes determining a gene's evolutionary trajectory.

I examined the distribution of known genetic variants in the *FAM72A* paralogs using data from the gnomAD v3.1.2 dataset (Chen et al. 2022). These variants were categorized based on their location in the genome's functional regions (Table 2). Counts of nonsense, synonymous, frameshift exonic variants, as well as 3’-UTR and splice site polymorphisms, were similar across all four *FAM72* genes. However, there was a distinct pattern of variation between *FAM72A-B* and *FAM72C-D*. Specifically, *FAM72A-B* had a notably higher number of promoter and intronic variants. Also, the higher ratio of missense to synonymous variants in *FAM72B* suggests possible balancing selection.

Table 2. Number of different classes of polymorphisms in FAM72A-D genes (data from gnomAD v3.1.2).

Location		*FAM2A*	*FAM72B*	*FAM72C*	*FAM72D*
Promoter		341	305	0	0
5'-UTR		17	142	6	9
Intron		1372	1556	989	940
Exon	Synonymous	4	7	1	3
	Missense	9	33	1	0
	Nonsense	0	1	1	0
	Frameshift	1	3	0	2
Splice site		1	1	0	0
3'-UTR		79	75	85	64

Since some protein domains are more tolerant to amino acid substitutions due to differences in evolutionary constraints on protein structure, I examined the distribution of protein-coding variants along the FAM72 peptides (Figure 10). As seen before, FAM72A-B proteins are characterized by a larger number of observed variants than FAM72C-D. In FAM72A-B proteins, there is a noticeable trend of increased variation in the C-terminus of the protein. Moreover, FAM72B stands out with an additional region of variability located between the 80th and 100th amino acid positions. The analysis of missense variants in *FAM72A-D* genes using data from the 1000 Genomes dataset showed that variations tend to be specific to individual superpopulations, as illustrated in Figure S3.

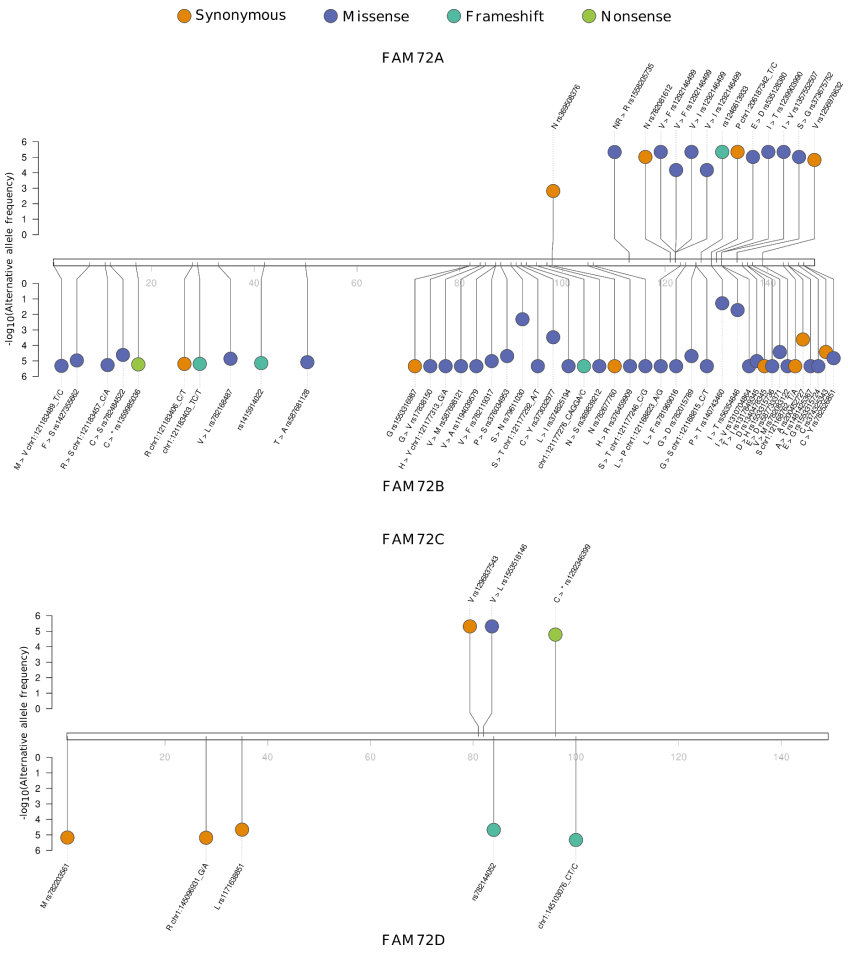

Figure 10. Distribution and frequency of coding variants within FAM72A-D genes. The y-axis displays the variant frequency as negative $\log_{10}$, while the x-axis denotes the amino acid positions impacted by these genetic variants.

3.3 Natural selection of *FAM72A* paralogs

3.3.1 Phylogenetic protein conservation

Selective pressures that act upon functionally important regions of a protein result in the preservation and maintenance of the protein's structural and functional integrity across diverse lineages. In particular, amino acids that are central to a protein's catalytic activity, ligand binding, or interaction with other macromolecules demonstrate a high degree of conservation that indicates strong purifying selection acting to eliminate deleterious mutations. Conversely, regions of a protein that are subject to less stringent

structural constraints, or that are involved in lineage-specific adaptations, may display signatures of positive selection, with an elevated rate of non-synonymous substitutions. These observations could indicate adaptive evolution, potentially driven by co-evolutionary arms races with interacting human or pathogen proteins.

To identify potential regions of functional or structural importance, I aligned FAM72A protein sequences from 58 taxa and identified stretches of high sequence conservation using 2 metrics of protein conservation. The results of the multiple sequence alignment and the conservation scores are visually presented in Figure 11. The analysis indicated that the FAM72A protein consists of regions of strong conservation, interspersed with variable regions. Specifically, the segments spanning amino acids 29-37, 66-80, and 93-105 demonstrated a notable degree of conservation. On the other end of the spectrum, the region around amino acids 134-145, located at the C-terminus of the protein, was identified as the most variable. Interestingly, the analysis also identified certain variable sites that appear to be lineage-specific amino acid substitutions. The most prominent example is the transition of amino acids at 15 distinct positions (16, 18, 27, 52, 62, 65, 106, 108, 111, 112, 10, 138, 139, 142, 144), which separates fish from other vertebrate groups. Here, the transition involves a switch either from or to hydrophilic amino acids, underlining a clear evolutionary divergence between these groups.

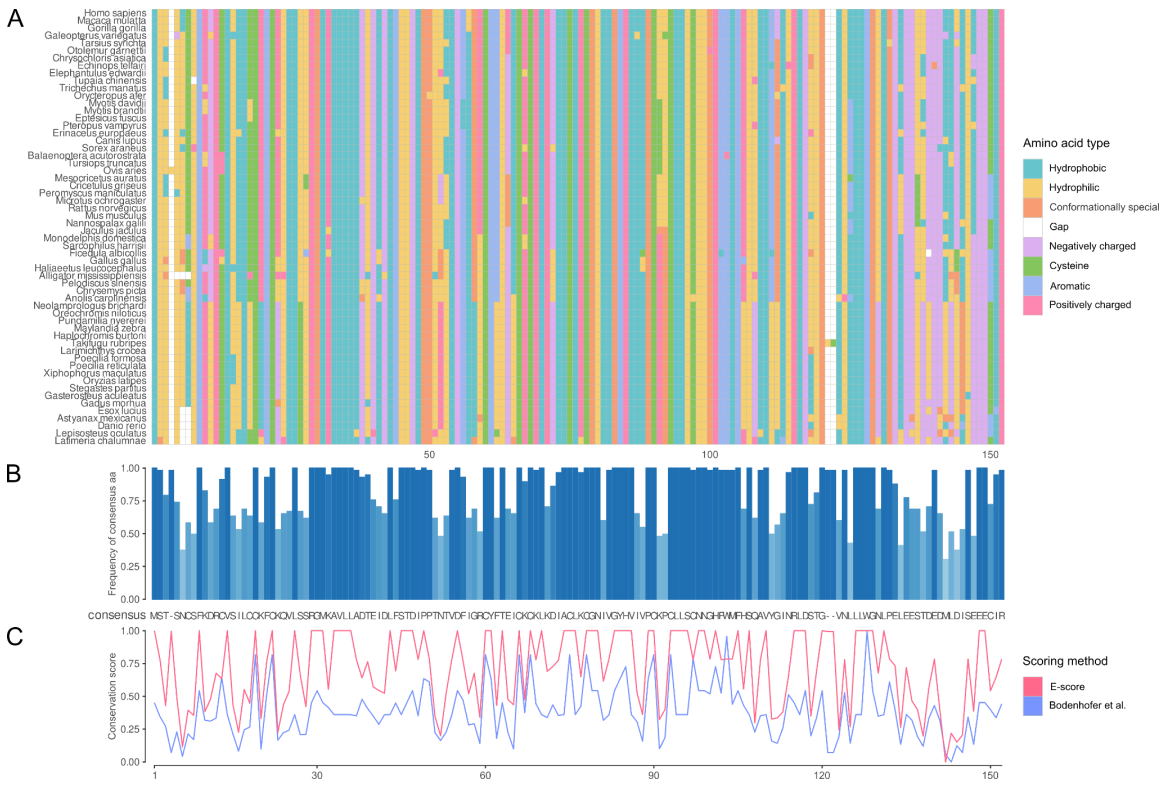

Figure 11. Phylogenetic conservation of FAM72A protein across vertebrates. A) Multiple sequence alignment of FAM72A proteins, and amino acids are colored according to the Zappo scheme. B) The consensus protein sequence and frequency of consensus amino acids in the alignment. C) Site-specific conservation scores based on Shannon entropy and substitution matrices.

3.3.2 Neutrality tests

Tajima's D, Fay and Wu's D, and Zeng's H are neutrality tests designed to detect deviations from neutral evolution in DNA sequences. Tajima's D involves comparing two different estimators of theta, where $\theta = 4N_e\mu$, and Ne is the effective population size and μ is the mutation rate. One estimate of θ is based on estimating the mean number of pairwise differences between sequences and is thus strongly influenced by alleles at an intermediate frequency ($\theta = \pi$), while the other estimator is based on the frequency of all segregating sites and is, therefore, more influenced by rare alleles (Watterson's theta, θw). At equilibrium the two estimates of θ are equal, and Tajima's D = 0, so that deviations from zero indicate potential non-neutral evolution or changes in N_e (Tajima 1989). Zeng's

H, on the other hand, contrasts rare and high-frequency variants to detect signs of recovery after a selective sweep (Zeng et al. 2006). Fay and Wu's D highlights an imbalance between high-frequency and intermediate-frequency alleles, incorporating ancestral allele information (Fay and Wu 2000). Extreme negative values of these metrics indicate signatures of positive selection.

Using the 1000 Genome dataset, I calculated three neutrality metrics and two diversities to test for signatures of a recent selective sweep at the *FAM72* loci. Null distributions for the metrics were calculated empirically across the entirety of chromosome 1 in non-overlapping windows of 17,387 bp in each superpopulation. The value of the observed metric for each *FAM72* paralog by superpopulation was considered significant if it lays in the top or bottom 5% of distribution. The empirical distributions created by this process reflect a range of values primarily influenced by the demographic dynamics of each population, serving as a reference to interpret the calculated neutrality tests' results.

The empirical distributions of the metrics were similar across superpopulations (Figure 12). However, the distributions in African populations were shifted to the right indicating higher genetic variation overall along chromosome 1. Among the five neutrality metrics applied, Tajima's D and Zeng's E consistently fell within the lowest 5% tail of the empirical distribution for the *FAM72C* gene across almost all superpopulations. However, Tajima's D values for the African population and Zeng's E values for the European population were exceptions to this trend. In addition, *FAM72D* had low Tajima's D values in four populations, with South Asians as the exception, and low Zeng's E values with Americans as the exception. For *FAM72A*, a low Tajima's D value

was observed in the American population, along with a low Zeng's E value in African populations. *FAM72B*, on the other hand, only demonstrated a low Zeng's E value in the South Asian population.

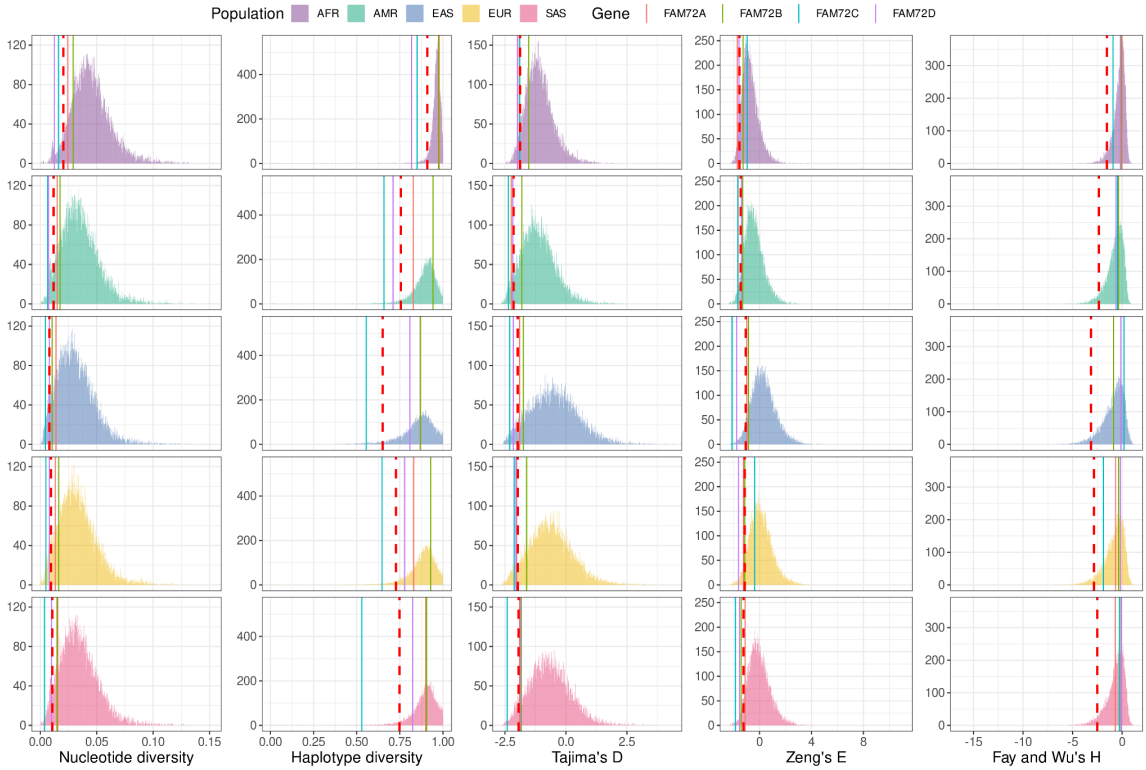

Figure 12. Comparative distributions of nucleotide diversity, haplotype diversity, and neutrality test metrics for *FAM72A-D* genes across superpopulations. Rows correspond to individual populations, while columns denote specific genetic metrics. The four solid lines in each graph signify the population-specific estimates for the *FAM72A-D* metrics, with the red dashed line marking the lower 5% tail threshold.

To analyze regional variations in Tajima's D values around the *FAM72-SRAGP2* loci, including a ±100 kb flanking region, I used overlapping 5kb sliding windows, shifting each by 500 bp (Figure 13). Globally, *FAM72A-SRGAP2* and *FAM72B-SRGAP2C* loci were characterized by shorter stretches of low Tajima's D values compared to other segmentally duplicated loci. Additionally, the downstream flanking region of *SRGAP2* paralogs displayed notable differences between these two pairs of loci. The former pair had a higher SNP density and Tajima's D values. Conversely, the

downstream flanking regions of *FAM72C-D* exhibited lower Tajima's D values in comparison to the analogous region of the older paralog pair.

Figure 13. Regional fluctuations of Tajima's D values in *FAM72-SRGAP2* ±100 kb loci. The upper part of each panel depicts Tajima's D variation in five superpopulations in 5 kb windows with a 500 bp shift. The dashed lines indicate standard Tajima's D cut-off values, indicating that a gene is not evolving neutrally. The lower part of the panel contains the canonical Ensembl gene structure.

Across all four *FAM72* paralogs, the second exon consistently demonstrated a drop in Tajima's D below -2. Additionally, there was a marked decrease in Tajima's D at the fourth exon of *FAM72B* and a less noticeable drop for the same exon in *FAM72A*. In contrast, the gene region between these exons displayed a sudden increase in Tajima's D. Meanwhile, for *FAM72C*, Tajima's D remained below -2 from the second exon to the middle of the third intron across three superpopulations. Tajima's D values remained above the cut-off across the *FAM72D* gene, except for a region around the second exon. Intriguingly, all *SRGAP2* paralogs, including the pseudogenized *SRGAP2D*, showed low Tajima's D values around the exons and within certain intron regions. Contrary to other paralogs, which showed a narrow spike in Tajima's D values in the middle of the second intron, *SRGAP2* had a ~33 kb stretch with high Tajima's D values. Overall, the observed regional Tajima's D patterns were more consistent among *SRGAP2* paralogs compared to *FAM72* genes.

3.3.3 Integrated haplotype score

The integrated haplotype score (iHS) is a statistic that identifies signals of recent positive selection in the genome based on patterns of linkage disequilibrium. The iHS compares the extent of haplotype homozygosity between derived and ancestral alleles at a given genomic position. In the case of a selective sweep, the frequency of beneficial alleles increases, leading to an extended region of a long haplotype around this variant that will slowly decay by recombination. Specifically, extended haplotype homozygosity (EHH) measures the probability of two randomly chosen haplotypes being identical by state over a specific distance. The iHH (integrated EHH) is then derived by integrating the EHH values to capture the area under the EHH curve. The iHS is the normalized log

ratio of iHH values for the derived versus the ancestral allele. Under neutrality, iHS values are expected to center around zero, whereas strong positive or negative values indicate selection.

Using phased 1000 Genomes data for chromosome 1 combined with the ancestral allele information from Ensembl, I calculated iHS scores across 26 human populations. I limited the inclusion criteria to SNPs with a minor allele frequency (MAF) greater than 1% and availability of ancestral allele state. MAF filtering led to a substantial reduction, resulting in only 20% of the original variants being eligible for iHS calculation.

The calculated iHS scores for the *FAM72-SRGAP2* loci, including a ± 10 kb flanking region, are depicted in Figure 14. Overall, the loci demonstrated a sparse distribution of significant iHS scores, with a majority of significant scores being associated with *SRGAP2* paralogs. African populations exhibited the largest number of selective sweep signals, although there were isolated signals in South Asian and American populations. Notably, European populations showed no such signals. Unfortunately, iHS calculation was not possible for *FAM72C-SRGAP2D* because of low nucleotide diversity and long monomorphic sequence stretches.

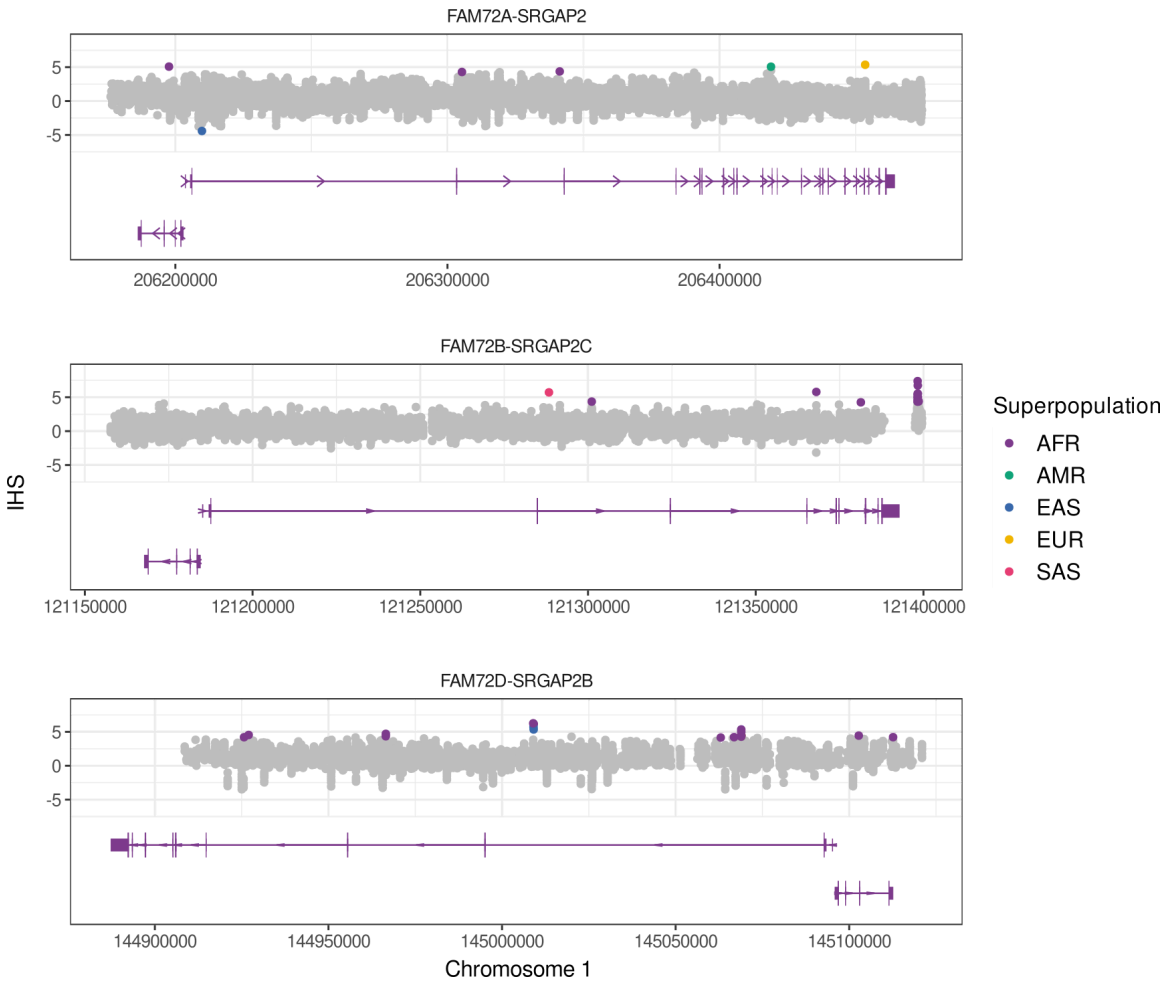

Figure 14 Population-specific integrated haplotype scores at *FAM72-SRGAP2* loci, extending ± 10 kb from the genes. Grey dots depict iHS values with an adjusted p-value greater than 0.05, while statistically significant iHS scores are denoted with colors corresponding to their respective superpopulation. Canonical gene structures are provided for *SRGAP2* (longer) and *FAM72* (shorter) paralogs as reference.

Focusing specifically on the *FAM72A-D* genes, only three statistically significant iHS scores were detected (Table 3). All three of these variants, residing within *FAM72A* and *FAM72D* genes, displayed significance exclusively in African populations. Two variants were situated in intronic regions, while the third was found in a 3'-UTR region.

Table 3. Genomic localization of statistically significant iHS scores.

Gene	Genomic position	Population	Superpopulation	Genomic context	iHS	-log10 (adj. p-value)
FAM72A	chr1:206197663	MSL	AFR	Intronic	5.079	2.499
FAM72D	chr1:145112636	GWD	AFR	3'-UTR	4.215	1.417
	chr1:145102742	YRI	AFR	Intronic	4.444	1.552

3.3.4 Long-term balancing selection ($\beta^{(1)}$ score)

Long-term balancing selection (LTBS) is an evolutionary process, maintaining allelic polymorphisms within a population across extensive temporal spans. LTBS may be explained by heterozygote advantage, negative frequency-dependent selection, or through ecological and spatially variable selection pressures. Balancing selection acts to preserve genetic diversity within populations, contributing to the adaptive potential of species in fluctuating environments (Bitarello et al. 2023).

Using the same dataset previously utilized for the iHS score analysis, I obtained standardized $\beta^{(1)}$ values, categorizing the upper 2% as extreme values. There was a notable variation in the number of significant values across different loci, as depicted in Figure 15. *FAM72A-SRGAP2* locus presented 611 significant $\beta^{(1)}$ values, predominantly concentrated between introns 10 and 15 of *SRGAP2*, whereas no signals were detected in the *FAM72A* gene body. The lowest number of statistically significant findings was observed in the *FAM72B-SRGAP2C* locus, with just 10 signals situated in intron 3 of *SRGAP2C*. In *FAM72C-SRGAP2D* and *FAM72D-SRGAP2B* 150 and 220 significant values were observed, respectively.

In *FAM72C*, 59 extreme values were detected (Table S5). Broadly, the potential signals of balancing selection are aggregated within two distinct regions. The initial

region spanned 2kb within intron 2 of the gene, encompassing 56 extreme values from African, American, and European populations. The second cluster of genomic positions, exclusively detected in the Luhya population from Webuye, Kenya, was situated in the third intron of *FAM72C*. In the case of *FAM72D*, three genomic positions were identified within a 541 bp region, where these variants were concurrently categorized as intronic variants of *FAM72D* and 2kb upstream variants of *SRGAP2B*.

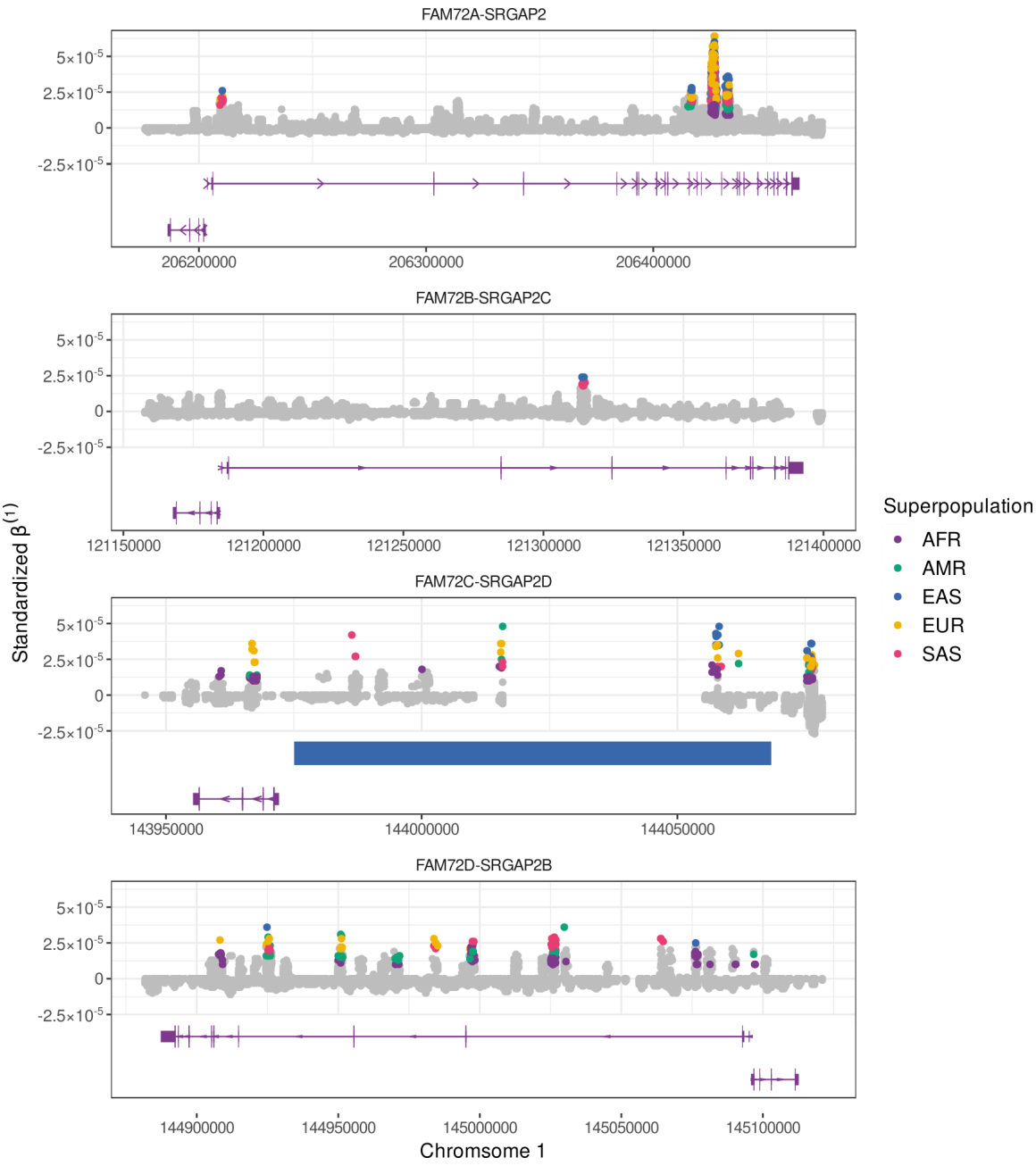

Figure 15. Population-specific $\beta^{(1)}$ values at *FAM72-SRGAP2* loci, extending ± 10 kb from the genes. Grey dots depict $\beta^{(1)}$ values within the lower 98%, while extreme $\beta^{(1)}$ values from the top 2% are denoted with colors corresponding to their respective superpopulation. Canonical gene structures are provided for *SRGAP2* (longer) and *FAM72* (shorter) paralogs as reference.

3.3.5 Linkage Disequilibrium (LD)

Linkage disequilibrium (LD) patterns reflect the action of evolutionary forces and demographic processes acting genome-wide or locally. For example, when a beneficial mutation arises, it can undergo a selective sweep, increasing in frequency along with the nearby alleles, resulting in a characteristic pattern of high LD. This pattern reflects the strength and recency of selection, with stronger and more recent sweeps creating larger regions of LD. Conversely, purifying selection removes deleterious mutations, which can also influence the LD patterns in the surrounding region. Balancing selection, on the other hand, maintains multiple alleles in the population, producing high LD due to the consistent frequencies of the selected alleles. To identify regions affected by these processes, I calculated correlation coefficients for every SNP pair in four *FAM72-SRGAP2* loci.

The analysis provided a detailed look at how the average pairwise LD (linkage disequilibrium) correlation coefficient varies significantly across different populations and genes. We observed a range from 0.01 to 0.134 in these coefficients, with an overall average of 0.039 across all populations and genes. Focusing on the individual *FAM72* genes, the Peruvian population showed the highest LD correlation for both *FAM72A* (r^2 = 0.134) and *FAM72B* (r^2 = 0.087), while the Gambian Mandinka population had the lowest for *FAM72A* (r^2 = 0.025), and Sri Lankan Tamils for *FAM72B* (r^2 = 0.031). For *FAM72C*, the Yoruba population displayed the highest LD (r^2 = 0.068), with Sri Lankan Tamils having the lowest (r^2 = 0.011). In the case of *FAM72D*, Colombians had the highest LD (r^2 = 0.039), and Southern Han Chinese had the lowest (r^2 = 0.01).

When the average pairwise LD correlation coefficients between the different *FAM72* genes were compared, a nearly two-fold difference between the *FAM72A-B* and *FAM72C-D* pairs was observed, with the former having higher values. However, when considering the median and interquartile range, *FAM72B* (0.00059, 0.0055) differed from the other paralogs (0.00024-0.00028, 0.001-0.0012).

Additionally, I looked at the physical location of SNP pairs that were in significant linkage disequilibrium ($r^2 \geq 0.75$). Here, the *FAM72C-D* genes had fewer highly correlated genetic variants compared to the *FAM72A-B* pair. Interestingly, these LD "hotspots" were mainly found in non-coding DNA regions, pointing to potential regions crucial for regulating gene expression. I also explored how the number of strong LD SNP pairs varied across different superpopulations. From this, a pattern emerged showing that African populations had a higher number and greater variability of these pairs, followed by South Asian populations, while European populations had the least variability (Figure S4).

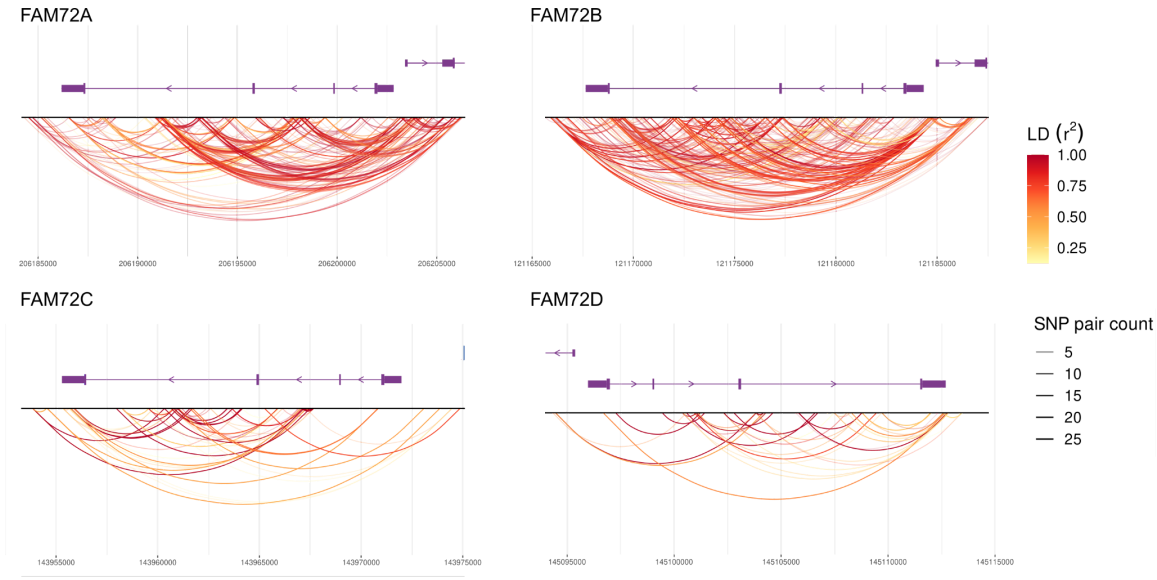

Figure 16. Distribution of variant pairs in strong LD ($r^2 > 0.75$) across *FAM72* paralogs. The black horizontal line represents a genomic region of *FAM72A* paralog with a ± 2kb flanking region. The gene structure of *FAM72* paralogs, alongside a truncated version of *SRGAP2* paralogs, is also depicted, according to the Ensembl gene model. The curves are color-coded to represent the LD r^2 values, while their intensity indicates the prevalence of a particular SNP pair with a strong LD association across various populations.

Next, my goal was to understand how the regulatory landscape is connected to the regions rich in LD (linkage disequilibrium) that were identified, especially since almost all the highest LD r^2 pairs were localized in non-coding parts of the genes. I focused on variants that consistently showed up in the top 2% LD r^2 across different populations. From these, I selected genomic positions that were in the upper 10% in terms of the number of extreme LD pairs in each gene and grouped them into 100 bp LD hotspots. I then used the Ensembl Regulatory Build (release 110) by Zerbino et al. (2015) to annotate these intervals.

In line with the previous observations, the *FAM72A-B* genes had a larger number of LD hotspots. The same high LD regions in the *FAM72A* gene were present across all five superpopulations. However, in some non-African superpopulations, there were

intervals in other paralogs that lacked strong LD signals. Through region-wise annotation, two promoters and two enhancer regions were identified — both regulatory features were only observed together in *FAM72B*. On the other hand, no regulatory elements were detected in the *FAM72D* gene, even though one of the LD regions showed the presence of transcription factor (TF) binding motifs (Table 4).

Using cell-type specific experimental epigenetic profiles integrated in the Ensembl Regulatory Build, I determined that both promoter regions are active in most cells, while both enhancers were mostly devoid of epigenetic marks (Table 4). Notably, the enhancer from *FAM72C* consistently showed an activated state in certain types of immune cells, such as monocytes, M0, M1, and M2 macrophages, but was either inactive or repressed in other immune cells. Conversely, the enhancer in *FAM72B* was almost entirely inactive, displaying signs of active chromatin only in human umbilical vein endothelial cells.

Table 4. Regulatory landscape of *FAM72A-D* LD hotspots.

Gene	Genomic region	P, n*	SP[þ]	Ensembl regulatory features	Regulatory feature activity, n				TF binding motifs
					A[⊙]	P[⊣⊢]	I[▫]	R[♦]	
FAM72A	chr1:206203200-206203300	7	All	Promoter (ENSR00000018900)	105	3	10	0	ETV2::FOXI1, CELK1::FOXI1, CERF::FOXI1, CFOXO1::ELF1, CFOXO1::ELK1, CFLI1::FOXI1, CETV5::FOXI1
	chr1:206203900-206204000	8	All						ELK1::HOXA3, CETV5::HOXA2, CETV5::EVX1, CHOXB2::ETV1, CHOXB2::ETV4, TEAD4::PAX5
FAM72B	chr1:121172100-121172200	5	All	Enhancer (ENSR00000930116)	2	0	111	4	FOXL1, CFOXC1, CFOXC2
	chr1:121184500-121184700	8	All	Promoter (ENSR00000012566)	100	3	15	0	E2F1::EOMES, KLF14, CSP4, CKLF16, CSP8, CSP3, CSP1, E2F3::TBX21, HOXB2::TBX21, HOXB2::RFX5
FAM72C	chr1:143953800-143953900	13	All	Enhancer (ENSR00000930227)	24	1	86	6	TBX19, CT, PAX6, GCM1::PITX1, GCM2::HES7, GCM2::SOX15, GCM2::PITX1
FAM72D	chr1:145102300-145102400	5	All	N/A	N/A	N/A	N/A	N/A	TEAD4::CLOCK, CTEAD4::HES7, TEAD4::PAX5

***** The count of populations in which a specified genomic segment contains positions identified among the upper 10% in terms of the prevalence of strong linkage disequilibrium (LD) pairings.

[þ] Superpopulations that encompass the populations listed in column "P".

[⊙] A regulatory element displays an active epigenetic signature, which can include evidence of open chromatin.

[⊣⊢] A regulatory element contains an epigenetic signature with the potential to be activated.

[▫] A regulatory element does not contain analyzed epigenetic marks.

[♦] A regulatory element contains epigenetic marks of expression repression

Chapter 4: Discussion

4.1 Phylogenetics and emergence of human-specific *FAM72A* paralogs

Globally, the estimates for the splits between *FAM72A* and *FAM72B-D*, and between *FAM72B* and *FAM72C-D* caused by duplication events align with those suggested by Dennis et al. (2012) — at 3.4 million years and 2.4 million years, respectively, with a slightly younger BEAST estimate for the former — 3.16 Mya. Additionally, the estimated age of a duplication that formed either *FAM72C* or *FAM72D* is older than the one from Dennis et al. (2012) — 1.54 Mya (95% HPD 0.51-2.83 Mya) versus 1 Mya (96%CI 0.4–1.3 Mya). However, due to the overlap in the wide credible intervals in the present analysis, a confident sequence of gene duplication events remains unknown. Yet, given the robust monophyletic nature of the *FAM72A* paralogs, a likely sequence of events inferred from sequence divergence would be *FAM72A*'s duplication giving rise to *FAM72B*, which subsequently led to the genesis of either both *FAM72C* and *FAM72D* or solely *FAM72C*, which then later duplicated into *FAM72D*.

The relaxed clock model estimated an average mutation rate of 6.8×10^{-4} substitutions per site per million years, although there is considerable uncertainty in this estimate, as indicated by the wide credible intervals ranging from 3.1×10^{-4} to 11.1×10^{-4}. This rate contrasts with direct, genome-wide estimates of mutation rates from studies of gorillas (6.5×10^{-4}, 95% CI 5.1×10^{-4}-7.9×10^{-4}), chimpanzees (6.4×10^{-4}, 95% CI 5.6×10^{-4}-7.3×10^{-4}) and humans (4.3×10^{-4}) by (Besenbacher et al. 2019). The discrepancy between these rates, along with the extensive high posterior density (HPD) intervals, might reflect highly divergent mutation rates across different clades, suggesting faster

evolutionary processes in the gorilla and chimpanzee lineages and notably slower rates in the clades containing *FAM72C* and *FAM72D*.

Another notable observation is the young age of modern *FAM72A-D* haplotypes, which seem to diverge in two waves: modern haplotypes of *FAM72A* and *FAM72B* started to emerge around 450 thousand years ago (Kya) before the emergence of *Homo sapiens* 300 Kya (Hublin et al. 2017; Scerri et al. 2018), whereas *FAM72C* and *FAM72D* haplotypes begun diverging around 190 Kya, coinciding with the emergence of anatomically modern humans (Stringer 2016).

The greatest strength and limitation of estimating phylogenetic divergence times using BEAST lies in the fossil calibrations, which are incorporated as prior probabilities on the ages of nodes. These calibrations are critical to molecular clock analyses, and empirical studies often face uncertainty in both the position and number of these calibration points. As a result, these factors can significantly influence the outcome of the analysis (Paradis 2013; Duchêne, Lanfear, and Ho 2014).

Another issue in Bayesian phylogenetic analyses is caused by the marginal priors for node ages that might deviate from the specified calibrations. The priors guide the Bayesian analysis by providing a starting point based on previous knowledge or assumptions. However, when multiple calibrations are not entirely consistent with each other either because of overlapping age intervals that they suggest for different nodes or because the actual tree topology does not align perfectly with the calibrated points, the marginal priors tend to deviate from the intended or specified calibrations, leading to inaccurate estimates of divergence timing (Heled and Drummond 2012).

In the present study, I limited fossil calibration to two prior distributions. For the calibration of the human-chimpanzee split, I used an unusually wide uniform distribution covering a span of 10 Myr and extensively overlapping with the prior for Homininae. This range is broader than what is typically used in the field (Dennis et al. 2012; de Manuel et al. 2016; Fontsere et al. 2022), and the rationale for these values is discussed by Vries and Beck (2023). To potentially refine the estimates, incorporating additional calibration points could be beneficial. For example, including *FAM72* genes from archaic hominins in the analysis might provide more calibration nodes.

The broad credible intervals (HPDs) encountered in divergence time estimates may stem from analyzing the full gene sequence as a single entity. Variability in base composition and mutation rates between the coding and noncoding regions of a gene can result in skewed divergence time estimates (Foster 2004; Kainer and Lanfear 2015). To address this issue, a more accurate approach could be adopted: partitioning the gene into its introns and exons, and further subdividing exons into untranslated regions (UTRs) and coding sequences (CDS). Partitioning allows the identification of the most likely substitution models for each distinct segment. Applying these models to each partition prior to the combined analysis in BEAST could potentially enhance the accuracy of the estimated divergence times.

4.2 Diversity of *FAM72A* paralogs

The observed low average pairwise Fst values (0.061 for *FAM72A*, 0.076 for *FAM72B*, 0.045 for *FAM72C*, and 0.052 for *FAM72D*) among the 26 populations from the 1000 Genomes dataset suggest a high homogeneity across the populations and loci examined. This pattern of homogeneity could be attributed to the shared recent ancestry

of human populations and possibly limited evolutionary divergence due to genetic drift or ongoing gene flow (Rosenberg et al. 2002; Weise et al. 2022; Z. Xu et al. 2022). The distinctively higher Fst value observed for the *FAM72B* gene between East Asian and South Asian populations may indicate unique selective forces or historical demographic events shaping its genetic variation differently (Campbell and Tishkoff 2008).

Population clustering analyses suggest that the *FAM72A* paralogs' genetic variation primarily correlates with biogeographical ancestry. Yet, the genetic variation in the *FAM72D* gene also correlates significantly with the Out-of-Africa model. The clustering pattern of African populations reflects genetic diversity within the continent and points to extensive gene flow among African populations. The Out-of-Africa model posits a genetic bottleneck effect during the migration from Africa approximately 60-75 Kya (Mallick et al. 2016; van Eeden et al. 2021). However, the variability in partial R^2 of this model for the *FAM72A-C* genes implies that factors beyond a simple bottleneck effect, such as different evolutionary forces, may have influenced these genes.

The haplotype analyses highlight the differentiation in the genetic architecture of *FAM72* paralogs. The predominance of certain haplotypes for *FAM72C* and *FAM72D* across diverse populations suggests purifying selection (Figure 9). These haplotypes are notably old and broadly represented, which does not align with the bottleneck effect typically associated with the out-of-Africa migration. This indicates that other evolutionary mechanisms have been at play. (Figure S2B).

Examining the spatial distribution of single nucleotide variants, *FAM72A* and *FAM72B* have a higher number of intronic variants compared to *FAM72C* and *FAM72D*. The young age of the *FAM72C-D* genes explains this difference. However, the abundance

of promoter mutations in *FAM72A-B* may result from relaxed selection, potentially leading to the diverse isoforms of these genes. Another potential explanation was proposed by Kutzner et al. (2015), who predicted the emergence of a new long-coding RNA in this region. Additionally, observations by Fraimovitch and Hagai (2023) that transcription factor binding motifs in promoters of young mammalian duplicate genes are initially quite similar but tend to diverge significantly over time also lend weight to the hypothesis that the observed promoter mutations could drive the evolution of new gene functions. As these promoters diverge, they may acquire unique regulatory features that allow the expression of gene copies to become specialized or adapted to new roles — a process that could be ongoing in the *FAM72* gene family.

FAM72B is further distinguished from other paralogs by accumulation of 5'-UTR and nonsynonymous coding variants, which might suggest ongoing relaxed selection of this gene. Additionally, the nonrandom pattern of nonsynonymous substitutions, particularly between the conserved 80^{th} and 100^{th} amino acid positions, hints at a potential ongoing pseudogenization process. For example, the Cys92Tyr variant caused by rs373032977, disrupts a predicted metal-binding site (Stewart and Bhagwat 2022), emphasizing the possibility of functional loss in this gene in the future.

In contrast to *FAM72A-B*, *FAM72C-D* are extremely conserved. Along with purifying selection, the observed sequence preservation of these genes can be maintained by an extensive interlocus gene conversion (IGC). IGC is a homologous recombination process that can cause non-allelic genes to become more similar to each other, effectively "homogenizing" sequences across different loci. Indeed, Vollger et al. (2023) assessed the distribution and frequency of interlocus gene conversion across human segmental

duplications and identified contrasting patterns of IGC between *FAM72* paralogs (Figure 17). They observed large-scale, bidirectional IGC events that span several dozens of kilobases in *FAM72C-D*. This extensive gene conversion could act to synchronize genetic variants between the paralogs, maintaining their sequence conservation. On the other hand, the *FAM72A-B* genes appear to undergo much smaller-scale IGC events, limited to regions around 1 kilobase. Furthermore, the asymmetric roles in IGC among these genes are notable. *FAM72B*'s status as a donor in IGC events being only once, while the other genes participated as donors more frequently (6-8 times), suggests differential evolutionary pressures or functional constraints on these genes. Meanwhile, *FAM72C* was an acceptor in IGC events 12 times, whereas the others played this role 2-4 times each.

The diversity analysis conducted faced certain constraints, primarily due to the decision to use phased VCF files. These files inherently lack singleton data because singletons cannot be reliably phased. Consequently, this absence likely leads to an underestimation of the true extent of population differentiation. Another limitation arises from the composition of the gnomAD v3.1 dataset, which has a disproportionate representation of European genomes. This bias potentially underestimates the full spectrum of the gene's global diversity. Moreover, the gnomAD dataset is not harmonized with the 1000 Genomes Project in terms of biogeographical ancestry or population, making direct comparisons unfeasible.

To enhance the robustness of future analyses, several improvements can be made. First, calculating Fst values at individual genomic positions could pinpoint specific genomic positions that contribute to genetic differentiation. Additionally, assessing

population-specific Fst and pairwise Fst values could aid in disentangling the evolutionary history of the genes from the current population structure. This comprehensive approach, which considers both historical and present-day population dynamics was advocated by Kitada, Nakamichi, and Kishino (2021).

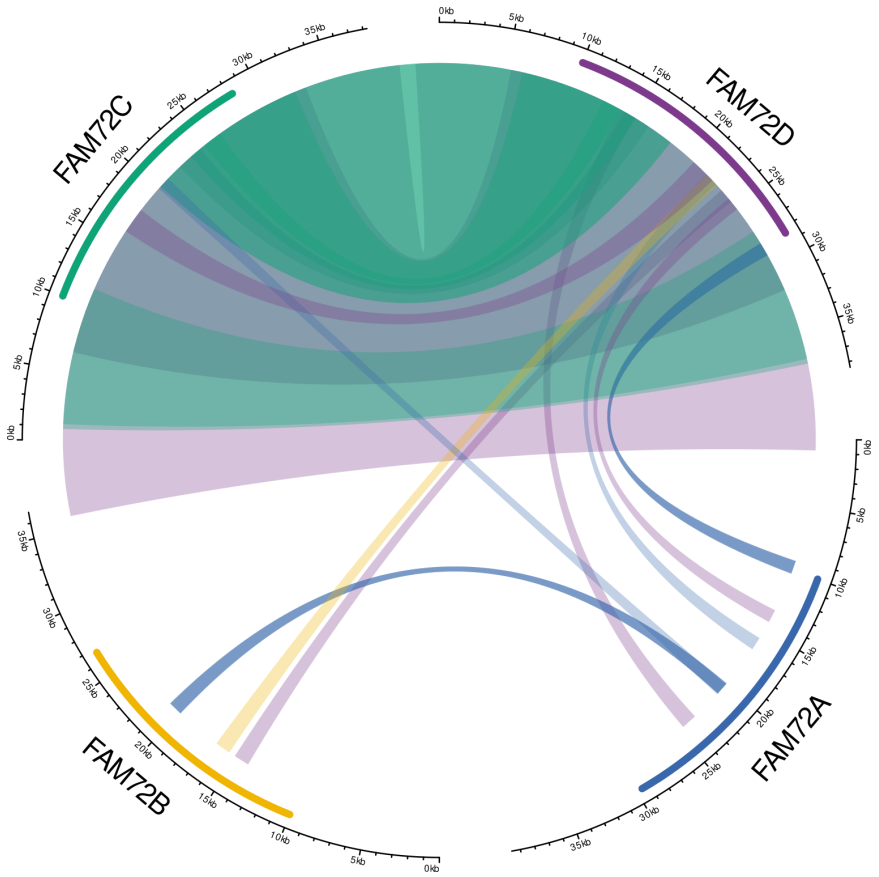

Figure 17. Interlocus gene conversion among *FAM72A* paralogs. The genes with flanking regions are depicted. The individual gene bodies are differentiated by a unique colour. Ribbons that link the loci signify specific IGC occurrences, with the color of each ribbon denoting the donor gene sequence and the width indicating the extent of the genomic region that underwent conversion.

4.3 Natural selection of FAM72A paralogs

The examination of FAM72A protein conservation across 58 taxa highlights regions within the proteins that show high conservation, suggesting these are critical for cellular functions. Specifically, amino acid sequences at positions 29-37, 66-80, and

93-105 are highly conserved, hinting at their role in catalytic activity, ligand binding, or macromolecule interaction. The functional importance of these regions is underscored by the presence of predicted metal-binding sites and by studies in mice showing that alterations in these sites disrupt binding with UNG2 (Stewart and Bhagwat 2022). Moreover, 4 out of 5 fixed amino acid differences of *FAM72* paralogs are located in extremely conserved protein regions. The d_N/d_S ratio of fixed substitutions in *FAM72C-D* relative to the ancestral *FAM72A* gene is 3 and 2, respectively. Kondrashov et al. (2002) found that most paralogs are characterized by symmetrical selection, and have $d_N/d_S << 1$. These findings may indicate a positive or relaxed selection (Zwonitzer et al. 2022), but the small number of amino acid differences precludes definitive conclusions.

The analyses of Tajima's D, Fay and Wu's H, and Zeng's E distributions suggest recent selective sweeps at *FAM72* loci, particularly affecting *FAM72C* and *FAM72D* genes. Low Tajima's D values across populations may reflect a narrowing of genetic diversity, possibly due to strong selection or population growth. Zeng's E and Tajima's D point to genes in a post-sweep recovery phase, where a selected haplotype becomes predominant and then begins to accumulate new genetic variants. In contrast, Fay and Wu's H identifies ongoing selective sweeps, placing *FAM72C* and *FAM72D* in the post-sweep stage (Zeng et al. 2006).

A regional analysis around *FAM72-SRAGP2* loci illustrates a mosaic of genetic diversity. The short stretches of low Tajima’s D values associated with the *FAM72A* and *FAM72B* loci indicate localized regions of low genetic diversity around exons. Conversely, the stretches of low Tajima’s D values throughout the whole *FAM72C* gene

may indicate the selection acting on noncoding sequences (Carlson et al. 2005; Naidoo et al. 2018).

The results from the integrated haplotype score (iHS) analysis indicate signals of recent positive selection, predominantly within the *SRGAP2* paralogs across the sampled human populations. These findings align with the existing literature that documents the role of *SRGAP2* in neural development and the evolution of cognitive functions in humans (Dennis et al. 2012; Sporny et al. 2017). The significance of the *SRGAP2* gene family in the evolution of human-specific traits may be inferred from the selective sweep signals predominantly found in African populations. The absence of significant iHS signals in European populations suggests a possible population-specific selective history or different selection pressures in non-African populations. However, the lack of iHS for *FAM72C-SRGAP2D* due to low nucleotide diversity underlines the potential limitations of using iHS in regions with monomorphic stretches, which can mask the presence of selective sweeps.

The $\beta^{(1)}$ score analysis suggests the presence of long-term balancing selection within the *SRGAP2* loci and *FAM72C* and *FAM72D* genes. The density of significant $\beta^{(1)}$ values in intronic regions of *FAM72-SRGAP2* loci may indicate spatio-temporal variation of fitness effects and selective pressures across the tissues and developmental stages (Wegmann, Dupanloup, and Excoffier 2008). The observed signals in intron 2 of *FAM72C* across diverse populations and the specific cluster in the Luhya population may point to global and population-specific ecological or pathogen-related selection pressures (Leffler et al. 2013). Additionally, the detected signals in noncoding regions of *FAM72C* and *FAM72D* may result from genotype-by-genotype interactions between humans and

viral pathogens leading to negative frequency-dependent selection (NFDS) (Bitarello et al. 2018; Råberg 2023). During arm race between viral pathogens and a host organism, pathogens typically evolve to target the most common host genotypes. As these prevalent host genotypes become more susceptible to infection, they are naturally selected against and decrease in frequency. This prompts the pathogens to adapt to other host genotypes, which, in turn, also become targeted and selected against. NFDS perpetuates a dynamic where multiple host alleles at specific genetic loci are retained over extended periods because the pathogen continually shifts its focus from one common host genotype to another, preventing any single allele from becoming universally dominant (Ebert and Fields 2020).

The linkage disequilibrium (LD) analysis across the *FAM72*A paralogs showed considerable variation in LD patterns among populations and between different genes. The lower LD in African populations such as the Gambian Mandinka might indicate a more extensive and older population history, consistent with the high genetic diversity that is characteristic of many African populations (Campbell and Tishkoff 2008). The outstanding prevalence of high LD in the Peruvian population can either reflect extensive admixture (Medina-Muñoz et al. 2023) or signatures of positive pathogen-driven selection that was reported in coast and rainforest Peruvian populations (Caro-Consuegra et al. 2022).

LD hotspots, primarily in noncoding regions, may be implicated in gene regulation. The fact that these hotspots were conserved across superpopulations for *FAM72A* but varied for other paralogs may reflect widespread preservation of the regulatory profile of the ancestral gene, but expression divergence of the derived genes

(Loker and Mann 2022; Aubé, Nielly-Thibault, and Landry 2023). The presence of transcription factor binding motifs in LD regions without identified regulatory elements in *FAM72D* suggests alternative regulatory mechanisms, such as chromatin remodelling, eQTLs or alternative splicing, may be at play. Ensembl Regulatory Build annotations reveal that LD hotspots in *FAM72C* are linked to immune cell function, while those in *FAM72B* are largely inactive, pointing to a possible diversification in function among the *FAM72* paralogs, potentially related to immune response.

The analysis of neutral test statistics faces a significant challenge due to the use of phased VCF files that lack singletons, which are crucial as these statistics rely on the variation in the site frequency spectrum. Moreover, the current approach to identifying linkage disequilibrium (LD) hotspots might benefit from a less biased *ad hoc* approach for cut-off choice. For instance, the LDna package (Kemppainen et al. 2015) represents LD data as a network graph, with loci as nodes and the LD correlation coefficients as edges. This package applies network analysis to discern clusters of loci that exhibit a denser or stronger LD with each other than with the rest of the network.

The methodological limitations inherent to short-read sequencing are amplified when analyzing segmentally duplicated genes with high levels of sequence similarity. In such contexts, differentiating between nearly identical sequences becomes a significant challenge, with short reads often leading to incorrect read mapping. Alignment ambiguity presents an additional obstacle. With multiple possible locations for short reads to map to, there is inherent uncertainty in variant calling. This uncertainty is particularly detrimental when detecting selection signals, as it can obscure the signatures of evolutionary pressures acting on these genes (Vollger et al. 2022). To overcome these issues, a shift

towards long-read sequencing techniques would be beneficial. For instance, the Institute of Molecular Pathology (Vienna, Austria) generated and recently released 1027 whole-genome sequences using Oxford Nanopore sequencing.

An issue in tests for selective sweeps concerning recently duplicated genes is the potential misinterpretation of their genetic signatures. Both selective sweeps and genetic drift can lead to similar patterns of reduced heterozygosity. In selective sweeps, this reduction occurs as beneficial mutations spread throughout a population, diminishing variation in neighbouring genomic regions due to linkage disequilibrium. Conversely, genetic drift can result in analogous patterns, especially within small populations where chance can rapidly fix a duplicated gene. Fixations by drift can create extended regions of genetic uniformity similar to those seen in selective sweeps, as the duplicated segment will be identical by descent. The neighbouring alleles can rise in frequency in tandem with the duplication, imitating the hitchhiking effect observed in selective sweeps. Selective sweeps are typically distinguished by an excess of rare alleles because a positively selected allele increases in frequency faster than recombination can introduce new variations. The fixation of a duplicated gene can lead to a similar distortion in the allele frequency spectrum (Kondrashov et al. 2002; Teshima, Coop, and Przeworski 2006; Thornton 2007).

4.4 General discussion

The analyses performed throughout this study reveal distinct evolutionary patterns between the gene pairs *FAM72A-B* and *FAM72C-D* (Figure 18). Phylogenetic evidence suggests that the *FAM72A-B* genes, along with the *SRGAP2* and *SRGAP2C* genes, were subject to strong selection pressures related to neocortex development up until about 450

Kya, coinciding with the emergence of *Homo sapiens* (Hublin et al. 2017) This period appears to have been followed by a relaxation in these selective pressures leading to haplotype diversification. Conversely, the *FAM72C* and *FAM72D* genes continued to evolve under constraint for an additional 200 thousand years, undergoing divergence around the dawn of anatomically modern humans (Stringer 2016). The retention of the *FAM72C* and *FAM72D* genes seems not to be linked to their cooperation with the co-duplicated counterparts in brain development, as the related genes *SRGAP2B* and *SRGAP2D* are either functionally redundant, highly variable, or have pseudogenized.

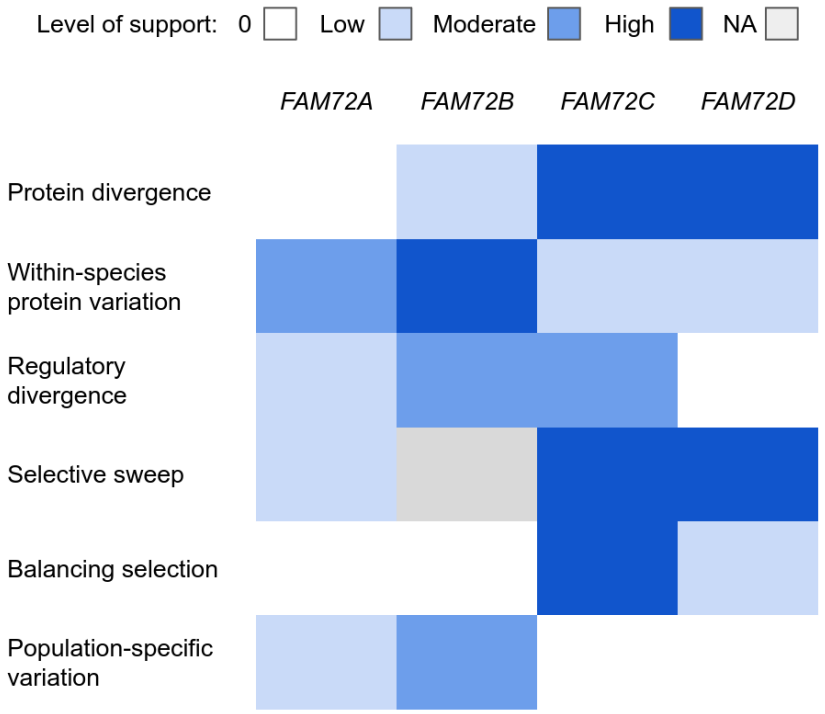

Figure 18. The summary of evidence for natural selection of *FAM72A* paralogs generated in this thesis.

To account for the retention of the *FAM72A* gene duplicates, two hypotheses are considered. The first, known as the "escape from adaptive conflict" (EAC) model, states that the ancestral gene had two essential but conflicting functions. Post-duplication, each gene could have specialized to improve one of these functions. The rapid build-up of

fixed amino acid changes in highly conserved regions of *FAM72C* and *FAM72D*, but not in the non-conserved C-terminus, might signal an evolutionary adaptation that favours the retention of mutations beneficial to one of the ancestral gene's roles. However, this model requires evidence of enhanced function in the ancestral *FAM72A*, which is not apparent since there is neither coding divergence from the sister Homininae lineages (Figure 11) nor divergence in gene expression (Shew et al. 2021).

The second hypothesis, the "innovation-amplification-divergence" (IAD) model, suggests that the ancestral gene had a primary essential function and a secondary, non-essential activity. If environmental changes made the secondary function advantageous, gene duplication would offer an immediate evolutionary benefit, likely leading to rapid fixation. Following fixation, the new gene could then specialize in the advantageous secondary function (Andersson, Jerlström-Hultqvist and Näsvall 2015). While the specific secondary activity of the ancestral *FAM72A* gene is not yet identified, its expression in lymphoid tissues, which are the primary sites of activity for the derived *FAM72C-D* genes, suggests that such a function exists. The d_N/d_S ratio pointing to functional divergence in *FAM72C-D* and the low genetic variability combined with low Tajima's D and Zeng's E values could indicate a fixation due to selective sweeps. Furthermore, a high frequency of balancing selection signals in the intronic regions of *FAM72C* and an LD hotspot in an intronic enhancer suggest regulatory divergence, potentially leading to specific expression in immune tissues.

The present analysis provides an initial look at the complex evolutionary trajectory of the human-specific *FAM72A* gene duplicates, but it is by no means exhaustive. The evolutionary framework suggested for the recent history of the *FAM72*

family is grounded in genomic data and theoretical models, but these require additional empirical corroboration. Future research may diverge along two principal approaches. Firstly, refining the present findings by utilizing data that is more comprehensive and less prone to bias will be essential. More advanced computational techniques could refine our understanding of the questions raised in this study. In-depth genomic analysis with improved data will enable a more detailed reconstruction of the evolutionary history of the *FAM72A* paralogs. Secondly, an in-depth examination of the cellular roles of the duplicated genes is required. The identified signals of balancing selection and selective sweep, along with the linkage disequilibrium (LD) hotspots discovered, need further investigation for the expression quantitative trait loci (eQTLs), which may be the driving forces of selection. Furthermore, these genomic regions should be analyzed for chromatin immunoprecipitation sequencing (ChIP-Seq) peaks in relevant to the hypotheses tissues and cells, such as neural progenitor cells and those involved in adaptive and innate immunity, using publicly accessible data sets.

After gene duplication events, it's common for paralogs to exhibit different coexpression relationships due to changes in their regulatory regions, leading to different patterns of gene expression. For example, if one paralog of a duplicated gene pair is highly coexpressed with genes involved in neurogenesis, while the other is coexpressed with genes in antiviral response, it suggests a functional divergence where each paralog has become integrated into different cellular processes. Gene coexpression networks (GCNs), where nodes represent genes, and edges represent significant coexpression relationships, typically inferred from large-scale transcriptomic datasets, can quantify the extent to which the expression of genes is synchronized across multiple conditions. These

measures are then used to build a network that can be analyzed to identify clusters or modules of highly interconnected genes, suggesting a group of genes working together in the same biological process. Thus, to address the functional divergence of *FAM72A* paralogs, gene co-expression network of transcriptomes of neuronal progenitor cells, activated B cells from germinal centers, and cells infected by a wide range of human herpesviruses (herpes simplex virus type 1 or 2, cytomegalovirus, Epstein-Barr virus) can be generated, followed determining whether these genes belong to the same or different co-expression modules.

Conclusions

In conclusion, this study of the recent evolutionary history of human-specific *FAM72A* paralogs provides support for Hypothesis 1, demonstrating a sequential duplication process where the ancestral *FAM72A* gene was duplicated to *FAM72B*, subsequently giving rise to *FAM72C* and *FAM72D* (Figure 7). The evidence from Bayesian divergence time estimation, structural conservation and divergence across these paralogs support this hypothesis.

Hypothesis 2 anticipates the presence of opposing selective forces acting on the *FAM72A* paralogs, with *FAM72A* and *FAM72B* showing patterns consistent with neutral evolution or balancing selection, and *FAM72C* and *FAM72D* exhibiting signs of positive selection. While findings indicate the overall asymmetric evolution of *FAM72A-B* and *FAM72C-D* pairs, the more complex evolutionary patterns were observed. Specifically, *FAM72A* and *FAM72B* paralogs appear to have been under relaxed selection allowing for variation in amino acid composition, while *FAM72C* and *FAM72D* paralogs show evidence recent selective sweeps. Yet, the findings suggest a complex mosaic of evolutionary forces: while the protein functions have been preserved at the amino acid level after fixation, there has been a simultaneous preservation of genetic variability in non-coding regions, potentially contributing to regulatory flexibility and adaptability.

Hypothesis 3 predicts emergence of population-specific adaptive haplotypes in the *FAM72A* paralogs, potentially driven by their role in the immune response. This study, however, did not find definitive proof of such population-specific haplotypes that confer localized advantage. While there is a suggestion of distinct haplotypes in *FAM72B* between East Asian and South Asian populations, the lack of clear adaptive signatures

may point to a more complex interplay between evolutionary forces and demographic history.

Further research, incorporating deeper genomic analysis and exploration into gene expression, is required to fully disentangle the evolutionary history of the *FAM72* gene family. More focused experimental and computational work could shed light on population-specific variations and environmental interactions of *FAM72A* paralogs in the context of the immune response.

References

Akiva, Pinchas, Amir Toporik, Sarit Edelheit, Yifat Peretz, Alex Diber, Ronen Shemesh, Amit Novik, and Rotem Sorek. 2006. “Transcription-Mediated Gene Fusion in the Human Genome.” Genome Research 16 (1): 30–36. https://doi.org/10.1101/gr.4137606.

Anderson, Marti J. 2001. “A New Method for Non-Parametric Multivariate Analysis of Variance.” Austral Ecology 26 (1): 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x.

Andersson, Dan, Jon Jerlström-Hultqvist, and Joakim Näsvall. 2015. “Evolution of New Functions De Novo and from Preexisting Genes.” Cold Spring Harbor Perspectives in Biology 7 (June). https://doi.org/10.1101/cshperspect.a017996.

Andrä, Paul. 2020. “Analysis and Functional Characterization in Embryonic Mouse Neocortex of a Set of Human-Specific Genes Expressed in Neural Progenitor Cells of Fetal Human Neocortex.” Doctor of Medicine, Dresden: Dresden University of Technology.

Arrigo, Nils, and Michael S Barker. 2012. “Rarely Successful Polyploids and Their Legacy in Plant Genomes.” Current Opinion in Plant Biology, Genome studies molecular genetics, 15 (2): 140–46. https://doi.org/10.1016/j.pbi.2012.03.010.

Atkinson, Elizabeth Grace, Amanda Jane Audesse, Julia Adela Palacios, Dean Michael Bobo, Ashley Elizabeth Webb, Sohini Ramachandran, and Brenna Mariah Henn. 2018. “No Evidence for Recent Selection at FOXP2 among Diverse Human Populations.” Cell 174 (6): 1424-1435.e15. https://doi.org/10.1016/j.cell.2018.06.048.

Aubé, Simon, Lou Nielly-Thibault, and Christian R. Landry. 2023. “Evolutionary Trade-off and Mutational Bias Could Favor Transcriptional over Translational Divergence within Paralog Pairs.” PLOS Genetics 19 (5): e1010756. https://doi.org/10.1371/journal.pgen.1010756.

Bailey, J. A., A. M. Yavor, H. F. Massa, B. J. Trask, and E. E. Eichler. 2001. “Segmental Duplications: Organization and Impact within the Current Human Genome Project Assembly.” Genome Research 11 (6): 1005–17. https://doi.org/10.1101/gr.gr-1871r.

Barkman, Todd, and Jianzhi Zhang. 2009. “Evidence for Escape from Adaptive Conflict?” Nature 462 (7274): E1–E1. https://doi.org/10.1038/nature08663.

Benayoun, Bérénice A., Elizabeth A. Pollina, Duygu Ucar, Salah Mahmoudi, Kalpana Karra, Edith D. Wong, Keerthana Devarajan, et al. 2014. “H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency.” Cell 158 (3): 673–88. https://doi.org/10.1016/j.cell.2014.06.027.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” Journal of the Royal Statistical Society: Series B (Methodological) 57 (1): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Besenbacher, Søren, Christina Hvilsom, Tomas Marques-Bonet, Thomas Mailund, and Mikkel Heide Schierup. 2019. “Direct Estimation of Mutations in Great Apes Reconciles Phylogenetic Dating.” Nature Ecology & Evolution 3 (2): 286–92. https://doi.org/10.1038/s41559-018-0778-x.

Bitarello, Bárbara D, Débora Y C Brandt, Diogo Meyer, and Aida M Andrés. 2023. “Inferring Balancing Selection From Genome-Scale Data.” Genome Biology and Evolution 15 (3): evad032. https://doi.org/10.1093/gbe/evad032.

Bodenhofer, Ulrich, Enrico Bonatesta, Christoph Horejš-Kainrath, and Sepp Hochreiter. 2015. “Msa: An R Package for Multiple Sequence Alignment.” Bioinformatics 31 (24): 3997–99. https://doi.org/10.1093/bioinformatics/btv494.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” Bioinformatics 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Bouckaert, Remco, Timothy G. Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, et al. 2019. “BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis.” PLoS Computational Biology 15 (4): e1006650. https://doi.org/10.1371/journal.pcbi.1006650.

Byrska-Bishop, Marta, Uday S. Evani, Xuefang Zhao, Anna O. Basile, Haley J. Abel, Allison A. Regier, André Corvelo, et al. 2022. “High-Coverage Whole-Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios.” Cell 185 (18): 3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.

Campbell, Michael C., and Sarah A. Tishkoff. 2008. “AFRICAN GENETIC DIVERSITY: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping.” Annual Review of Genomics and Human Genetics 9: 403–33. https://doi.org/10.1146/annurev.genom.9.081307.164258.

Cantsilieris, Stuart, Susan M. Sunkin, Matthew E. Johnson, Fabio Anaclerio, John Huddleston, Carl Baker, Max L. Dougherty, et al. 2020. “An Evolutionary Driver of Interspersed Segmental Duplications in Primates.” Genome Biology 21 (1): 202. https://doi.org/10.1186/s13059-020-02074-4.

Carelli, Francesco Nicola, Takashi Hayakawa, Yasuhiro Go, Hiroo Imai, Maria Warnefors, and Henrik Kaessmann. 2016. “The Life History of Retrocopies Illuminates the Evolution of New Mammalian Genes.” Genome Research 26 (3): 301–14. https://doi.org/10.1101/gr.198473.115.

Carlson, Christopher S., Daryl J. Thomas, Michael A. Eberle, Johanna E. Swanson, Robert J. Livingston, Mark J. Rieder, and Deborah A. Nickerson. 2005. “Genomic Regions Exhibiting Positive Selection Identified from Dense Genotype Data.” Genome Research 15 (11): 1553–65. https://doi.org/10.1101/gr.4326505.

Caro-Consuegra, Rocio, Maria A Nieves-Colón, Erin Rawls, Verónica Rubin-de-Celis, Beatriz Lizárraga, Tatiana Vidaurre, Karla Sandoval, et al. 2022. “Uncovering Signals of

Positive Selection in Peruvian Populations from Three Ecological Regions." Molecular Biology and Evolution 39 (8): msac158. https://doi.org/10.1093/molbev/msac158.

Chahwan, Richard, Winfried Edelmann, Matthew D Scharff, and Sergio Roa. 2012. "AIDing Antibody Diversity by Error-Prone Mismatch Repair." Seminars in Immunology 24 (4): 293–300. https://doi.org/10.1016/j.smim.2012.05.005.

Charrier, Cécile, Kaumudi Joshi, Jaeda Coutinho-Budd, Ji-Eun Kim, Nelle Lambert, Jacqueline de Marchena, Wei-Lin Jin, et al. 2012. "Inhibition of SRGAP2 Function by Its Human-Specific Paralogs Induces Neoteny during Spine Maturation." Cell 149 (4): 923–35. https://doi.org/10.1016/j.cell.2012.03.034.

Chen, Siwei, Laurent C. Francioli, Julia K. Goodrich, Ryan L. Collins, Masahiro Kanai, Qingbo Wang, Jessica Alföldi, et al. 2022. "A Genome-Wide Mutational Constraint Map Quantified from Variation in 76,156 Human Genomes." bioRxiv.

Clark, James W., and Philip C. J. Donoghue. 2018. "Whole-Genome Duplication and Plant Macroevolution." Trends in Plant Science 23 (10): 933–45. https://doi.org/10.1016/j.tplants.2018.07.006.

Conrad, Bernard, and Stylianos E. Antonarakis. 2007. "Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease." Annual Review of Genomics and Human Genetics 8 (1): 17–35. https://doi.org/10.1146/annurev.genom.8.021307.110233.

Cortesi, Fabio, Zuzana Musilová, Sara M. Stieb, Nathan S. Hart, Ulrike E. Siebeck, Martin Malmstrøm, Ole K. Tørresen, et al. 2015. "Ancestral Duplications and Highly Dynamic Opsin Gene Evolution in Percomorph Fishes." Proceedings of the National Academy of Sciences of the United States of America 112 (5): 1493–98. https://doi.org/10.1073/pnas.1417803112.

Crow, James Franklin, and Motoo Kimura. 1970. An Introduction to Population Genetics Theory. Burgess Publishing Company.

Cunningham, Fiona, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, et al. 2022. "Ensembl 2022." Nucleic Acids Research 50 (D1): D988–95. https://doi.org/10.1093/nar/gkab1049.

Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of SAMtools and BCFtools." GigaScience 10 (2): giab008. https://doi.org/10.1093/gigascience/giab008.

Davis, Jerel C., and Dmitri A. Petrov. 2005. "Do Disparate Mechanisms of Duplication Add Similar Genes to the Genome?" Trends in Genetics: TIG 21 (10): 548–51. https://doi.org/10.1016/j.tig.2005.07.008.

Dehal, Paramvir, and Jeffrey L. Boore. 2005. "Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate." PLoS Biology 3 (10): e314. https://doi.org/10.1371/journal.pbio.0030314.

Dennis, Megan Y., Lana Harshman, Bradley J. Nelson, Osnat Penn, Stuart Cantsilieris, John Huddleston, Francesca Antonacci, et al. 2017. “The Evolution and Population Diversity of Human-Specific Segmental Duplications.” Nature Ecology & Evolution 1 (3): 1–10. https://doi.org/10.1038/s41559-016-0069.

Dennis, Megan Y., Xander Nuttle, Peter H. Sudmant, Francesca Antonacci, Tina A. Graves, Mikhail Nefedov, Jill A. Rosenfeld, et al. 2012. “Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication.” Cell 149 (4): 912–22. https://doi.org/10.1016/j.cell.2012.03.033.

Dixon, Philip. 2003. “VEGAN, a Package of R Functions for Community Ecology.” Journal of Vegetation Science 14 (6): 927–30. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x.

Dougherty, Max L., Jason G. Underwood, Bradley J. Nelson, Elizabeth Tseng, Katherine M. Munson, Osnat Penn, Tomasz J. Nowakowski, Alex A. Pollen, and Evan E. Eichler. 2018. “Transcriptional Fates of Human-Specific Segmental Duplications in Brain.” Genome Research 28 (10): 1566–76. https://doi.org/10.1101/gr.237610.118.

Duchêne, Sebastián, Robert Lanfear, and Simon Y. W. Ho. 2014. “The Impact of Calibration and Clock-Model Choice on Molecular Estimates of Divergence Times.” Molecular Phylogenetics and Evolution 78 (September): 277–89. https://doi.org/10.1016/j.ympev.2014.05.032.

Eeden, Gerald van, Caitlin Uren, Marlo Möller, and Brenna M Henn. 2021. “Inferring Recombination Patterns in African Populations.” Human Molecular Genetics 30 (R1): R11–16. https://doi.org/10.1093/hmg/ddab020.

Fay, J. C., and C. I. Wu. 2000. “Hitchhiking under Positive Darwinian Selection.” Genetics 155 (3): 1405–13. https://doi.org/10.1093/genetics/155.3.1405.

Feng, Qinghua, John V. Moran, Haig H. Kazazian, and Jef D. Boeke. 1996. “Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition.” Cell 87 (5): 905–16. https://doi.org/10.1016/S0092-8674(00)81997-2.

Feng, Y, C Li, J Steward, P Barbulescu, N Desivo, A Álvarez-Quilón, R Pezo, et al. 2021. “FAM72A Antagonizes UNG2 to Promote Mutagenic Repair during Antibody Maturation.” Nature In Press.

Fiddes, Ian T., Gerrald A. Lodewijk, Meghan Mooring, Colleen M. Bosworth, Adam D. Ewing, Gary L. Mantalas, Adam M. Novak, et al. 2018. “Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis.” Cell 173 (6): 1356-1369.e22. https://doi.org/10.1016/j.cell.2018.03.051.

Fischer, Jan, Eduardo Fernández Ortuño, Fabio Marsoner, Annasara Artioli, Jula Peters, Takashi Namba, Christina Eugster Oegema, Wieland B. Huttner, Julia Ladewig, and Michael Heide. 2022. “Human-Specific ARHGAP11B Ensures Human-like Basal Progenitor Levels in Hominid Cerebral Organoids.” EMBO Reports 23 (11): e54728. https://doi.org/10.15252/embr.202254728.

Fontsere, Claudia, Martin Kuhlwilm, Carlos Morcillo-Suarez, Marina Alvarez-Estape, Jack D. Lester, Paolo Gratton, Joshua M. Schmidt, et al. 2022. “Population Dynamics and Genetic Connectivity in Recent Chimpanzee History.” Cell Genomics 2 (6): 100133. https://doi.org/10.1016/j.xgen.2022.100133.

Force, A, M Lynch, F B Pickett, A Amores, Y L Yan, and J Postlethwait. 1999. “Preservation of Duplicate Genes by Complementary, Degenerative Mutations.” Genetics 151 (4): 1531–45.

Fossati, Matteo, Rocco Pizzarelli, Ewoud R. Schmidt, Justine V. Kupferman, David Stroebel, Franck Polleux, and Cécile Charrier. 2016. “SRGAP2 and Its Human-Specific Paralog Co-Regulate the Development of Excitatory and Inhibitory Synapses.” Neuron 91 (2): 356–69. https://doi.org/10.1016/j.neuron.2016.06.013.

Foster, Peter G. 2004. “Modeling Compositional Heterogeneity.” Systematic Biology 53 (3): 485–95. https://doi.org/10.1080/10635150490445779.

Fraimovitch, Evgeny, and Tzachi Hagai. 2023. “Promoter Evolution of Mammalian Gene Duplicates.” BMC Biology 21 (1): 80. https://doi.org/10.1186/s12915-023-01590-6.

Garrison, Erik, and Gabor Marth. 2012. “Haplotype-Based Variant Detection from Short-Read Sequencing.” arXiv. https://doi.org/10.48550/arXiv.1207.3907.

Gautier, Mathieu, Alexander Klassmann, and Renaud Vitalis. 2017. “Rehh 2.0: A Reimplementation of the R Package Rehh to Detect Positive Selection from Haplotype Structure.” Molecular Ecology Resources 17 (1): 78–90. https://doi.org/10.1111/1755-0998.12634.

Gibbs, Richard A., George M. Weinstock, Michael L. Metzker, Donna M. Muzny, Erica J. Sodergren, Steven Scherer, Graham Scott, et al. 2004. “Genome Sequence of the Brown Norway Rat Yields Insights into Mammalian Evolution.” Nature 428 (6982): 493–521. https://doi.org/10.1038/nature02426.

Glinos, Dafni A., Garrett Garborcauskas, Paul Hoffman, Nava Ehsan, Lihua Jiang, Alper Gokden, Xiaoguang Dai, et al. 2022. “Transcriptome Variation in Human Tissues Revealed by Long-Read Sequencing.” Nature 608 (7922): 353–59. https://doi.org/10.1038/s41586-022-05035-y.

Goudet, Jérôme. 2005. “Hierfstat, a Package for r to Compute and Test Hierarchical F-Statistics.” Molecular Ecology Notes 5 (1): 184–86. https://doi.org/10.1111/j.1471-8286.2004.00828.x.

Guan, Yuanfang, Maitreya J Dunham, and Olga G Troyanskaya. 2007. “Functional Analysis of Gene Duplications in Saccharomyces Cerevisiae.” Genetics 175 (2): 933–43. https://doi.org/10.1534/genetics.106.064329.

Guo, Chunguang, Xiaodong Zhang, Stephen P Fink, Petra Platzer, Keith Wilson, James K. V. Willson, Zhenghe Wang, and Sanford D Markowitz. 2008. “Ugene, a Newly Identified Protein That Is Commonly over-Expressed in Cancer, and That Binds Uracil

DNA-Glycosylase." Cancer Research 68 (15): 6118–26. https://doi.org/10.1158/0008-5472.CAN-08-1259.

Hahn, Matthew W. 2009. "Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates." Journal of Heredity 100 (5): 605–17. https://doi.org/10.1093/jhered/esp047.

Hardwick, Simon A., Anoushka Joglekar, Paul Flicek, Adam Frankish, and Hagen U. Tilgner. 2019. "Getting the Entire Message: Progress in Isoform Sequencing." Frontiers in Genetics 10. https://www.frontiersin.org/articles/10.3389/fgene.2019.00709.

Hasegawa, Masami, Hirohisa Kishino, and Taka-aki Yano. 1985. "Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA." Journal of Molecular Evolution 22 (2): 160–74. https://doi.org/10.1007/BF02101694.

He, Xionglei, and Jianzhi Zhang. 2005. "Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution." Genetics 169 (2): 1157–64. https://doi.org/10.1534/genetics.104.037051.

Heide, Michael, Christiane Haffner, Ayako Murayama, Yoko Kurotaki, Haruka Shinohara, Hideyuki Okano, Erika Sasaki, and Wieland B. Huttner. 2020. "Human-Specific ARHGAP11B Increases Size and Folding of Primate Neocortex in the Fetal Marmoset." Science (New York, N.Y.) 369 (6503): 546–50. https://doi.org/10.1126/science.abb2401.

Heled, Joseph, and Alexei J. Drummond. 2012. "Calibrated Tree Priors for Relaxed Phylogenetics and Divergence Time Estimation." Systematic Biology 61 (1): 138–49. https://doi.org/10.1093/sysbio/syr087.

Heyworth, Paul G., Deborah Noack, and Andrew R. Cross. 2002. "Identification of a Novel NCF-1 (P47-Phox) Pseudogene Not Containing the Signature GT Deletion: Significance for A47 Degrees Chronic Granulomatous Disease Carrier Detection." Blood 100 (5): 1845–51. https://doi.org/10.1182/blood-2002-03-0861.

Ho, Nguyen Thi Thanh, Arne Kutzner, and Klaus Heese. 2019. "A Novel Divergent Gene Transcription Paradigm—the Decisive, Brain-Specific, Neural |-Srgap2–Fam72a-| Master Gene Paradigm." Molecular Neurobiology 56 (8): 5891–99. https://doi.org/10.1007/s12035-019-1486-5.

Ho, Nguyen Thi Thanh, Chinmay Satish Rahane, Subrata Pramanik, Pok-Son Kim, Arne Kutzner, and Klaus Heese. 2021. "FAM72, Glioblastoma Multiforme (GBM) and Beyond." Cancers 13 (5): 1025. https://doi.org/10.3390/cancers13051025.

Hoegg, Simone, Henner Brinkmann, John S. Taylor, and Axel Meyer. 2004. "Phylogenetic Timing of the Fish-Specific Genome Duplication Correlates with the Diversification of Teleost Fish." Journal of Molecular Evolution 59 (2): 190–203. https://doi.org/10.1007/s00239-004-2613-z.

Hublin, Jean-Jacques, Abdelouahed Ben-Ncer, Shara E. Bailey, Sarah E. Freidline, Simon Neubauer, Matthew M. Skinner, Inga Bergmann, et al. 2017. “New Fossils from Jebel Irhoud, Morocco and the Pan-African Origin of Homo Sapiens.” Nature 546 (7657): 289–92. https://doi.org/10.1038/nature22336.

Hsieh, PingHsun, Vy Dang, Mitchell R. Vollger, Yafei Mao, Tzu-Hsueh Huang, Philip C. Dishuck, Carl Baker, et al. 2021. “Evidence for Opposing Selective Forces Operating on Human-Specific Duplicated TCAF Genes in Neanderthals and Humans.” Nature Communications 12 (1): 5118. https://doi.org/10.1038/s41467-021-25435-4.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, et al. 2019. “eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses.” Nucleic Acids Research 47 (D1): D309–14. https://doi.org/10.1093/nar/gky1085.

Hughes, Graham M., Emma C. Teeling, and Desmond G. Higgins. 2014. “Loss of Olfactory Receptor Function in Hominin Evolution.” PLOS ONE 9 (1): e84714. https://doi.org/10.1371/journal.pone.0084714.

Hultqvist, Malin, Peter Olofsson, Jens Holmberg, B. Thomas Bäckström, Jesper Tordsson, and Rikard Holmdahl. 2004. “Enhanced Autoimmunity, Arthritis, and Encephalomyelitis in Mice with a Reduced Oxidative Burst Due to a Mutation in the Ncf1 Gene.” Proceedings of the National Academy of Sciences of the United States of America 101 (34): 12646–51. https://doi.org/10.1073/pnas.0403831101.

Human Protein Atlas. 2023. “RNA GTEx Tissue Gene Data.” https://www.proteinatlas.org/download/rna_tissue_gtex.tsv.zip.

Jiang, Zhaoshi, Haixu Tang, Mario Ventura, Maria Francesca Cardone, Tomas Marques-Bonet, Xinwei She, Pavel A. Pevzner, and Evan E. Eichler. 2007. “Ancestral Reconstruction of Segmental Duplications Reveals Punctuated Cores of Human Genome Evolution.” Nature Genetics 39 (11): 1361–68. https://doi.org/10.1038/ng.2007.9.

Jouffrey, V., A. S. Leonard, and S. E. Ahnert. 2021. “Gene Duplication and Subsequent Diversification Strongly Affect Phenotypic Evolvability and Robustness.” Royal Society Open Science 8 (6): 201636. https://doi.org/10.1098/rsos.201636.

Kainer, David, and Robert Lanfear. 2015. “The Effects of Partitioning on Phylogenetic Inference.” Molecular Biology and Evolution 32 (6): 1611–27. https://doi.org/10.1093/molbev/msv026.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, and Lars S. Jermiin. 2017. “ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates.” Nature Methods 14 (6): 587–89. https://doi.org/10.1038/nmeth.4285.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” Molecular Biology and Evolution 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.

Kemppainen, Petri, Christopher G. Knight, Devojit K. Sarma, Thaung Hlaing, Anil Prakash, Yan Naung Maung Maung, Pradya Somboon, Jagadish Mahanta, and Catherine Walton. 2015. “Linkage Disequilibrium Network Analysis (LDna) Gives a Global View of Chromosomal Inversions, Local Adaptation and Geographic Structure.” Molecular Ecology Resources 15 (5): 1031–45. https://doi.org/10.1111/1755-0998.12369.

Kimura, Motoo, and Tomoko Ohta. 1969. “The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population.” Genetics 61 (3): 763–71.

Kimura, Motoo. 1970. The Neutral Theory of Molecular Evolution. Cambridge University Press.

Kitada, Shuichi, Reiichiro Nakamichi, and Hirohisa Kishino. 2021. “Understanding Population Structure in an Evolutionary Context: Population-Specific FST and Pairwise FST.” G3 Genes|Genomes|Genetics 11 (11): jkab316. https://doi.org/10.1093/g3journal/jkab316.

Knaus, Brian J., and Niklaus J. Grünwald. 2017. “Vcfr: A Package to Manipulate and Visualize Variant Call Format Data in R.” Molecular Ecology Resources 17 (1): 44–53. https://doi.org/10.1111/1755-0998.12549.

Kondrashov, Fyodor A., and Alexey S. Kondrashov. 2006. “Role of Selection in Fixation of Gene Duplications.” Journal of Theoretical Biology 239 (2): 141–51. https://doi.org/10.1016/j.jtbi.2005.08.033.

Kutzner, Arne, Subrata Pramanik, Pok-Son Kim, and Klaus Heese. 2015. “All-or-(N)One – an Epistemological Characterization of the Human Tumorigenic Neuronal Paralogous FAM72 Gene Loci.” Genomics 106 (5): 278–85. https://doi.org/10.1016/j.ygeno.2015.07.003.

Langergraber, Kevin E., Kay Prüfer, Carolyn Rowney, Christophe Boesch, Catherine Crockford, Katie Fawcett, Eiji Inoue, et al. 2012. “Generation Times in Wild Chimpanzees and Gorillas Suggest Earlier Divergence Times in Great Ape and Human Evolution.” Proceedings of the National Academy of Sciences 109 (39): 15716–21. https://doi.org/10.1073/pnas.1211740109.

Leffler, Ellen M., Ziyue Gao, Susanne Pfeifer, Laure Ségurel, Adam Auton, Oliver Venn, Rory Bowden, et al. 2013. “Multiple Instances of Ancient Balancing Selection Shared between Humans and Chimpanzees.” Science (New York, N.Y.) 339 (6127): 1578–82. https://doi.org/10.1126/science.1234070.

Lenormand, Thomas, Thomas Guillemaud, Denis Bourguet, and Michel Raymond. 1998. “Appearance and Sweep of a Gene Duplication: Adaptive Response and Potential for New Functions in the Mosquito Culex Pipiens.” Evolution 52 (6): 1705–12. https://doi.org/10.2307/2411343.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” Bioinformatics (Oxford, England) 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li, Yupeng, Richard E. Higgs, Robert W. Hoffman, Ernst R. Dow, Xiong Liu, Michelle Petri, Daniel J. Wallace, et al. 2019. “A Bayesian Gene Network Reveals Insight into the JAK-STAT Pathway in Systemic Lupus Erythematosus.” Edited by Gang Han. PLOS ONE 14 (12): e0225651. https://doi.org/10.1371/journal.pone.0225651.

Liu, George E., Mario Ventura, Angelo Cellamare, Lin Chen, Ze Cheng, Bin Zhu, Congjun Li, Jiuzhou Song, and Evan E. Eichler. 2009. “Analysis of Recent Segmental Duplications in the Bovine Genome.” BMC Genomics 10 (1): 571. https://doi.org/10.1186/1471-2164-10-571.

Loker, Ryan, and Richard S. Mann. 2022. “Divergent Expression of Paralogous Genes by Modification of Shared Enhancer Activity through a Promoter-Proximal Silencer.” Current Biology 32 (16): 3545-3555.e4. https://doi.org/10.1016/j.cub.2022.06.069.

Lynch, Michael, and John S. Conery. 2000. “The Evolutionary Fate and Consequences of Duplicate Genes.” Science 290 (5494): 1151–55. https://doi.org/10.1126/science.290.5494.1151.

Maere, Steven, Stefanie De Bodt, Jeroen Raes, Tineke Casneuf, Marc Van Montagu, Martin Kuiper, and Yves Van de Peer. 2005. “Modeling Gene and Genome Duplications in Eukaryotes.” Proceedings of the National Academy of Sciences 102 (15): 5454–59. https://doi.org/10.1073/pnas.0501102102.

Mahjani, Behrang, Rebecca Birnbaum, Ariela Buxbaum Grice, Carolina Cappi, Seulgi Jung, Marina Natividad Avila, Abraham Reichenberg, et al. 2022. “Phenotypic Impact of Rare Potentially Damaging Copy Number Variation in Obsessive-Compulsive Disorder and Chronic Tic Disorders.” Genes 13 (10): 1796. https://doi.org/10.3390/genes13101796.

Makino, Takashi, and Aoife McLysaght. 2010. “Ohnologs in the Human Genome Are Dosage Balanced and Frequently Associated with Disease.” Proceedings of the National Academy of Sciences 107 (20): 9270–74. https://doi.org/10.1073/pnas.0914697107.

Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, et al. 2016. “The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations.” Nature 538 (7624): 201–6. https://doi.org/10.1038/nature18964.

Manuel, Marc de, Martin Kuhlwilm, Peter Frandsen, Vitor C. Sousa, Tariq Desai, Javier Prado-Martinez, Jessica Hernandez-Rodriguez, et al. 2016. “Chimpanzee Genomic Diversity Reveals Ancient Admixture with Bonobos.” Science (New York, N.Y.) 354 (6311): 477–81. https://doi.org/10.1126/science.aag2602.

Medina-Muñoz, Santiago G., Diego Ortega-Del Vecchyo, Luis Pablo Cruz-Hervert, Leticia Ferreyra-Reyes, Lourdes García-García, Andrés Moreno-Estrada, and Aaron P.

Ragsdale. 2023. “Demographic Modeling of Admixed Latin American Populations from Whole Genomes.” American Journal of Human Genetics 110 (10): 1804–16. https://doi.org/10.1016/j.ajhg.2023.08.015.

Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.” Molecular Biology and Evolution 37 (5): 1530–34. https://doi.org/10.1093/molbev/msaa015.

Moore, Richard C, and Michael D Purugganan. 2005. “The Evolutionary Dynamics of Plant Duplicate Genes.” Current Opinion in Plant Biology, Genome studies and molecular genetics / Plant biotechnology, 8 (2): 122–28. https://doi.org/10.1016/j.pbi.2004.12.001.

Naidoo, Thijessen, Per Sjödin, Carina Schlebusch, and Mattias Jakobsson. 2018. “Patterns of Variation in Cis-Regulatory Regions: Examining Evidence of Purifying Selection.” BMC Genomics 19 (1): 95. https://doi.org/10.1186/s12864-017-4422-y.

Näsvall, Joakim, Lei Sun, John R. Roth, and Dan I. Andersson. 2012. “Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence.” Science 338 (6105): 384–87. https://doi.org/10.1126/science.1226521.

Neuberger, Michael S., Reuben S. Harris, Javier Di Noia, and Svend K. Petersen-Mahrt. 2003. “Immunity through DNA Deamination.” Trends in Biochemical Sciences 28 (6): 305–12. https://doi.org/10.1016/S0968-0004(03)00111-7.

Nicholas, Thomas J., Ze Cheng, Mario Ventura, Katrina Mealey, Evan E. Eichler, and Joshua M. Akey. 2009. “The Genomic Architecture of Segmental Duplications and Associated Copy Number Variants in Dogs.” Genome Research 19 (3): 491–99. https://doi.org/10.1101/gr.084715.108.

Obenchain, Valerie, Michael Lawrence, Vincent Carey, Stephanie Gogarten, Paul Shannon, and Martin Morgan. 2014. “VariantAnnotation : A Bioconductor Package for Exploration and Annotation of Genetic Variants.” Bioinformatics 30 (14): 2076–78. https://doi.org/10.1093/bioinformatics/btu168.

O’Connell, Jared, Deepti Gurdasani, Olivier Delaneau, Nicola Pirastu, Sheila Ulivi, Massimiliano Cocca, Michela Traglia, et al. 2014. “A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness.” PLOS Genetics 10 (4): e1004234. https://doi.org/10.1371/journal.pgen.1004234.

Ohno, Susumu. 1970. Evolution by Gene Duplication. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-86659-3.

Okamura, Kohji, and Kenta Nakai. 2008. “Retrotransposition as a Source of New Promoters.” Molecular Biology and Evolution 25 (6): 1231–38. https://doi.org/10.1093/molbev/msn071.

Otsuka, Hiroshi, Akira Fukao, Yoshinori Funakami, Kent E. Duncan, and Toshinobu Fujiwara. 2019. “Emerging Evidence of Translational Control by AU-Rich Element-Binding Proteins.” Frontiers in Genetics 10. https://www.frontiersin.org/articles/10.3389/fgene.2019.00332.

Otto, Sarah P., and Paul Yong. 2002. “The Evolution of Gene Duplicates.” Advances in Genetics 46: 451–83. https://doi.org/10.1016/s0065-2660(02)46017-8.

Paradis, Emmanuel. 2010. “Pegas: An R Package for Population Genetics with an Integrated–Modular Approach.” Bioinformatics 26 (3): 419–20. https://doi.org/10.1093/bioinformatics/btp696.

Paradis, Emmanuel. 2013. “Molecular Dating of Phylogenies by Likelihood Methods: A Comparison of Models and a New Information Criterion.” Molecular Phylogenetics and Evolution 67 (2): 436–44. https://doi.org/10.1016/j.ympev.2013.02.008.

Paten, Benedict, Javier Herrero, Stephen Fitzgerald, Kathryn Beal, Paul Flicek, Ian Holmes, and Ewan Birney. 2008. “Genome-Wide Nucleotide-Level Mammalian Ancestor Reconstruction.” Genome Research 18 (11): 1829–43. https://doi.org/10.1101/gr.076521.108.

Paterson, A. H., J. E. Bowers, and B. A. Chapman. 2004. “Ancient Polyploidization Predating Divergence of the Cereals, and Its Consequences for Comparative Genomics.” Proceedings of the National Academy of Sciences 101 (26): 9903–8. https://doi.org/10.1073/pnas.0307901101.

Perry, George H., Nathaniel J. Dominy, Katrina G. Claw, Arthur S. Lee, Heike Fiegler, Richard Redon, John Werner, et al. 2007. “Diet and the Evolution of Human Amylase Gene Copy Number Variation.” Nature Genetics 39 (10): 1256–60. https://doi.org/10.1038/ng2123.

Perry, George H., Logan Kistler, Mary A. Kelaita, and Aaron J. Sams. 2015. “Insights into Hominin Phenotypic and Dietary Evolution from Ancient DNA Sequence Data.” Journal of Human Evolution 79 (February): 55–63. https://doi.org/10.1016/j.jhevol.2014.10.018.

Pfeifer, Bastian, Ulrich Wittelsbürger, Sebastian E. Ramos-Onsins, and Martin J. Lercher. 2014. “PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R.” Molecular Biology and Evolution 31 (7): 1929–36. https://doi.org/10.1093/molbev/msu136.

Płuciennik, Alicja, Michał Stolarczyk, Maria Bzówka, Agata Raczyńska, Tomasz Magdziarz, and Artur Góra. 2018. “BALCONY: An R Package for MSA and Functional Compartments of Protein Variability Analysis.” BMC Bioinformatics 19 (1): 300. https://doi.org/10.1186/s12859-018-2294-z.

Råberg, Lars. 2023. “Human and Pathogen Genotype-by-Genotype Interactions in the Light of Coevolution Theory.” PLOS Genetics 19 (4): e1010685. https://doi.org/10.1371/journal.pgen.1010685.

Rambaut, Andrew, Alexei J Drummond, Dong Xie, Guy Baele, and Marc A Suchard. 2018. “Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7.” Systematic Biology 67 (5): 901–4. https://doi.org/10.1093/sysbio/syy032.

Rapoport, IA. 1940. “Mnogokratnye Linejnye Povtoreniya Uchastkov Khromosom i Ikh Evolyucionnoe Znachenie.[Multiple Linear Repeats of Chromosome Segments and Their Evolutionary Significance].” Zh. Obshchej Biologii 1: 235–70.

Rastogi, Shruti, and David A Liberles. 2005. “Subfunctionalization of Duplicated Genes as a Transition State to Neofunctionalization.” BMC Evolutionary Biology 5 (April): 28. https://doi.org/10.1186/1471-2148-5-28.

Regier, Allison A., Yossi Farjoun, David E. Larson, Olga Krasheninina, Hyun Min Kang, Daniel P. Howrigan, Bo-Juen Chen, et al. 2018. “Functional Equivalence of Genome Sequencing Analysis Pipelines Enables Harmonized Variant Calling across Human Genetics Projects.” Nature Communications 9 (1): 4038. https://doi.org/10.1038/s41467-018-06159-4.

Renaud, Gabriel. 2018. “Glactools: A Command-Line Toolset for the Management of Genotype Likelihoods and Allele Counts.” Bioinformatics 34 (8): 1398–1400. https://doi.org/10.1093/bioinformatics/btx749.

Renganathan, Senthil, Subrata Pramanik, Rajasekaran Ekambaram, Arne Kutzner, Pok-Son Kim, and Klaus Heese. 2021. “Identification of a Chemotherapeutic Lead Molecule for the Potential Disruption of the FAM72A-UNG2 Interaction to Interfere with Genome Stability, Centromere Formation, and Genome Editing.” Cancers 13 (22): 5870. https://doi.org/10.3390/cancers13225870.

Rogier, Mélanie, Jacques Moritz, Isabelle Robert, Chloé Lescale, Vincent Heyer, Arthur Abello, Ophélie Martin, et al. 2021. “Fam72a Enforces Error-Prone DNA Repair during Antibody Diversification.” Nature 600 (7888): 329–33. https://doi.org/10.1038/s41586-021-04093-y.

Rosenberg, Noah A., Jonathan K. Pritchard, James L. Weber, Howard M. Cann, Kenneth K. Kidd, Lev A. Zhivotovsky, and Marcus W. Feldman. 2002. “Genetic Structure of Human Populations.” Science (New York, N.Y.) 298 (5602): 2381–85. https://doi.org/10.1126/science.1078311.

Russel, Patricio Maturana, Brendon J Brewer, Steffen Klaere, and Remco R Bouckaert. 2019. “Model Selection and Parameter Inference in Phylogenetics Using Nested Sampling.” Systematic Biology 68 (2): 219–33. https://doi.org/10.1093/sysbio/syy050.

Savva, Renos. 2020. “The Essential Co-Option of Uracil-DNA Glycosylases by Herpesviruses Invites Novel Antiviral Design.” Microorganisms 8 (3): 461. https://doi.org/10.3390/microorganisms8030461.

Scerri, Eleanor M. L., Mark G. Thomas, Andrea Manica, Philipp Gunz, Jay T. Stock, Chris Stringer, Matt Grove, et al. 2018. “Did Our Species Evolve in Subdivided

Populations across Africa, and Why Does It Matter?" Trends in Ecology & Evolution 33 (8): 582–94. https://doi.org/10.1016/j.tree.2018.05.005.

Schmidt, Ewoud R E, Justine V Kupferman, Michelle Stackmann, and Franck Polleux. 2019. "The Human-Specific Paralogs SRGAP2B and SRGAP2C Differentially Modulate SRGAP2A-Dependent Synaptic Development." Scientific Reports 9 (1): 18692. https://doi.org/10.1038/s41598-019-54887-4.

Shao, Yong, Long Zhou, Fang Li, Lan Zhao, Bao-Lin Zhang, Feng Shao, Jia-Wei Chen, et al. 2023. "Phylogenomic Analyses Provide Insights into Primate Evolution." Science 380 (6648): 913–24. https://doi.org/10.1126/science.abn6919.

She, Xinwei, Ze Cheng, Sebastian Zöllner, Deanna M. Church, and Evan E. Eichler. 2008. "Mouse Segmental Duplication and Copy Number Variation." Nature Genetics 40 (7): 909–14. https://doi.org/10.1038/ng.172.

Shen, Wei, Shuai Le, Yan Li, and Fuquan Hu. 2016. "SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation." PLOS ONE 11 (10): e0163962. https://doi.org/10.1371/journal.pone.0163962.

Shew, Colin J, Paulina Carmona-Mora, Daniela C Soto, Mira Mastoras, Elizabeth Roberts, Joseph Rosas, Dhriti Jagannathan, Gulhan Kaya, Henriette O'Geen, and Megan Y Dennis. 2021. "Diverse Molecular Mechanisms Contribute to Differential Expression of Human Duplicated Genes." Molecular Biology and Evolution 38 (8): 3060–77. https://doi.org/10.1093/molbev/msab131.

Siewert, Katherine M., and Benjamin F. Voight. 2017. "Detecting Long-Term Balancing Selection Using Allele Frequency Correlation." Molecular Biology and Evolution 34 (11): 2996–3005. https://doi.org/10.1093/molbev/msx209.

Sikosek, Tobias, Hue Sun Chan, and Erich Bornberg-Bauer. 2012. "Escape from Adaptive Conflict Follows from Weak Functional Trade-Offs and Mutational Robustness." Proceedings of the National Academy of Sciences 109 (37): 14888–93. https://doi.org/10.1073/pnas.1115620109.

Spofford, Janice B. 1969. "Heterosis and the Evolution of Duplications." The American Naturalist 103 (932): 407–32.

Sporny, Michael, Julia Guez-Haddad, Annett Kreusch, Sivan Shakartzi, Avi Neznansky, Alice Cross, Michail N. Isupov, Britta Qualmann, Michael M. Kessels, and Yarden Opatowsky. 2017. "Structural History of Human SRGAP2 Proteins." Molecular Biology and Evolution 34 (6): 1463–78. https://doi.org/10.1093/molbev/msx094.

Stewart, Jessica A., and Ashok S. Bhagwat. 2022. "A Redox-Sensitive Iron-Sulfur Cluster in Murine FAM72A Controls Its Ability to Degrade the Nuclear Form of Uracil-DNA Glycosylase." DNA Repair 118 (October): 103381. https://doi.org/10.1016/j.dnarep.2022.103381.

Sudmant, Peter H., John Huddleston, Claudia R. Catacchio, Maika Malig, LaDeana W. Hillier, Carl Baker, Kiana Mohajeri, et al. 2013. “Evolution and Diversity of Copy Number Variation in the Great Ape Lineage.” Genome Research 23 (9): 1373–82. https://doi.org/10.1101/gr.158543.113.

Tajima, F. 1989. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” Genetics 123 (3): 585–95.

Tang, Haibao, Xiyin Wang, John E. Bowers, Ray Ming, Maqsudul Alam, and Andrew H. Paterson. 2008. “Unraveling Ancient Hexaploidy through Multiply-Aligned Angiosperm Gene Maps.” Genome Research 18 (12): 1944–54. https://doi.org/10.1101/gr.080978.108.

Teshima, Kosuke M., Graham Coop, and Molly Przeworski. 2006. “How Reliable Are Empirical Genomic Scans for Selective Sweeps?” Genome Research 16 (6): 702–12. https://doi.org/10.1101/gr.5105206.

Thornton, Kevin R. 2007. “The Neutral Coalescent Process for Recent Gene Duplications and Copy-Number Variants.” Genetics 177 (2): 987–1000. https://doi.org/10.1534/genetics.107.074948.

Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. “Tissue-Based Map of the Human Proteome.” Science 347 (6220): 1260419. https://doi.org/10.1126/science.1260419.

Van de Peer, Yves, Steven Maere, and Axel Meyer. 2009. “The Evolutionary Significance of Ancient Genome Duplications.” Nature Reviews Genetics 10 (10): 725–32. https://doi.org/10.1038/nrg2600.

Van de Peer, Yves, Eshchar Mizrachi, and Kathleen Marchal. 2017. “The Evolutionary Significance of Polyploidy.” Nature Reviews Genetics 18 (7): 411–24. https://doi.org/10.1038/nrg.2017.26.

Vance, Zoe, and Aoife McLysaght. 2023. “Ohnologs and SSD Paralogs Differ in Genomic and Expression Features Related to Dosage Constraints.” Genome Biology and Evolution 15 (10): evad174. https://doi.org/10.1093/gbe/evad174.

Vollger, Mitchell R., Philip C. Dishuck, William T. Harvey, William S. DeWitt, Xavi Guitart, Michael E. Goldberg, Allison N. Rozanski, et al. 2023. “Increased Mutation and Gene Conversion within Human Segmental Duplications.” Nature 617 (7960): 325–34. https://doi.org/10.1038/s41586-023-05895-y.

Vollger, Mitchell R., Xavi Guitart, Philip C. Dishuck, Ludovica Mercuri, William T. Harvey, Ariel Gershman, Mark Diekhans, et al. 2022. “Segmental Duplications and Their Variation in a Complete Human Genome.” Science 376 (6588): eabj6965. https://doi.org/10.1126/science.abj6965.

Vries, Dorien de, and Robin M. D. Beck. 2023. “Twenty-Five Well-Justified Fossil Calibrations for Primate Divergences.” Palaeontologia Electronica 26 (1): 1–52. https://doi.org/10.26879/1249.

Walsh, J. B. 1995. “How Often Do Duplicated Genes Evolve New Functions?” Genetics 139 (1): 421–28.

Wegmann, Daniel, Isabelle Dupanloup, and Laurent Excoffier. 2008. “Width of Gene Expression Profile Drives Alternative Splicing.” PLOS ONE 3 (10): e3587. https://doi.org/10.1371/journal.pone.0003587.

Weir, B. S., and C. Clark Cockerham. 1984. “Estimating F-Statistics for the Analysis of Population Structure.” Evolution 38 (6): 1358–70. https://doi.org/10.2307/2408641.

Weise, Jessica A., Jillian Ng, Robert F. Oldt, Joy Viray, Kelly L. McCulloh, David Glenn Smith, and Sreetharan Kanthaswamy. 2022. “Genetic Differentiation between and within Northern Native American Language Groups: An Argument for the Expansion of the Native American CODIS Database.” Forensic Sciences Research 7 (4): 662–72. https://doi.org/10.1080/20961790.2021.1963088.

Xu, Shuangbin, Lin Li, Xiao Luo, Meijun Chen, Wenli Tang, Li Zhan, Zehan Dai, Tommy T. Lam, Yi Guan, and Guangchuang Yu. 2022. “Ggtree: A Serialized Data Object for Visualization of a Phylogenetic Tree and Annotation Data.” iMeta 1 (4): e56. https://doi.org/10.1002/imt2.56.

Yang, Ziheng. 1994. “Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods.” Journal of Molecular Evolution 39 (3): 306–14. https://doi.org/10.1007/BF00160154.

Yao, Yao, Lorenzo Carretero-Paulet, and Yves Van de Peer. 2019. “Using Digital Organisms to Study the Evolutionary Consequences of Whole Genome Duplication and Polyploidy.” PloS One 14 (7): e0220257. https://doi.org/10.1371/journal.pone.0220257.

Yokoyama, Shozo. 2008. “Evolution of Dim-Light and Color Vision Pigments.” Annual Review of Genomics and Human Genetics 9 (1): 259–82. https://doi.org/10.1146/annurev.genom.9.081307.164228.

Zeng, Kai, Yun-Xin Fu, Suhua Shi, and Chung-I Wu. 2006. “Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants.” Genetics 174 (3): 1431–39. https://doi.org/10.1534/genetics.106.061432.

Zerbino, Daniel R., Steven P. Wilder, Nathan Johnson, Thomas Juettemann, and Paul R. Flicek. 2015. “The Ensembl Regulatory Build.” Genome Biology 16 (1): 56. https://doi.org/10.1186/s13059-015-0621-5.

Zhang, Liang, Jacqueline Wax, Renliang Huang, Frank Petersen, and Xinhua Yu. 2022. “Meta-Analysis and Systematic Review of the Association between a Hypoactive NCF1 Variant and Various Autoimmune Diseases.” Antioxidants (Basel, Switzerland) 11 (8): 1589. https://doi.org/10.3390/antiox11081589.

Zwonitzer, Kendra D., Erik N. K. Iverson, James J. Sterling, Ryan J. Weaver, Bradley A. Maclaine, and Justin C. Havird. 2022. “Disentangling Positive vs. Relaxed Selection in Animal Mitochondrial Genomes.” bioRxiv. https://doi.org/10.1101/2022.10.05.510972.

Supplementary data

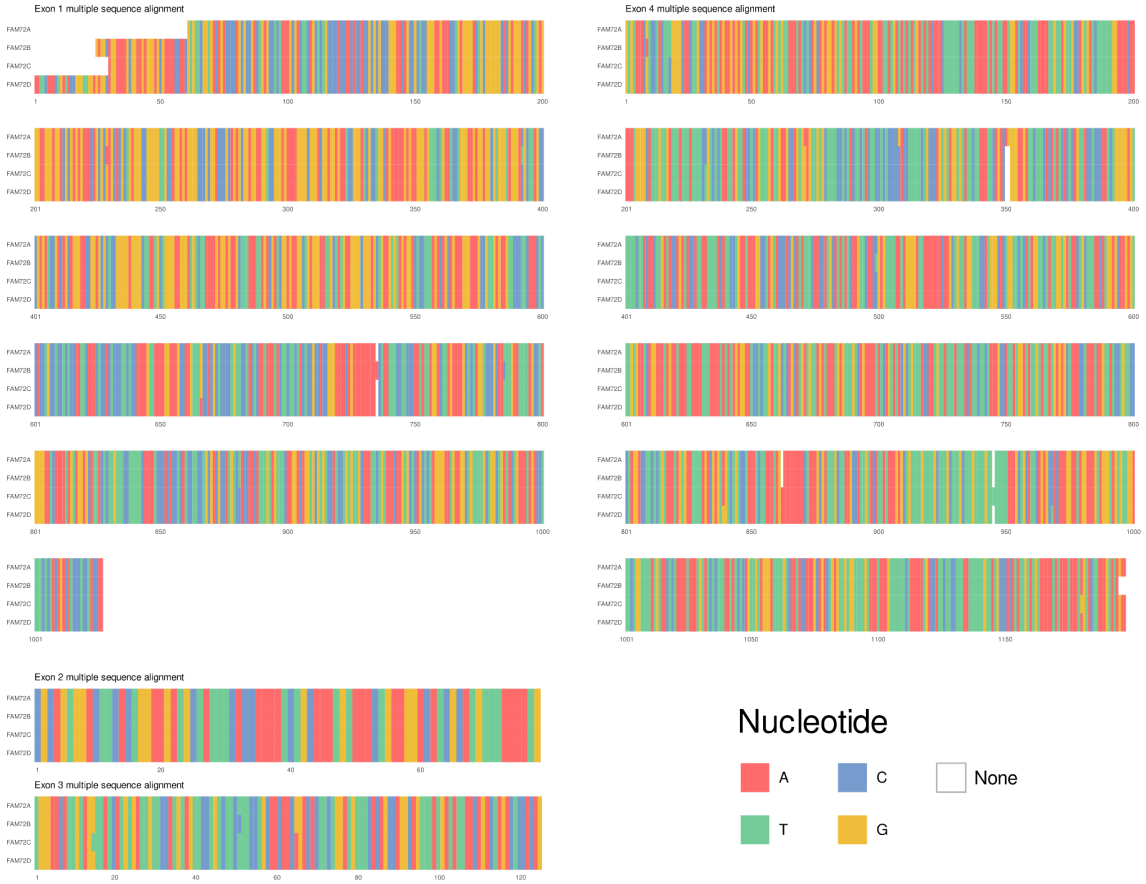

Figure S1. Multiple sequence alignment of *FAM72A-D* exons

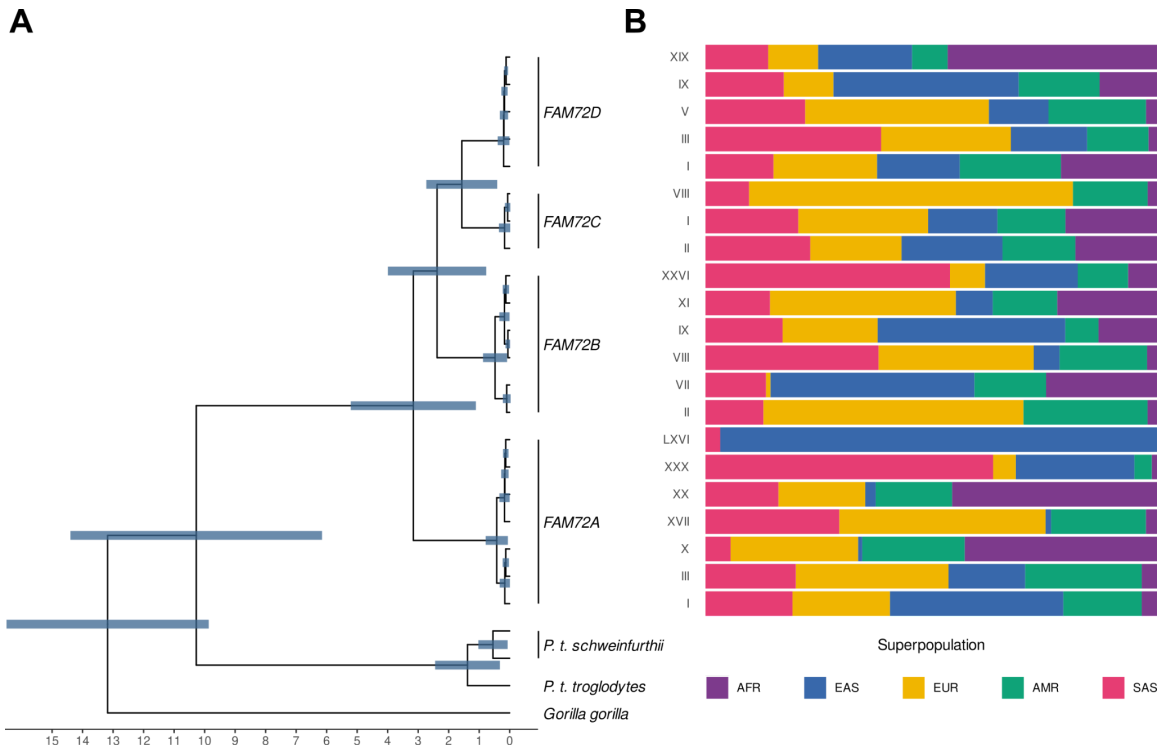

Figure S2. Genetic Lineage and Ancestry Distribution of *FAM72*. A. Phylogenetic tree illustrating the evolutionary divergence among the *FAM72* genes of great apes. B. Ancestry composition of the predominant human FAM72 haplotypes, positioned in relation to their branches on the phylogenetic tree.

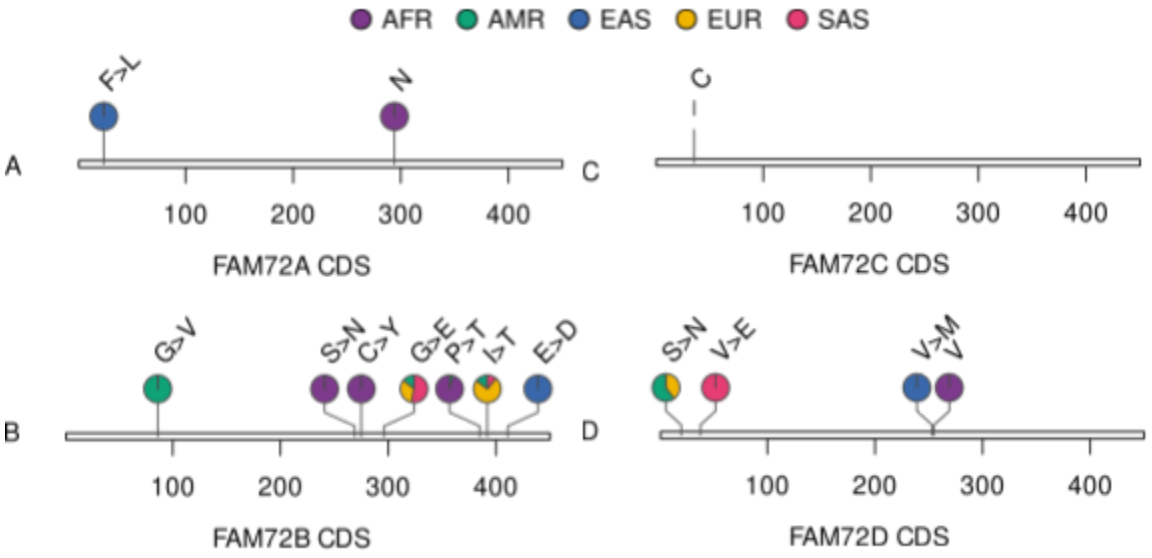

Figure S3. Distribution of coding FAM72A-D mutations across five superpopulations from the 1000 Genomes dataset. Variants are color-coded by their consequence, pie charts are proportional to the frequency of a variant in a given superpopulation

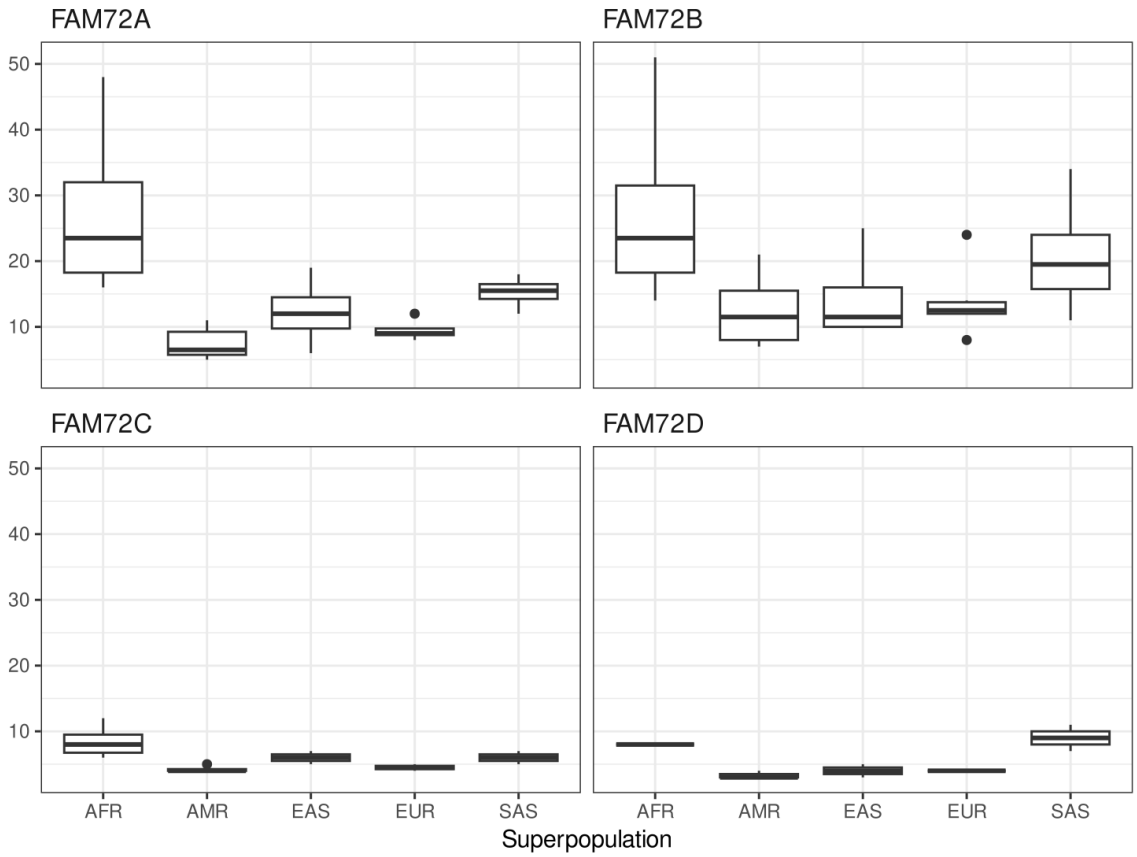

Figure S4. Number of distinct SNPs with extreme LD values in populations. The barplots representing superpopulations contain population counts of distinct genomic positions in extreme LD.

Table S1. Links to primary datasets used for the analysis.

Dataset	Link
High coverage WGS 1000 Genomes phased multisample VCF for chromosome 1	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/CCDG_14151_B01_GRM_WGS_2020-08-05_chr1.filtered.shapeit2-duohmm-phased.vcf.gz
1000 Genomes .ped file containing pedigrees and ancestry information	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/20130606_g1k_3202_samples_ped_population.txt
Ensembl variation VCF for chromosome 1	ftp://ftp.ensembl.org/pub/current_variation/vcf/homo_sapiens/homo_sapiens-chr1.vcf.gz
Ensembl 110 regulatory features for *Homo sapiens*	ftp://ftp.ensembl.org/pub/release-110/regulation/homo_sapiens/homo_sapiens.GRCh38.Regulatory_Build.regulatory_features.20221007.gff.gz
Ensembl 110 transcription factor binding motif features for *Homo sapiens*	ftp://ftp.ensembl.org/pub/release-110/regulation/homo_sapiens/MotifFeatures/homo_sapiens.GRCh38.motif_features.gff.gz
gnomAD v3.1 VCF for chromosome 1	https://storage.googleapis.com/gcp-public-data--gnomad/release/3.1.2/vcf/genomes/gnomad.genomes.v3.1.2.sites.chr1.vcf.bgz
Chimpanzee FASTQ files	https://www.ebi.ac.uk/ena/browser/view/PRJEB15086

Table S2. Description of 1000 genomes populations and the number of single nucleotide polymorphic sites identified in each population.

Superpopulation	Population code	Population description	Sample size	Number of polymorphic SNPs				
				Chromosome 1	FAM72A	FAM72B	FAM72C	FAM72D
AFR (African)	ACB	African Caribbean in Barbados	96	1618772	86	112	45	34
	ASW	African Ancestry in Southwest US	61	1456092	76	93	38	26
	ESN	Esan in Nigeria	99	1488298	76	79	38	29
	GWD	Gambian in Western Division, The Gambia	113	1574502	83	88	42	35
	LWK	Luhya in Webuye, Kenya	99	1522712	85	91	36	28
	MSL	Mende in Sierra Leone	85	1477686	75	90	42	32
	YRI	Yoruba in Ibadan, Nigeria	108	1541853	69	91	41	35
AMR (American)	CLM	Colombian in Medellin, Colombia	94	1183641	60	62	30	27
	MXL	Mexican Ancestry in Los Angeles, California	64	992144	35	60	21	16
	PEL	Peruvian in Lima, Peru	85	953031	36	52	26	16
	PUR	Puerto Rican in Puerto Rico	104	1343916	57	86	45	35
EAS (East Asian)	CDX	Chinese Dai in Xishuangbanna, China	93	798230	41	42	29	22
	CHB	Han Chinese in Beijing, China	103	824783	43	31	19	22
	CHS	Han Chinese South	105	897218	47	42	27	32
	JPT	Japanese in Tokyo, Japan	104	772818	30	30	19	30
	KHV	Kinh in Ho Chi Minh City, Vietnam	99	868685	41	40	27	27
EUR (European)	CEU	Utah residents (CEPH) with Northern and Western European ancestry	99	935214	52	52	18	26
	FIN	Finnish in Finland	99	827111	45	43	14	18

	GBR	British in England and Scotland	91	856167	37	44	17	20
	IBS	Iberian populations in Spain	107	1050948	44	63	21	26
	TSI	Toscani in Italia	107	924690	28	55	22	20
SAS (South Asian)	BEB	Bengali in Bangladesh	86	1004183	47	52	27	30
	GIH	Gujarati Indian in Houston, TX	103	950558	46	53	26	22
	ITU	Indian Telugu in the UK	102	968233	49	49	21	25
	PJL	Punjabi in Lahore, Pakistan	96	1021591	44	51	26	26
	STU	Sri Lankan Tamil in the UK	102	980794	50	54	23	26

Table S3. List of models tested by ModelFinder. The list is sorted by BIC scores. Plus signs denote the 95% confidence sets, minus signs denote significant exclusion.

Model	LogL	AIC	w-AIC	AICc	w-AICc	BIC	w-BIC
HKY+F+I	-30546.801	61197.601	2.5×10^{-3}	61197.886	2.52×10^{-3}	61607.012	1.65×10^{-1}
		-		-		+	
HKY+F+G4	-30546.812	61197.625	2.47×10^{-3}	61197.910	2.49×10^{-3}	61607.036	1.63×10^{-1}
		-		-		+	
TPM3+F+G4	-30541.891	61189.781	1.25×10^{-1}	61190.077	1.25×10^{-1}	61607.066	1.61×10^{-1}
		+		+		+	
TPM3u+F+G4	-30541.891	61189.782	1.25×10^{-1}	61190.078	1.25×10^{-1}	61607.066	1.61×10^{-1}
		+		+		+	
TPM3+F+I	-30541.895	61189.789	1.24×10^{-1}	61190.085	1.25×10^{-1}	61607.073	1.6×10^{-1}
		+		+		+	
TPM3u+F+I	-30541.896	61189.792	1.24×10^{-1}	61190.087	1.24×10^{-1}	61607.076	1.6×10^{-1}
		+		+		+	
K3Pu+F+I	-30545.065	61196.130	5.21×10^{-3}	61196.425	5.23×10^{-3}	61613.414	6.72×10^{-3}
		-		-		-	
K3Pu+F+G4	-30545.072	61196.145	5.17×10^{-3}	61196.441	5.19×10^{-3}	61613.429	6.67×10^{-3}
		-		-		-	
TN+F+I	-30546.371	61198.741	1.41×10^{-3}	61199.037	1.42×10^{-3}	61616.025	1.82×10^{-3}
		-		-		-	
TIM3+F+G4	-30541.442	61190.884	7.18×10^{-2}	61191.191	7.17×10^{-2}	61616.041	1.81×10^{-3}
		+		+		-	
TN+F+G4	-30546.382	61198.765	1.4×10^{-3}	61199.061	1.4×10^{-3}	61616.049	1.8×10^{-3}
		-		-		-	
TIM3+F+I	-30541.446	61190.892	7.15×10^{-3}	61191.199	7.14×10^{-2}	61616.050	1.8×10^{-3}
		+		+		-	
HKY+F+I+G4	-30546.508	61199.016	1.23×10^{-3}	61199.311	1.24×10^{-3}	61616.300	1.59×10^{-3}
		-		-		-	
TPM3+F+I+G4	-30541.598	61191.197	6.14×10^{-2}	61191.504	6.13×10^{-2}	61616.354	1.54×10^{-3}
		+		+		-	
TPM3u+F+I+G4	-30541.599	61191.199	6.13×10^{-2}	61191.506	6.12×10^{-2}	61616.356	1.54×10^{-3}
		+		+		-	
TPM2+F+I	-30546.739	61199.478	9.77×10^{-4}	61199.774	9.81×10^{-4}	61616.762	1.26×10^{-3}
		-		-		-	
TPM2u+F+I	-30546.739	61199.478	9.77×10^{-4}	61199.774	9.81×10^{-4}	61616.762	1.26×10^{-3}
		-		-		-	
TPM2+F+G4	-30546.749	61199.498	9.67×10^{-4}	61199.794	9.71×10^{-4}	61616.783	1.25×10^{-3}
		-		-		-	
TPM2u+F+G4	-30546.749	61199.498	9.67×10^{-4}	61199.794	9.71×10^{-4}	61616.783	1.25×10^{-3}
		-		-		-	
TIM+F+I	-30544.639	61197.278	2.93×10^{-3}	61197.585	2.93×10^{-3}	61622.436	7.39×10^{-5}
		-		-		-	
TIM+F+G4	-30544.647	61197.294	2.91×10^{-3}	61197.601	2.91×10^{-3}	61622.451	7.33×10^{-5}
		-		-		-	
K3Pu+F+I+G4	-30544.768	61197.535	2.58×10^{-3}	61197.842	2.58×10^{-3}	61622.693	6.50×10^{-5}
		-		-		-	
TVM+F+G4	-30540.966	61191.932	4.25×10^{-2}	61192.251	4.22×10^{-2}	61624.963	2.09×10^{-5}
		-		-		-	

TVM+F+I	-30540.98	61191.960	4.19×10^{-2}	61192.278	4.16×10^{-2}	61624.990	2.06×10^{-5}
TN+F+I+G4	-30546.079	61200.159	6.95×10^{-4}	61200.465	6.94×10^{-4}	61625.316	1.75×10^{-5}
TIM3+F+I+G4	-30541.151	61192.302	3.53×10^{-2}	61192.620	3.51×10^{-2}	61625.332	1.74×10^{-5}
TIM2+F+I	-30546.295	61200.589	5.6×10^{-4}	61200.896	5.6×10^{-4}	61625.747	1.41×10^{-5}
TIM2+F+G4	-30546.305	61200.610	5.55×10^{-4}	61200.917	5.54×10^{-4}	61625.768	1.40×10^{-5}
TPM2u+F+I+G4	-30546.446	61200.891	4.82×10^{-4}	61201.198	4.81×10^{-4}	61626.049	1.21×10^{-5}
TPM2+F+I+G4	-30546.446	61200.891	4.82×10^{-4}	61201.198	4.81×10^{-4}	61626.049	1.21×10^{-5}
TIM+F+I+G4	-30544.342	61198.683	1.45×10^{-3}	61199.002	1.44×10^{-3}	61631.714	7.14×10^{-7}
GTR+F+G4	-30540.532	61193.064	2.41×10^{-2}	61193.394	2.38×10^{-2}	61633.968	2.31×10^{-7}
GTR+F+I	-30540.548	61193.096	2.37×10^{-2}	61193.426	2.34×10^{-2}	61634.000	2.28×10^{-7}
TVM+F+I+G4	-30540.668	61193.336	2.11×10^{-2}	61193.666	2.08×10^{-2}	61634.240	2.02×10^{-7}
TIM2+F+I+G4	-30546.003	61202.006	2.76×10^{-4}	61202.324	2.74×10^{-4}	61635.036	1.36×10^{-7}
GTR+F+I+G4	-30540.236	61194.472	1.19×10^{-2}	61194.814	1.17×10^{-2}	61643.250	2.23×10^{-09}
GTR+F	-30553.937	61217.874	9.89×10^{-8}	61218.192	9.82×10^{-8}	61650.905	4.86×10^{-11}
GTR+F+R2	-30548.113	61210.226	4.53×10^{-6}	61210.567	4.45×10^{-6}	61659.003	8.47×10^{-13}
GTR+F+R3	-30543.716	61205.432	4.98×10^{-70}	61205.798	4.82×10^{-5}	61669.956	3.54×10^{-15}
F81+F+G4	-30702.186	61506.372	2.23×10^{-70}	61506.646	2.27×10^{-70}	61907.909	7.56×10^{-67}
F81+F+I	-30702.208	61506.416	2.18×10^{-70}	61506.690	2.22×10^{-70}	61907.954	7.39×10^{-67}
F81+F+I+G4	-30701.905	61507.810	1.09×10^{-70}	61508.095	1.10×10^{-70}	61917.221	7.19×10^{-69}
K2P+I	-30738.24	61574.480	3.62×10^{-85}	61574.734	3.72×10^{-85}	61960.271	3.22×10^{-78}
K2P+G4	-30738.391	61574.781	3.12×10^{-85}	61575.035	3.20×10^{-85}	61960.573	2.77×10^{-78}
K3P+I	-30737.011	61574.022	4.56×10^{-85}	61574.285	4.65×10^{-85}	61967.686	7.91×10^{-80}
K3P+G4	-30737.161	61574.322	3.92×10^{-85}	61574.586	4.00×10^{-85}	61967.986	6.81×10^{-80}
TNe+I	-30737.86	61575.720	1.95×10^{-85}	61575.983	1.99×10^{-85}	61969.384	3.38×10^{-80}
K2P+I+G4	-30737.951	61575.902	1.78×10^{-85}	61576.166	1.82×10^{-85}	61969.566	3.09×10^{-80}

TNe+G4	-30738.011	61576.022	1.68×10^{-85}	61576.286	1.71×10^{-85}	61969.687	2.91×10^{-80}
TIM3e+I	-30735.571	61573.142	7.07×10^{-85}	61573.416	7.18×10^{-85}	61974.680	2.40×10^{-81}
TIM3e+G4	-30735.783	61573.565	5.73×10^{-85}	61573.839	5.81×10^{-85}	61975.103	1.94×10^{-81}
TIMe+I	-30736.651	61575.302	2.40×10^{-85}	61575.576	2.44×10^{-85}	61976.840	8.14×10^{-82}
K3P+I+G4	-30736.718	61575.436	2.25×10^{-85}	61575.710	2.28×10^{-85}	61976.973	7.61×10^{-82}
TIMe+G4	-30736.798	61575.597	2.07×10^{-85}	61575.871	2.10×10^{-85}	61977.135	7.02×10^{-82}
TIM2e+I	-30737.516	61577.031	1.01×10^{-85}	61577.305	1.03×10^{-85}	61978.569	3.43×10^{-82}
TNe+I+G4	-30737.572	61577.144	9.56×10^{-86}	61577.418	9.71×10^{-86}	61978.682	3.24×10^{-82}
TIM2e+G4	-30737.692	61577.384	8.48×10^{-86}	61577.658	8.61×10^{-86}	61978.922	2.87×10^{-82}
TVMe+I	-30734.667	61573.334	6.43×10^{-85}	61573.619	6.49×10^{-85}	61982.745	4.25×10^{-83}
TVMe+G4	-30734.789	61573.578	5.69×10^{-85}	61573.863	5.74×10^{-85}	61982.989	3.76×10^{-83}
TIM3e+I+G4	-30735.279	61574.558	3.48×10^{-85}	61574.843	3.52×10^{-85}	61983.969	2.30×10^{-83}
TIMe+I+G4	-30736.36	61576.719	1.18×10^{-85}	61577.004	1.19×10^{-85}	61986.130	7.82×10^{-84}
TIM2e+I+G4	-30737.231	61578.461	4.95×10^{-86}	61578.746	5.00×10^{-86}	61987.872	3.27×10^{-84}
TVMe+I+G4	-30734.372	61574.744	3.18×10^{-85}	61575.040	3.19×10^{-85}	61992.028	4.10×10^{-85}
SYM+I	-30734.446	61574.892	2.95×10^{-85}	61575.188	2.96×10^{-85}	61992.176	3.80×10^{-85}
SYM+G4	-30734.561	61575.123	2.63×10^{-85}	61575.419	2.64×10^{-85}	61992.407	3.39×10^{-85}
SYM+I+G4	-30734.151	61576.302	1.46×10^{-85}	61576.609	1.46×10^{-85}	62001.460	3.67×10^{-87}
JC+I	-30889.226	61874.453	2.64×10^{-150}	61874.696	2.72×10^{-150}	62252.370	1.20×10^{-141}
JC+G4	-30889.237	61874.474	2.61×10^{-150}	61874.717	2.69×10^{-150}	62252.392	1.19×10^{-141}
JC+I+G4	-30888.921	61875.843	1.32×10^{-150}	61876.096	1.35×10^{-150}	62261.634	1.17×10^{-143}

Table S4. Summary of *FAM72* divergence dating using optimized relaxed clock model in BEAST.

	Mean	Standard error	Standard deviation	Median	95% HPD lower	95% HPD upper	ESS
Posterior	-30607.40	0.56	12.30	-30607.20	-30631.30	-30582.60	479.47
Likelihood	-30562.00	0.11	4.26	-30561.70	-30570.60	-30554.20	1637.67
Prior	-45.42	0.55	11.72	-45.07	-68.40	-21.78	449.21
Tree Likelihood	-30562.00	0.11	4.26	-30561.70	-30570.60	-30554.20	1637.67
Tree.height	13.18	0.02	1.71	13.22	9.76	16.39	8778.71
Tree.treeLength	48.32	0.21	7.49	48.08	33.93	62.68	1243.47
kappa	5.18	0.02	0.51	5.14	4.20	6.18	851.52
FreqParameter.1	0.29	0.00	0.00	0.29	0.28	0.30	451.80
FreqParameter.2	0.22	0.00	0.00	0.22	0.21	0.23	407.16
FreqParameter.3	0.21	0.00	0.00	0.21	0.20	0.22	316.96
FreqParameter.4	0.28	0.00	0.00	0.28	0.27	0.29	412.75
ProportionInvariant	0.74	0.00	0.06	0.75	0.62	0.84	340.59
MutationRate	1.00	0.00	0.00	1.00	1.00	1.00	NaN
CalibratedYuleModel	-7.45	0.40	7.26	-7.72	-21.63	6.45	326.74
BirthRateY	2.10	0.04	0.87	1.93	0.73	3.83	417.91
logP(mrca(Homininae))	-1.96	0.01	0.63	-1.72	-3.23	-1.51	9001.00
MRCA.Age(Homininae), Myr	13.18	0.02	1.71	13.22	9.76	16.39	8778.71
logP(mrca(Hominini))	-2.34	0.00	0.00	-2.34	-2.34	-2.34	1.25
MRCA.Age(Hominini), Myr	10.27	0.07	2.17	10.33	6.06	14.30	1116.24
Monophyletic(*Homo sapiens*)	0.98	0.00	0.16	1.00	1.00	1.00	9001.00
MRCA.Age(Homo sapiens), Myr	3.16	0.06	1.08	3.02	1.25	5.36	371.35
Monophyletic(*Pan troglodytes*)	0.99	0.00	0.09	1.00	1.00	1.00	9001.00
MRCA.Age(Pan troglodytes), Myr	1.38	0.03	0.58	1.29	0.44	2.57	521.87
Monophyletic(FAM72B-D)	0.96	0.00	0.20	1.00	1.00	1.00	8973.03
MRCA.Age(FAM72B-D), Myr	2.38	0.04	0.85	2.27	0.89	4.12	394.85
Monophyletic(FAM72C-D)	0.96	0.00	0.19	1.00	1.00	1.00	9001.00
MRCA.Age(FAM72C-D), Myr	1.57	0.03	0.63	1.48	0.51	2.83	454.87
Monophyletic(FAM72A)	0.98	0.00	0.13	1.00	1.00	1.00	8897.93
MRCA.Age(FAM72A), Myr	0.42	0.01	0.21	0.38	0.10	0.83	536.77
Monophyletic(FAM72B)	0.98	0.00	0.12	1.00	1.00	1.00	9001.00
MRCA.Age(FAM72B), Myr	0.48	0.01	0.22	0.44	0.15	0.93	582.29
Monophyletic(FAM72C)	0.98	0.00	0.12	1.00	1.00	1.00	8656.62
MRCA.Age(FAM72C), Myr	0.17	0.00	0.12	0.14	0.01	0.38	1190.35

Monophyletic(FAM72D)	0.97	0.00	0.16	1.00	1.00	1.00	9001.00
MRCA.Age(FAM72D), Myr	0.20	0.00	0.12	0.17	0.03	0.42	828.01
ORCucldMean	0.00068	0.00001	0.00022	0.00064	0.00031	0.00111	273.69
ORCsigma	0.541	0.005	0.136	0.535	0.282	0.801	883.74
ORCRatesStat.mean	0.00	0.00	0.00	0.00	0.00	0.00	1367.00
ORCRatesStat.variance	0.0006	0.0000	0.0001	0.0006	0.0004	0.0008	402.42
ORCRatesStat.coefficientOfVariation	0.56	0.01	0.16	0.54	0.28	0.87	510.45

ESS — Effective Sample Size

HPD — Highest Probability Density

Table S5. Genomic locations with extreme $\beta^{(1)}$ scores found in specific populations.

Gene	Genomic position	Population	Superpopulation	$\beta^{(1)}$	Percentile of standardized $\beta^{(1)}$
FAM72C	chr1:143966782	ASW	AFR	3.735	100
	chr1:143966823			3.735	100
	chr1:143967050			3.757	100
	chr1:143967670			2.818	99
	chr1:143967050	ESN	AFR	2.642	99
	chr1:143967102			2.642	99
	chr1:143966370	GWD	AFR	4.115	100
	chr1:143966825			3.804	99
	chr1:143966840			3.804	99
	chr1:143966853			3.804	99
	chr1:143967050			3.583	99
	chr1:143967110			3.601	99
	chr1:143967315			3.592	99
	chr1:143967490			3.738	99
	chr1:143967788			3.738	99
	chr1:143966823	IBS	EUR	6.947	100
	chr1:143966825			7.92	100
	chr1:143966840			7.92	100
	chr1:143967202			6.803	100
	chr1:143967289			5.116	100
	chr1:143967295			5.116	100
	chr1:143967410			5.116	100
	chr1:143960403	LWK	AFR	4.05	100
	chr1:143960740			4.101	100
	chr1:143960797			5.11	100
	chr1:143966370			3.659	99
	chr1:143966853			3.522	99
	chr1:143967050			3.453	99
	chr1:143967063			3.468	99
	chr1:143967315			3.436	99
	chr1:143967788			3.589	99
	chr1:143966370	MSL	AFR	3.9	100
	chr1:143967050			3.783	99
	chr1:143967110			3.783	99
	chr1:143967315			3.812	99
	chr1:143967670			3.288	99
	chr1:143966370	PUR	AMR	4.067	99
	chr1:143966823			3.797	99

	chr1:143966825			3.797	99
	chr1:143966840			3.797	99
	chr1:143966853			3.797	99
	chr1:143966857			3.797	99
	chr1:143967016			3.774	99
	chr1:143967202			3.648	99
	chr1:143967788			4.124	99
	chr1:143967050	YRI	AFR	3.615	99
	chr1:143967099			3.852	100
	chr1:143967110			3.845	99
	chr1:143967152			3.897	100
	chr1:143967155			3.897	100
	chr1:143967161			3.897	100
	chr1:143967178			3.897	100
	chr1:143967202			3.897	100
	chr1:143967205			3.897	100
	chr1:143967315			3.561	99
	chr1:143967495			4.052	100
	chr1:143967651			4.052	100
	chr1:143967670			3.295	99
	chr1:143967788			3.933	100
FAM72D	chr1:145097288	GWD	AFR	2.7	99
	chr1:145096747	MXL	AMR	2.769	99
	chr1:145097083	YRI	AFR	2.75	99