ESCI upf. School of International Studies

Bachelor's Degree in Bioinformatics (UPF-UPC-UB-UAB)
Final Degree Project

Quantifying positive selection on the human X chromosome and its impact on autism

ARIADNA SAEZ [1] **AND KASPER MUNCH** [2]

[1] *School of International Studies, Pompeu Fabra University, Pg. de Pujades 1, 08003 Barcelona, ES*
[2] *Bioinformatics Research Centre, Aarhus University, Universitetsbyen 81, Building 1872, 8000 Aarhus, DK*

ABSTRACT

Past studies have primarily focused on identifying autism genes in the 22 autosomal chromosomes while neglecting the more complicated analysis of the X chromosome. However, recent research on autism and brain anatomy suggests that positive selection may play a role in shaping genetic variation on this chromosome due to the overlap of genes involved in both spermatogenesis and brain development. This overlap creates potential for positive selection on genes for their role in spermatogenesis to (possibly negatively) affect their function in brain development.
By analyzing patterns of genetic diversity, we intend to identify positively selected genes on the human X chromosome across several populations to answer: (1) if positively selected genes are enriched for ASD genes and (2) if positive selection is more common in genes that are active in both spermatids and brain compared to those that are only active in brain (possibly due to meiotic drive).
These positively selected ASD-related genes expressed in both tissues might contribute to the risk or development of autism and neurodevelopmental disorders and by investigating them, we aim to illuminate the evolutionary forces driving this genetic variation.

Supplementary information: *Supplementary material is available through the following link*

1. INTRODUCTION

Autism [1], also known as autism spectrum disorder (ASD), is a neurodevelopmental condition of variable severity with lifelong effects that can be recognized from early childhood. It is chiefly characterized by difficulties with social interaction and communication and restricted or repetitive thought and behavior patterns. According to the World Health Organization (WHO), around 1% of the world's population has autism spectrum disorder – over 78,000,000 people [2]. In the US, one in every 36 children suffers from ASD. Specifically, in Denmark, the proportion of the Danish population diagnosed with autism is rising quickly and is expected to reach at least 3% soon. Known to be caused by genetic factors, autism is known to run in families, indicating a heritable aspect of the disorder. Environmental factors might also be involved [3]. However, the precise genetic mechanisms involved in ASD development are complex and not fully understood.
A recent study showed that the X chromosome explains 20% of neuro-anatomical variation relevant to ASD, although it represents only 5% of the genome [4]. Even though findings highlight the X chromosome's role in certain aspects of autism-related brain folding [5], efforts to link genes to autism rarely consider X-linked genes [6]. This is due to the different numbers of X chromosomes in males and females [7], complicating such genetic studies. The X chromosome's distinct inheritance pattern is what

makes it significant in ASD cases. Typically, females have two X genes (XX) while males have one X gene and one Y gene (XY). When a male has a genetic variation (SNP) linked to ASD on his X chromosome, he does not experience the compensatory impact of having another X chromosome, as females do. This lack of redundancy in males can magnify the effects of X-linked genetic variations [8], potentially enhancing their impact on ASD.

Positive selection plays a crucial role in shaping the genetic landscape of the X chromosome [9]. The ability to quantify positive selection allows the identification of genes and variants that have undergone adaptive changes over time, leading to an increase in allele frequency. Positive selection refers to the evolutionary process by which advantageous traits or alleles become more common in a population due to their contribution to fitness or survival. It may result from two categories of variants. First, those that increase male fitness by improving spermatogenesis, thus enhancing reproductive success. Second, those that do not directly increase male fitness but promote their own transmission. Unlike autosomes, the X chromosome faces the evolutionary pressure of meiotic drive [10], a phenomenon where certain genetic elements, often termed selfish genes, promote their natural selection by killing off Y-bearing sperm cells. These selfish genes exploit the asymmetry of sex chromosome inheritance, potentially leading to an increase in their own transmission to offspring despite detrimental effects on other tissues and the overrepresentation of certain genetic variants associated with ASD on the X chromosome. Thus, studying spermatogenesis is particularly relevant since genes expressed in sperm cells are more likely to be under positive selection compared to other tissues due to their crucial role in reproduction and the potential influence of meiotic drive.

Understanding the mechanisms of positive selection requires examining the role and generation of genetic diversity, mostly influenced by recombination. Recombination [11], the process by which genetic material is exchanged between homologous chromosomes during meiosis, plays a crucial role in shuffling genetic diversity within a population. This diversity is essential for positive selection to act upon, as it provides the raw material for advantageous traits to emerge and spread through a population. Investigating the dynamics of recombination on the X chromosome provides insights into how positive selection shapes haplotype structures. Haplotypes are combinations of alleles at multiple loci on a single chromosome that are inherited together. Particularly on the X chromosome, where males possess only one copy, transmission of genetic risk for neurodevelopmental disorders can be highly affected.

Instances of recurrent bursts of strong natural selection on the human and primate X chromosome indicate that a selected variant and its environment on the genetic basis fly to high frequency, most recently at 45-55,000 Before Present (BP); this led to strong selection at fourteen different loci and in effect changed allele frequencies for 10% of the chromosome [12]. The scale of this phenomenon is exclusive to the X chromosome, which strongly suggests that such selection acts on gene function in male spermatogenesis; the process by which male germ cells undergo mitosis and meiosis to produce mature sperm cells in the testes. A proteomic comparison between the brain and testis in a recent study has indicated a relationship between spermatogenesis and neurodevelopment. From the total of 14315 and 15687 proteins constituting the human brain and testis proteome in the whole genome, respectively, 13442 are common to both tissues. This corresponds to more than 80% of genes expressed in both the brain and testis [13].

This overlap between brain genes also active in male sperm cells produces the same proteins, although the proteins play different roles in the two tissues. These genes may be frequently and strongly promoted by natural selection for their roles in spermatogenesis, even if the same gene variants are mildly deleterious to brain function and development. In other words, the effect of positive selection can result in a higher prevalence of certain genetic variants that are advantageous for reproduction but may also contribute to the risk of ASD.

The broader significance of the X chromosome in autism is highlighted by the fact that there is more selection on the X chromosome, most likely on spermatid genes, and we think meiotic drive is responsible. Meiotic drive can be very strong and mask deleterious effects on other tissues.

2. OBJECTIVES

This project aims to identify positively selected genes on the X chromosome, (H_1) determine if positively selected genes are enriched for ASD genes, and (H_2) determine whether positive selection is more common in genes active in both spermatids and brain compared to those only active in brain.

We speculate there is a link between genes involved in spermatogenesis and autism, as the altered brain structures observed in ASD might be influenced by the same genetic variants that affect spermatogenesis. The dynamic changes in the brain, driven by neural migration, particularly related to how neural connections and pathways develop and function, may be influenced by genes that also play critical roles in the development and function of sperm cells, suggesting that certain genetic traits linked to ASD may confer advantages in terms of reproductive success, potentially contributing to the persistence of ASD-related genetic variants in populations.

Overall, the main goal is to understand how positive selection, genetic variation, and molecular mechanisms can influence ASD risk and development, focusing on genes located on the X chromosome and their potential roles in spermatid and brain development.

3. METHODS

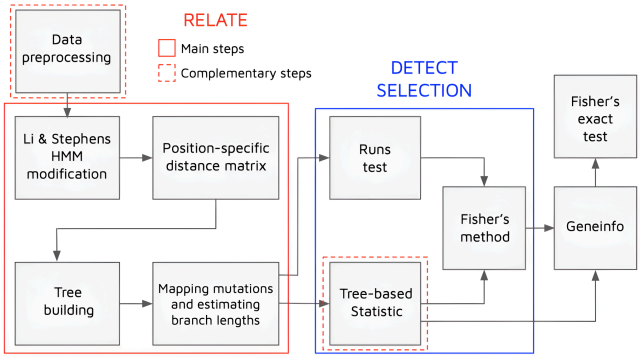

Figure 1: Project pipeline. Steps followed to identify genes under positive selection in the X chromosome and perform GO Enrichment Analysis.

3.1 . Data collection and preprocessing

In this study, we used the ancestral sequence of the human genome as it existed in the common ancestor of all humans. This dataset encompasses a representation of the genomic sequence that theoretically reflects the genetic material of the common ancestor of the entire human population. The common ancestor is a hypothetical individual from the distant past, and this data, which contains ancestral allele information, is sourced from the GRCh38 genome assembly from Ensembl [14].

Furthermore, we also acquired Variant Call Format (VCF) files specifically for phased data from the 1000 Genomes Project [15], a groundbreaking international initiative that aimed to catalog human genetic variations. These files encompass SNP information, with allelic data organized by haplotype, for individuals represented within the 1000 Genomes dataset. This phasing process involves determining the parental origin of alleles, offering a more comprehensive insight into genetic variations among individuals. Sequence index files associated with high-coverage data were also retrieved to easily navigate and align sequence data for downstream analysis.

While the dataset contains files for 26 populations, our analysis is centered on subsets, including 5 African, 4 European, and 5 East Asian populations. Moreover, we ultimately opted to incorporate the Puerto Rican population into our analysis due to its recognized high degree of admixture [16], [17]. The result of admixture is a population with a diverse genetic makeup, reflecting contributions from multiple ancestral sources.

Once all data was gathered, we proceeded to remove non-biallelic SNPs and flip haplotypes according to the ancestral genome, ensuring that the ancestral allele is consistently denoted by 0. Following this step, we filtered SNPs, adjusted distances between them using a genomic mask, and generated SNP annotations required for certain add-on modules.

Exclusively X chromosome data was utilized for the analysis.

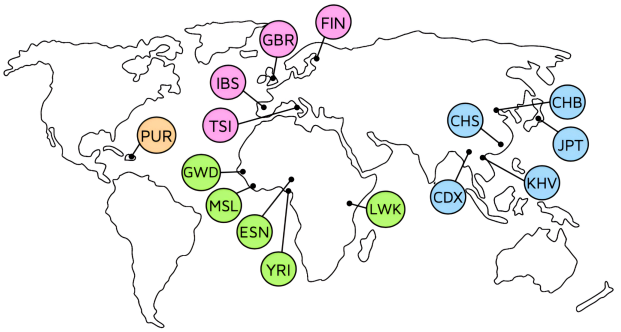

Figure 2: Data collection distribution. Displayed in yellow is PUR: Puerto Rican in Puerto Rico. In pink are FIN: Finnish in Finland; GBR: British from England and Scotland; IBS: Iberian populations in Spain; TSI: Toscani in Italia. In green are GWD: Gambian in Western Division - Mandinka; MSL: Mende in Sierra Leone; ESN: Esan in Nigeria; YRI: Yoruba in Ibadan, Nigeria; LWK: Luhya in Webuye, Kenya. In blue are CDX: Chinese Dai in Xishuangbanna, China; CHB: Han Chinese in Beijing, China; CHS: Han Chinese South, China; JPT: Japanese in Tokyo, Japan; KHV: Kinh in Ho Chi Minh City, Vietnam.

3.2. Relate

The inference of positive selection is the most challenging and time-consuming aspect of this project, warranting a detailed explanation below.

Relate [18] is a scalable software that estimates genome-wide genealogies for thousands of samples, producing adaptable tree structures that account for changes in local ancestry caused by recombination. Relate firstly, identifies a genealogical tree at each site in the genome and describes ancestral relationships among sequences but not their coalescence times. Secondly, these times are estimated after mutations are mapped to branches of these trees, allowing for variable population sizes simultaneously inferred from the data, to produce complete genealogies. [19] Once the ancestral reference genome is aligned with the SNPs data set, Relate follows mainly 4 steps to identify trees.

3.2.1. Li and Stephens HMM Modification

Li and Stephens HMM [20] is a statistical model used to infer haplotype structures and reconstruct the genetic distance among samples in a genomic region. Operating within a hidden Markov modeling framework, the model utilizes a hidden state representation for unobserved haplotypes and observes genetic markers, such as SNPs, across a genomic region. By considering the linkage disequilibrium patterns observed in populations, the Li and Stephens HMM models recombination rates (r) and mutation probabilities (p), providing valuable insights into haplotype structures. It can pinpoint recombination hotspots (regions of the genome with elevated recombination rates) and capture genetic diversity within populations.

In this software, a modification of Li and Stephens HMM [21] is applied. This modification takes into account ancestral and derived states at each SNP and by using the human ancestral sequence, we were able to determine

those states. Ancestral states represent the original genetic characteristics whereas derived states signify altered or mutated genetic features in the context of evolutionary changes. Thereby, the model is modified to determine the probability that one sample copied its genetic information from another sample in the vicinity of the focal SNP, meaning they are closer in terms of evolution. In other words, a higher posterior probability of a reference sequence implies a smaller evolutionary distance to that sequence.

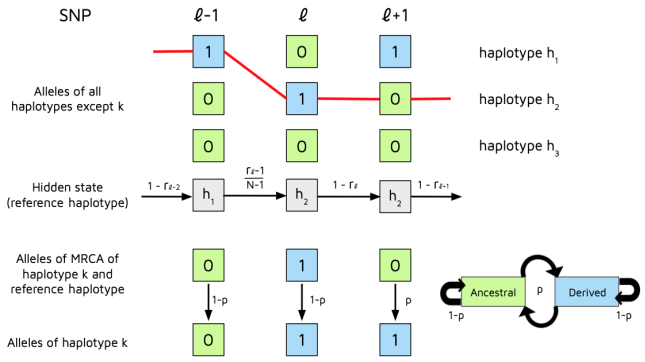

Figure 3: Li and Stephens HMM modified version overview. The figure above shows a schematic example of the modified Li and Stephens HMM applied to haplotype k, which has alleles 0, 1, 1 at loci ℓ-1, ℓ, ℓ+1. The emission (bottom right) and transition probabilities (on top of arrows between hidden states) shown correspond to the path indicated by the red solid line.

3.2.2. Position-specific distance matrix

Posterior probabilities, determined by HMM based on observed data, serve as key inputs in constructing a distance matrix within the Relate framework. This matrix estimates genetic differences between pairs of samples by evaluating the relative order of coalescence events at each genomic position.

While the trees generated vary along the genome, the method heavily relies on nearby SNPs to reconstruct the tree at a specific position. The HMM intuitively reconstructs haplotypes as mosaics [22] of other sample haplotypes, accommodating variations in the copying process and depicting changes in haplotypes as recombination events. After applying the HMM, at a given SNP position ℓ, each haplotype has a probability of being copied from another, producing a position-specific distance matrix.

Appropriately rescaled, this matrix provides insights into the count of mutations unique to one haplotype and those shared from a common ancestor in another. In the absence of recombinations, entries in this matrix converge to the limit of mutations. With recombination, the matrix stores a local count of derived mutations, showcasing that closely related haplotypes tend to have fewer mismatches over extended regions, resulting in smaller distances in the matrix.

3.2.3. Tree building

A hierarchical clustering algorithm is used to build the binary tree from the distance matrix, similar to the Unweighted Pair Group Method with Arithmetic Mean (UPGMA). Data points are grouped in tree-like structures using hierarchical clustering, with leaves representing individual points and internal nodes representing clusters of similar data. UPGMA assumes that all branches have a constant rate of evolution, but in hierarchical clustering algorithms, it may vary in their measurement of cluster dissimilarity. Linkage criteria such as complete, single, or average linkage are used to determine how clusters will merge.

In this particular algorithm, the sequence of haplotype co-occurrence is the driving force behind hierarchical clustering. Notably, it deviates from strict average-linkage criteria seen in UPGMA in several ways. The selection of pairs of clades for merging is not solely based on the order of coalescence but also considers a symmetrized score in the distance matrix, introducing a more sophisticated pairing criterion. After merging, the distance matrix is updated using a weighted sum influenced by clade sizes, a departure from UPGMA's typical unweighted averaging. Additionally, the algorithm is explicitly designed for robustness, accommodating potential inconsistencies, complex recombination histories, and model assumption violations. The algorithm further relaxes conditions for identifying clade pairs to coalesce, acknowledging potential errors or complexities in the data, and providing a more adaptive approach for handling diverse datasets. This dynamic strategy ensures the construction of a binary tree that effectively captures the underlying evolutionary relationships within the dataset.

3.2.4. Mapping mutations to branches and estimating branch lengths

Once the tree structure is set up, mutations match with the branch that has shared descendants with those carrying the derived allele. To handle errors better, a general rule helps with this matching. However, some mutations, such as repeats or error-prone spots, may not neatly fit into just one branch. For these cases, we find the smallest group of branches needed to get back all the data. If a mutation fits into the tree only when we see the derived allele as ancestral (and the other way around), we switch the alleles at that spot.

To optimize computational efficiency and avoid frequent tree construction, trees initiate from the 5' end of a region or chromosome. They progress sequentially, generating a new tree when a SNP is flipped or can not be uniquely assigned to a branch. Ultimately, equivalent branches in neighboring trees prompt the application of the Markov Chain Monte Carlo (MCMC) algorithm to gauge branch lengths. This algorithm assumes a coalescent prior and a single panmictic population [23].

3.3. Detecting positive selection

3.3.1. Tree-based statistic

In addition to the main functions of Relate, we incorporated an additional module to perform the last crucial step to infer selection. This module focuses on investigating positive selection by examining the frequency of each SNP and the number of lineages observed in the corresponding phylogenetic tree over time. The rationale behind this approach lies in the expectation that positive selection leads to the rapid spread and fixation of advantageous mutations, resulting in an increased number of descendant lineages. P-values were computed by evaluating the likelihood of obtaining a result from a given sample, by determining the expected number of ancestral and derived events at the end based on initial conditions.

Under the null hypothesis, we assumed that there is no positive selection acting on the SNPs, and the observed patterns are consistent with neutral evolution. Conversely, the alternative hypothesis posits that certain SNPs exhibit evidence of positive selection, characterized by significant deviations in SNP frequencies and lineage numbers from what would be expected under neutral evolution.

According to GWAS, the commonly accepted genome-wide significance threshold in whole-genome SNP-based studies is 5×10^{-8} [24], [25]. Derived from Bonferroni correction [26], it adjusts for the number of tests to control the family-wise error rate. Since focusing specifically on the X chromosome (approximately $\frac{1}{20}$ of the whole genome), the number of independent tests is reduced compared to the entire genome. Therefore, we used a less stringent threshold by effectively increasing it by a factor of 20:

$$Threshold = 5 \times 10^{-8} \times 20 = 10^{-6} \quad (1)$$

For convenience and interpretability, we converted p-values to $-log_{10}(p - value)$, setting our significance threshold at 6. Thus, all SNPs above this threshold will be considered under positive selection.

3.3.2. Runs test

Beyond computing p-values using Relate's own test for positive selection, we aimed to calculate additional statistics. We also wanted to consider the order in which coalescent events occurred. For each SNP, we combined information related to genealogical relationships and mutation rates (obtained using Relate) to obtain tree sequences in tskit format [27]. From there, we performed a runs test in which the number of ancestral and derived events happening in a row were taken into account. The runs test [28] (similar to the Wald-Wolfowitz test [29] but for shorter sequences) is primarily used to check the randomness of a sequence. It is an agnostic approach as it considers the sequence and frequency of these events without prior assumptions.

If we look at a sequence of binary numbers (being 0 ancestral and 1 derived event), we may pose the question of how many "runs" (sub-sequences consisting of all ones or all zeroes) normally occur in a random sequence of a certain number of zeroes and ones. To decide whether the number of found runs is improbable, one first has to determine the distribution of runs as a function of the counts of zeroes and ones (n_1 and n_2). This distribution may be derived from combinatorial calculations. From these considerations, we could infer that the probability $p(r)$ of observing r runs, given n_1 and n_2, can be calculated as follows:

If r is odd:

$$p(r) = \frac{\binom{n_1-1}{k}\binom{n_2-1}{k-1}+\binom{n_1-1}{k-1}\binom{n_2-1}{k}}{\binom{n_1+n-2}{n_1}} \quad (2)$$

If r is even:

$$p(r) = \frac{2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{\binom{n_1+n_2}{n_1}} \quad (3)$$

Where $k = \frac{r-1}{2}$ and $k = \frac{r}{2}$, respectively. k is an intermediary variable used to simplify the calculation of run probabilities by breaking down the total number of runs into manageable combinatorial components.

A small number of runs indicates insufficient mixing, while a high number of runs points to systematic ordering effects in the sequence. If the number of actually detected runs differs significantly from the expected number of runs, it raises the suspicion that the sequence is not random. The presence of extended runs predominantly supports the scenario of positive selection as the most plausible explanation since negative selection prohibits variants from ever reaching high frequency.

3.3.3. Fisher's method

To obtain accurate results and perhaps identify new positively selected SNPs that Relate may not have detected, we decided to employ Fisher's combined probability test [30] using both p-values obtained from the Tree-based statistic (Relate) and Runs test. Fisher's method is a technique for data fusion or "meta-analysis" used to combine the results from several independent tests bearing upon the same overall hypothesis (H_0). It combines extreme value probabilities (p-values) from each test into one test statistic (X^2) using the formula:

$$X_{2k}^2 = -2\sum_{i=1}^{k} \ln p_i, \quad (4)$$

where p_i is the p-value for the i^{th} hypothesis test. When the p-values tend to be small, the test statistic X^2 will be large, which suggests that the null hypotheses are not

true for every test. When all the null hypotheses are true, and the p_i (or their corresponding test statistics) are independent, X^2 has a chi-squared distribution [31] with $2k$ degrees of freedom, where k is the number of tests being combined. This fact can be used to determine the p-value for X^2. By combining p-values, we enhance our analysis' robustness and comprehensiveness. This integration of diverse information sources increases the power and reliability of our findings, providing a statistically sound approach to detecting significant signals of positive selection.

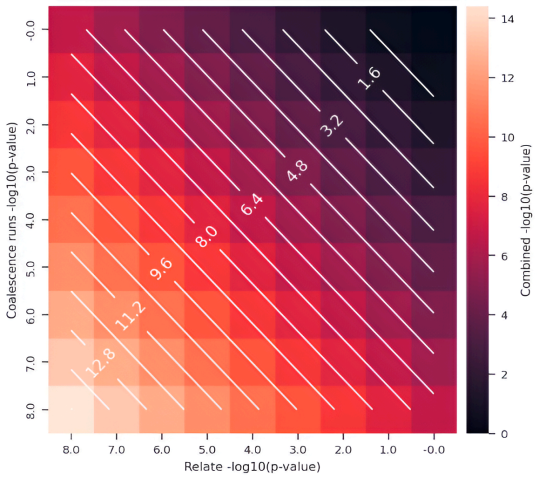

Figure 4: Two combined p-values under Fisher's method. For convenience and interpretability, $-log_{10}(p-value)$ were used in this project. High $-log_{10}(p-value)$ (lightest boundary) means strong evidence against the null hypothesis.

3.4. Geneinfo

Geneinfo [32] is a package designed for Jupyter notebooks, offering utility functions for working with gene information, GO ontologies, and gene networks. Using Geneinfo, we mapped the genomic locations of significant SNPs to nearby genes (using a window size of 1000bp), thereby identifying the positively selected candidate genes. Then, we obtained genes' function and genomic position, and related diseases. All this information is directly retrieved from GeneCards [33] by Geneinfo.

3.5. Fisher's exact test

A crucial statistical technique used to determine whether there are nonrandom correlations between two category variables is the Fisher's Exact Test [34], [35]. In contrast to the chi-squared test, Fisher's Exact Test is especially helpful in situations with small sample sizes since it determines the precise probability of detecting the data under the null hypothesis without the need for large-sample approximations. We used Fisher's exact test to assess whether the number of obtained genes in both understudy categories (ASD-related and expressed in both sperm cells and brain) undergoing positive selection deviates from what would be expected by chance.
The standard threshold of $0,05$ was used [36].

	ASD	**NON-ASD**	Row total
SEL	a	b	a + b
NO SEL	c	d	c + d
Column total	a + c	b +d	a + b + c + d (=n)

Table 1: Contingency table used for applying Fisher's exact test. The table shows the number of genes under selection and not under selection about ASD-related genes and non-ASD-related genes. The values in the cells represent the count of genes in each category (a, b, c and d) and n represents the total.

Thus, the probability p of observing a specific set of values (a, b, c, d) is given by the following formula:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \quad (5)$$

We classified the found genes into the categories using an ASD-related gene list (Sfari Database, 2024 [37], [38]), a brain gene list (Human Protein Atlas, 2024 [39]), and a list of genes expressed in spermatogenesis (Meritxell Riera, personal communication, 2024).

3.6. Code and data availability

Data preprocessing and running of Relate were conducted within a GWF [40] workflow run in the GenomeDK cluster [41]. For further downstream analysis, Python scripts and Jupyter Notebooks were carried out. These resources, links to data, and all generated plots and figures (utilizing the R language and the ggplot2 package [42]) are available online at the following link: https://github.com/munch-group/ariadna-intern.git

4. RESULTS AND DISCUSSION

4.1. Detecting positive selection

After executing Relate, evidence of positive selection across the X chromosome was visualized for each population, including a threshold line to identify the significant variations [Supplementary Figure 1]. Upon observing exclusively SNPs above the threshold [Figure 5a], we identified 111, 42, 414, and 27 SNPs under positive selection in Africans, Europeans, PUR, and East Asians, respectively.
The variation in the number of positively selected SNPs suggests that different factors could have shaped the genetic makeup of these populations. Relate mistaking admixture for selection could explain the high number of positively selected SNPs in PUR. Africans have more than double the number of positively selected SNPs compared to Europeans and almost four times that of East Asians. African populations typically have higher genetic diversity due to a longer evolutionary history [43]. This diversity can provide more raw material for selection to act upon. Historical events such as bottlenecks and migrations [44] might have reduced Europeans' genetic diversity, influencing the number of positively selected SNPs. East

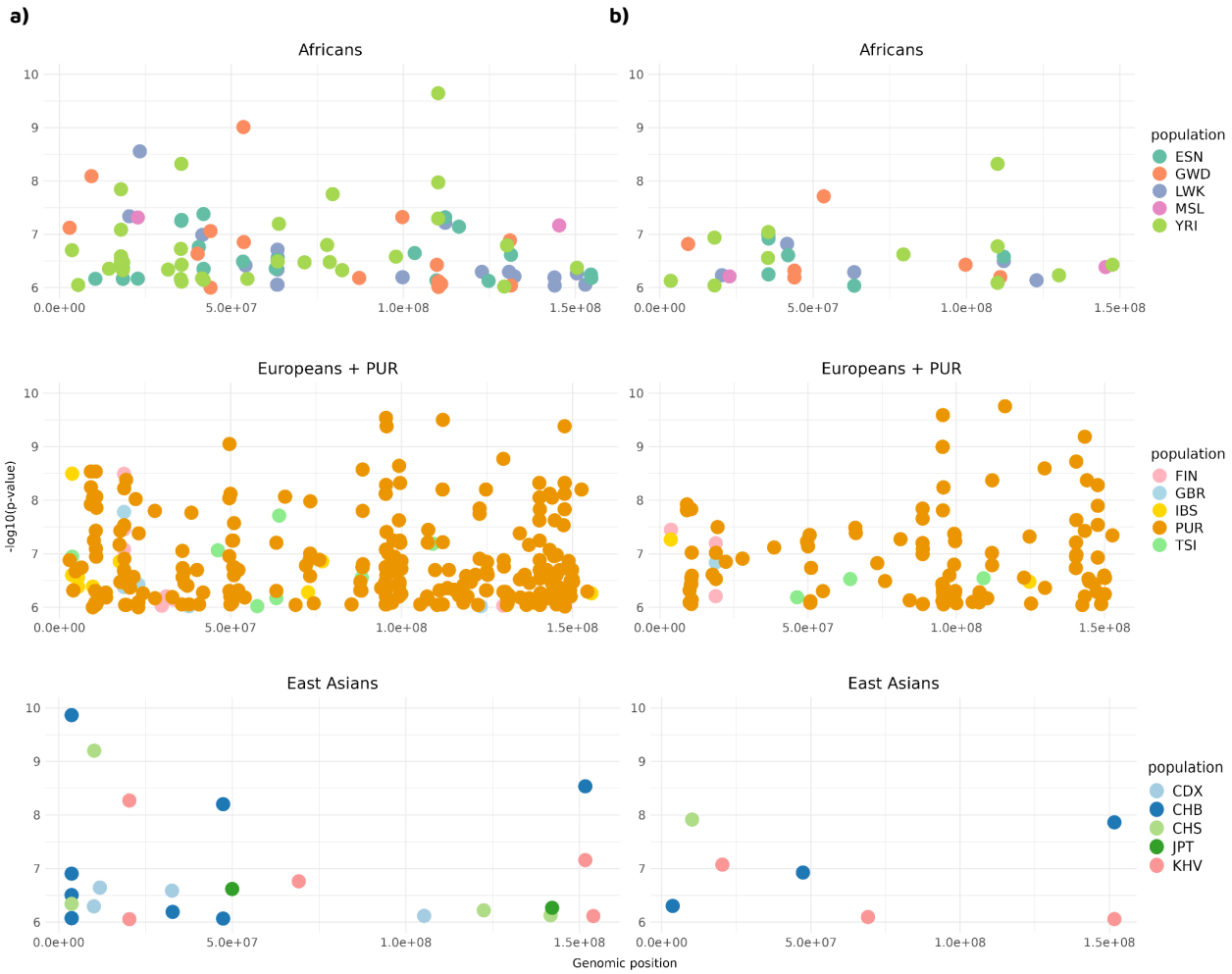

Figure 5: Evidence of positive selection in populations. Comparison of SNPs above threshold (6) along chromosome X in Africa, Europe, Puerto Rico, and East Asia populations obtained by: **a)** Tree-based statistic (from Relate) and **b)** Fisher's method. Each SNP, corresponding to a dot, is plotted when a mutation reaches a frequency of 2, indicating that it has persisted in the population for at least two copies. The higher the $-log_{10}(p - value)$ the stronger the evidence against the null hypothesis. PUR was included in the European plot due to its largest component of genetic ancestry being European, particularly from Spain, due to the colonial history and settlement patterns.

Asian populations may have experienced more stable environmental conditions over time, leading to fewer new adaptive pressures. After conducting the runs test, evidence for positive selection was visualized [Supplementary Figure 2] following the same process but did not obtain any SNP above the threshold. This lack of significant SNPs is not unexpected due to the low power of the runs test. The tree-based statistic likely has higher power and sensitivity to detect positive selection since it is tailored to look for specific patterns expected under positive selection. In contrast, the runs test is a general test for randomness and might not capture the specific signal of selection as effectively.

Since the runs test did not show sufficient power, we performed Fisher's method combining evidence from the Relate and runs test to get more robust and reliable results. Evidence of positive selection across the X chromosome was visualized for each population, including a threshold line to identify the significant variations [Supplementary Figure 3]. Upon observing exclusively SNPs above the threshold [Figure 5b], we identified 32, 13, 165, and 7 SNPs under positive selection in Africans, Europeans, PUR, and East Asians, respectively. As expected, we identified fewer SNPs compared to Relate alone because the p-values from the runs test were not sufficiently low, and when they were low, they often did not coincide with low p-values from the Relate analysis. This reduced the overall power of the combined analysis.

4.2. Identifying candidate genes

We identified genes overlapping significant SNPs in both Relate [Table 2] and Fisher's method [Supplementary Table 1] results above threshold by mapping positively selected SNPs to nearby genes using Geneinfo. After removing unannotated, microRNA and antisense genes, we found 35 and 9 positively selected and well-annotated candidate genes in Relate and Fisher's method, respectively. Genes exclusively found in PUR were omitted to avoid false signals of positive selection since

GENE	START POS	END POS	BRAIN	SPERM	ASD	POPULATION
ACSL4	109641334	109733257	YES	YES	NO	GWD
AMMECR1	110194185	110318085	YES	YES	NO	GWD, PUR
ARSL	2934520	2964341	NO	NO	NO	GWD, PUR
BCOR	40051245	40177277	YES	YES	NO	GWD
CASK	41514933	41665852	YES	YES	YES	ESN, LWK, YRI
CDKL5	18425607	18640196	YES	YES	YES	ESN, YRI
CLCN4	10156974	10237660	YES	YES	YES	CHS
CLCN5	49922595	50042541	YES	YES	NO	JPT, PUR
DMD	31119221	32155469	YES	YES	YES	CDX, CHB, FIN, PUR, YRI
ENOX2	130622324	130720491	YES	YES	NO	LWK, PUR
FAM120C	54068323	54183254	YES	YES	NO	LWK, PUR
FRMPD4	11822438	12724523	YES	YES	YES	CDX
G6PD	154531389	154547018	YES	YES	NO	ESN
GNL3L	54530218	54645854	YES	YES	NO	YRI
HUWE1	53532095	53680089	YES	YES	YES	GWD
IGSF1	131273505	131289306	YES	NO	NO	ESN, GWD
IL1RAPL1	28587445	29956718	YES	YES	YES	CDX, FIN
MAGT1	77825746	77895435	YES	YES	NO	YRI
MAMLD1	150361571	150514173	YES	YES	NO	LWK, YRI
NHS	17375199	17735994	YES	YES	NO	IBS, PUR
NYX	41447342	41475652	YES	NO	NO	YRI
PAK3	110944396	111217494	YES	YES	NO	GWD, PUR
PASD1	151563674	151676739	NO	NO	NO	CHB, KHV
PHKA1	72578813	72714306	YES	YES	NO	IBS
PRKX	3604339	3713649	YES	YES	NO	CHS, CHB, IBS, TSI, YRI
PTCHD1	23334395	23404374	YES	YES	YES	LWK
RAB33A	130110622	130184870	YES	YES	NO	YRI
RTL4	112083012	112457514	NO	NO	NO	ESN, LWK, PUR
TENM1	124375902	124963817	YES	YES	NO	ESN, IBS, PUR
TMEM164	110002368	110177788	YES	YES	NO	YRI
TMLHE	155489010	155536365	YES	YES	YES	IBS
WWC3	10015253	10144474	YES	YES	NO	CDX, PUR
XPNPEP2	129738978	129769536	NO	NO	NO	FIN, PUR
ZMYM3	71239623	71253721	YES	YES	NO	YRI
ZNF185	152898066	152973474	YES	YES	NO	LWK

Table 2: Genes under positive selection. This table provides information about the 35 candidate genes found in chromosome X across populations by Geneinfo given p-values obtained from Tree-based statistic (in Relate workflow). The 4th and 5th columns report whether the gene is expressed in the brain and whether it is expressed in both sperm cells, respectively. The 6th column reports whether the gene is associated with autism. The last column states in which populations these genes are found to be under positive selection.

very strong admixture violates the assumptions of the model Relate is based on. So the significant findings may be spurious. Admixture can mimic positive selection signals, leading to a mistaken interpretation of these patterns as evidence of selection rather than the result of admixture. The 9 candidate genes in Fisher's method list (CASK, CLCN4, PAK3, PASD1, PRKX, RAB33A, RTL4, TENM1, and TMEM164) were also present in Relate's list. Genes found by Relate might include both strong and more subtle signals, suggesting Relate might be more sensitive, while those identified only by Fisher's method are fewer candidates, but these may be more robust. These genes might represent stronger candidates for further study, as they show significance even when considering the null results from the runs test. Nevertheless, we decided to proceed with the 35 candidate genes identified by Relate to ensure more statistical power for further downstream analysis. Some genes to highlight [Table 2] are DMD and PRKX, which are shared across various populations reflecting historical

movements and genetic mixing, leading to the spread of adaptive traits. DMD and PRKX may carry adaptive traits that confer advantages in many environments or provide resistance to particular diseases. Furthermore, genes like ACSL4 and BCOR (only in GWD) might be under selection due to specific local environmental pressures in West Africa. Out of these 35 positively selected genes, 31 genes are expressed in the brain, and from these, 29 are expressed in both brain and sperm cells. 9 positively selected genes are associated with autism. From these ASD-related genes, CASK, HUWE1, FRMPD4, IL1RAPL1, and CLCN4 are associated with Intellectual Developmental Disorder and Syndromic and Non-Syndromic X-Linked Intellectual Disability. Mutations in PTCHD and TMLHE are associated directly with Autism X-Linked. Finally, CDKL5 is associated with Epileptic Encephalopathy 2, and DMD is associated with Muscular Dystrophy. All ASD-related genes under positive selection are expressed in both brain and sperm cells. This dual expression could imply that these genes are involved in critical pathways or mechanisms that are essential for both brain development/function and reproductive processes. We could also highlight the fact that CDKL5, HUWE1, FRMPD4 and, PTCHD1 are Differentially Expressed Genes (DEGs) in spermatogenesis. DEGs are candidate meiotic drivers because differentially expressed genes may exert a differential effect on X and Y-bearing sperm cells.

4.3. GO enrichment analysis

a)

	ASD	**NON-ASD**	Row total
SEL	9	26	35
NO SEL	68	1309	1377
Column total	77	1323	1412

b)

	SPERM	**NON-SPERM**	Row total
SEL	29	2	31
NO SEL	371	160	531
Column total	400	162	562

Table 3: Contingency tables used for applying Fisher's exact tests. Tables to test **a)** Hypothesis 1 (H_1) and **b)** Hypothesis 2 (H_2). Totals are **a)** the number of genes in the X chromosome (1412) and **b)** the number of brain genes in the X chromosome (562). P-values obtained were **a)** $5,95 \times 10^{-5}$ and **b)** $0,002$.

According to the standard threshold, both H_1 (Positively selected genes are enriched for ASD) and H_2 (Positive selection is more common in genes that are active in both spermatids and the brain compared to those that are only active in the brain) are accepted since the p-values obtained in Fisher Exact's tests are lower than $0,05$. The low p-value obtained in H_1 ($5,95 \times 10^{-5}$) [Table 3a] strongly indicates that positively selected genes are enriched for ASD, meaning that among the genes that have undergone positive selection, there is a higher proportion of genes associated with ASD compared to what would be expected by chance. Viewed in isolation, this might suggest that characteristics associated with ASD have an evolutionary benefit. This advantage may arise from enhanced sensory awareness or special cognitive skills, which may have been useful in ancestral environments. Secondly, it implies that cognitive traits common to people with ASD, including exceptional memory or focused attention, could have been beneficial for survival and reproduction in earlier circumstances, leading to their retention and spread in human populations. While ASD is often viewed in a negative light due to the challenges it presents, the presence of positively selected genes suggests that there may be a balance between the disadvantages and advantages conferred by these genetic variants. This balancing selection could imply that the benefits of the associated traits outweigh the challenges in certain contexts or environments.

However, the acceptance of H_2 [Table 3b] supports the idea of strong positive selection acting on genes for their role in spermatogenesis rather than brain development. Due to this dual function, there may be evolutionary trade-offs where the role of the gene in brain development may be influenced and even constrained by selection for favorable features in reproduction or selection on transmission distortion by meiotic drivers on X. Although this dual selection pressure may have complex impacts on brain development and function, it may also help explain why certain genetic variants that are advantageous in reproductive situations continue to exist. Although this is only indirect evidence, it bolsters the meiotic drive theory, suggesting that genes involved in spermatogenesis are particularly prone to positive selection due to their crucial role in ensuring successful reproduction. The evolutionary advantage conferred by these genes in spermatogenesis may inadvertently affect their neuronal roles, highlighting an interesting aspect of genetic regulation where reproductive success influences neurodevelopmental outcomes. Overall, genes involved in both reproduction and brain function are crucial and, their expression in both cell types does translate to stronger positive selection compared to brain-specific genes.

5. CONCLUSIONS

In this study, we showed the importance of considering the X chromosome and the interplay between spermatogenesis and brain development in understanding the genetic basis of ASD. When detecting positive selection we could see a wide variation in the number of SNPs found under selection across demographic groups. The high number of identified SNPs under selection could be explained by Relate mistaking admixture for selection in PUR and high genetic diversity and evolutionary history

in Africans. Conversely, the low numbers of selected SNPs in Europeans and East Asians could be explained by reductions in genetic diversity and more stable environments. This analysis revealed population-specific genetic adaptations as well as possible historical movements and genetic mixing, leading to the spread of adaptive traits.

Our final tests indicated that both hypotheses were true and thus, positive selection enriches ASD-related genes and is more common in genes that are active in both spermatids and the brain compared to those that are only active in the brain. This overlap might result in genes being subject to positive selection due to their role in spermatogenesis, which could have consequences for their function in the brain and contribute to the development of ASD. The altered brain structures observed in ASD could be influenced by the same genetic variants that affect spermatogenesis, highlighting a potential evolutionary trade-off.

The findings provide a foundation for more research into how positive selection and genetic variation on the X chromosome influence ASD risk and development. Further work to understand genes' roles in reproduction and brain function could include investigating in detail the Puerto Ricans, by being able to differentiate real positive selection signals from admixture. Furthermore, we could increase the number of samples by using all data from the 26 populations in the 1000 Genome Project and perhaps investigate further into primates. By adding more samples, we could lower the threshold and have more statistical power and robustness to detect positively selected genes and gain insights into the genetic basis of ASD.

6. ABBREVIATIONS

SNP: Single Nucleotide Polymorphism; ASD: Autism Spectrum Disorder; BP: Before Present; WHO: World Health Organization; VCF: Variant Call Format; PUR: Puerto Rican in Puerto Rico; FIN: Finnish in Finland; GBR: British from England and Scotland; IBS: Iberian populations in Spain; TSI: Toscani in Italia; GWD: Gambian in Western Division - Mandinka; MSL: Mende in Sierra Leone; ESN: Esan in Nigeria; YRI: Yoruba in Ibadan, Nigeria; LWK: Luhya in Webuye, Kenya; CDX: Chinese Dai in Xishuangbanna, China; CHB: Han Chinese in Beijing, China; CHS: Han Chinese South, China; JPT: Japanese in Tokyo, Japan; KHV: Kinh in Ho Chi Minh City, Vietnam; HMM: Hidden Markov Model; MCMC: Markov Chain Monte Carlo; UPGMA: Unweighted Pair Group Method with Arithmetic Mean; GO: Gene Ontology; DEGs: Differentially Expressed Genes

7. ACKNOWLEDGMENTS

Chiefly, I would like to express my sincere gratitude to my research supervisor, Kasper Munch, for his outstanding guidance, support, and patience. I would also like to extend my thanks to the entire BiRC team for providing me the opportunity to work and learn alongside such talented individuals.

8. REFERENCES

[1] H. Hodges, C. Fealko, and N. Soares, "Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation," *Transl. Pediatr.*, vol. 9, no. Suppl 1, pp. S55–S65, Feb. 2020, doi: 10.21037/tp.2019.09.09.

[2] "Autism." Accessed: Mar. 14, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders

[3] P. Chaste and M. Leboyer, "Autism risk factors: genes, environment, and gene-environment interactions," *Dialogues Clin. Neurosci.*, vol. 14, no. 3, pp. 281–292, Sep. 2012.

[4] T. T. Mallard *et al.*, "X-chromosome influences on neuroanatomical variation in humans," *Nat. Neurosci.*, vol. 24, no. 9, pp. 1216–1224, Sep. 2021, doi: 10.1038/s41593-021-00890-w.

[5] T. A. Nguyen, A. W. Lehr, and K. W. Roche, "Neuroligins and Neurodevelopmental Disorders: X-Linked Genetics," *Front. Synaptic Neurosci.*, vol. 12, p. 33, Aug. 2020, doi: 10.3389/fnsyn.2020.00033.

[6] E. J. Marco and D. H. Skuse, "Autism-lessons from the X chromosome," *Soc. Cogn. Affect. Neurosci.*, vol. 1, no. 3, pp. 183–193, Dec. 2006, doi: 10.1093/scan/nsl028.

[7] A. P. Arnold *et al.*, "The importance of having two X chromosomes," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 371, no. 1688, p. 20150113, Feb. 2016, doi: 10.1098/rstb.2015.0113.

[8] A. M. Casto, J. Z. Li, D. Absher, R. Myers, S. Ramachandran, and M. W. Feldman, "Characterization of X-Linked SNP genotypic variation in globally distributed human populations," *Genome Biol.*, vol. 11, no. 1, p. R10, 2010, doi: 10.1186/gb-2010-11-1-r10.

[9] K. R. Veeramah, R. N. Gutenkunst, A. E. Woerner, J. C. Watkins, and M. F. Hammer, "Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans," *Mol. Biol. Evol.*, vol. 31, no. 9, pp. 2267–2282, Sep. 2014, doi: 10.1093/molbev/msu166.

[10] J. B. Searle and F. P.-M. de Villena, "The evolutionary significance of meiotic drive," *Heredity*, vol. 129, no. 1, pp. 44–47, Jul. 2022, doi: 10.1038/s41437-022-00534-0.

[11] J. Stapley, P. G. D. Feulner, S. E. Johnston, A. W. Santure, and C. M. Smadja, "Recombination: the good, the bad and the variable," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 372, no. 1736, p. 20170279, Dec. 2017, doi: 10.1098/rstb.2017.0279.

[12] L. Skov *et al.*, "Extraordinary selection on the human X chromosome associated with archaic admixture," *Cell Genomics*, vol. 3, no. 3, p. 100274, Mar. 2023, doi: 10.1016/j.xgen.2023.100274.

[13] B. Matos, S. J. Publicover, L. F. C. Castro, P. J. Esteves, and M. Fardilha, "Brain and testis: more alike than previously thought?," *Open Biol.*, vol. 11, no. 6, p. 200322, doi: 10.1098/rsob.200322.

[14] "Index of /pub/release-109/fasta/ancestral_alleles." Accessed: Mar. 07, 2024. [Online]. Available: http://ftp.ensembl.org/pub/release-109/fasta/ancestral_alleles/
[15] "1000 Genomes | A Deep Catalog of Human Genetic Variation." Accessed: Mar. 07, 2024. [Online]. Available: https://www.internationalgenome.org/
[16] M. Via *et al.*, "History Shaped the Geographic Distribution of Genomic Admixture on the Island of Puerto Rico," *PLoS ONE*, vol. 6, no. 1, p. e16513, Jan. 2011, doi: 10.1371/journal.pone.0016513.
[17] H. Tang *et al.*, "Recent Genetic Selection in the Ancestral Admixture of Puerto Ricans," *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 626–633, Sep. 2007.
[18] "Relate Documentation." Accessed: Mar. 07, 2024. [Online]. Available: https://myersgroup.github.io/relate/
[19] L. Speidel, "A method for genome-wide genealogy estimation for thousands of samples," *Nat. Genet.*, vol. 51, 2019.
[20] "Haplotype matching in large cohorts using the Li and Stephens model | Bioinformatics | Oxford Academic." Accessed: May 21, 2024. [Online]. Available: https://academic.oup.com/bioinformatics/article/35/5/798/5079326
[21] H. Li and R. Durbin, "Inference of human population history from individual whole-genome sequences," *Nature*, vol. 475, no. 7357, pp. 493–496, Jul. 2011, doi: 10.1038/nature10231.
[22] A. Sanaullah, D. Zhi, and S. Zhang, "Minimal positional substring cover is a haplotype threading alternative to Li and Stephens model," *Genome Res.*, vol. 33, no. 7, pp. 1007–1014, Jul. 2023, doi: 10.1101/gr.277673.123.
[23] J. F. C. Kingman, "On the genealogy of large populations," *J. Appl. Probab.*, vol. 19, no. A, pp. 27–43, Jan. 1982, doi: 10.2307/3213548.
[24] J. Fadista, A. K. Manning, J. C. Florez, and L. Groop, "The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants," *Eur. J. Hum. Genet.*, vol. 24, no. 8, pp. 1202–1205, Aug. 2016, doi: 10.1038/ejhg.2015.269.
[25] "4.5.2: GWAS," Biology LibreTexts. Accessed: May 25, 2024. [Online]. Available: https://bio.libretexts.org/Courses/Clinton_College/BIO_300%3A_Introduction_to_Genetics_(Neely)/04%3A_Inheritance/4.05%3A__Linkage/4.5.02%3A_GWAS
[26] A. S. Kaler and L. C. Purcell, "Estimation of a significance threshold for genome-wide association studies," *BMC Genomics*, vol. 20, no. 1, p. 618, Jul. 2019, doi: 10.1186/s12864-019-5992-7.
[27] "Introduction — Tskit manual." Accessed: May 21, 2024. [Online]. Available: https://tskit.dev/tskit/docs/stable/introduction.html
[28] "Runs Test." Accessed: May 21, 2024. [Online]. Available: http://www.statistics4u.com/fundstat_eng/ee_runs_test.html
[29] "Wald–Wolfowitz runs test," *Wikipedia*. Apr. 06, 2024. Accessed: Jun. 16, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Wald%E2%80%93Wolfowitz_runs_test&oldid=1217467835
[30] "Fisher's method," *Wikipedia*. Feb. 28, 2024. Accessed: May 21, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Fisher%27s_method&oldid=1210712377
[31] "Chi-squared distribution," *Wikipedia*. Apr. 12, 2024. Accessed: May 21, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Chi-squared_distribution&oldid=1218540953
[32] "munch-group/geneinfo." munch-group, Apr. 12, 2024. Accessed: May 21, 2024. [Online]. Available: https://github.com/munch-group/geneinfo
[33] "GeneCards - Human Genes | Gene Database | Gene Search." Accessed: Jun. 08, 2024. [Online]. Available: https://www.genecards.org/
[34] "Fisher's exact test," *Wikipedia*. Feb. 28, 2024. Accessed: May 23, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Fisher%27s_exact_test&oldid=1210712147
[35] R. A. Fisher, "On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P," Jan. 1922, doi: 10.2307/2340521.
[36] J. Sauro and J. R. Lewis, "Chapter 5 - Is There a Statistical Difference between Designs?," in *Quantifying the User Experience*, J. Sauro and J. R. Lewis, Eds., Boston: Morgan Kaufmann, 2012, pp. 63–103. doi: 10.1016/B978-0-12-384968-7.00005-9.
[37] "Human Gene Module," SFARI Gene. Accessed: Jun. 17, 2024. [Online]. Available: https://gene.sfari.org/database/human-gene/
[38] E. Leitão *et al.*, "Systematic analysis and prediction of genes associated with monogenic disorders on human chromosome X," *Nat. Commun.*, vol. 13, no. 1, p. 6570, Nov. 2022, doi: 10.1038/s41467-022-34264-y.
[39] "Search: NOT tissue_category_rna:brain;Not detected - The Human Protein Atlas." Accessed: Jun. 17, 2024. [Online]. Available: https://www.proteinatlas.org/search/NOT+tissue_category_rna%3Abrain%3BNot+detected
[40] "gwf 2.0.5 documentation." Accessed: Mar. 07, 2024. [Online]. Available: https://gwf.app/
[41] "GenomeDK." Accessed: Mar. 07, 2024. [Online]. Available: https://genome.au.dk/
[42] H. Wickham, "Data Analysis," in *ggplot2: Elegant Graphics for Data Analysis*, H. Wickham, Ed., Cham: Springer International Publishing, 2016, pp. 189–201. doi: 10.1007/978-3-319-24277-4_9.
[43] F. Gomez, J. Hirbo, and S. A. Tishkoff, "Genetic Variation and Adaptation in Africa: Implications for Human Evolution and Disease," *Cold Spring Harb. Perspect. Biol.*, vol. 6, no. 7, p. a008524, Jul. 2014, doi: 10.1101/cshperspect.a008524.
[44] V. Margari *et al.*, "Extreme glacial cooling likely led to hominin depopulation of Europe in the Early Pleistocene," *Science*, vol. 381, no. 6658, pp. 693–699, Aug. 2023, doi: 10.1126/science.adf4445.