Evolutionary and comparative genomics of X-linked genes associated with Autism and Tourette syndrome

Carolina Pontes Cunha
Master in Forensic Genetics
Department of Biology
Faculty of Sciences of the University of Porto
2023/2024

Evolutionary and comparative genomics of X-linked genes associated with Autism and Tourette syndrome

Carolina Pontes Cunha

Dissertation carried out as part of the Master in Forensic Genetics

Department of Biology

2023/2024

Supervisor

Prof. Dr. Agostinho Antunes, FCUP/CIIMAR

Acknowledgements

This dissertation marks the culmination of a few years of hard work, dedication, persistence, and it would not be possible without the unwavering support and contribution of a group of people, who, in one way or another, helped and marked this journey.

To professor Agostinho Antunes, my supervisor, I would like to thank for the constant availability to help, and all the guidance given to me.

To Luana Ramos, who was an essential part of my journey, I would like to thank for the constant availability at any hour, for all the support, orientation and dedication throughout this project, but also for all emotional support and kindness that she gave me. I will always be grateful.

A thank you to all my family, especially my parents for always being my safe harbour, without their support and help I would not be able to do this.

To all my friends for always being there ready to help, but specifically to Sara, who was one of the most important parts in my journey, thank you for always being there, thank you for being my home away from home for two years, and for continuing being available and ready to help every time.

Finally, a thank you to everyone that, in a way or another, helped and accompanied me in this journey, I will never forget.

Resumo

É importante estudar os cromossomas sexuais pois têm um papel fundamental nas forças evolutivas que atuam no genoma, uma vez que têm um padrão de transmissão único. O cromossoma X em específico tem características muito interessantes. Neste cromossoma estão localizados inúmeros genes relacionados com o desenvolvimento cerebral, incluindo vários que foram ao longo dos anos associados com patologias hereditárias humanas, nomeadamente a síndrome de Tourette e o autismo. O objetivo deste trabalho é explorar a evolução de alguns destes genes, identificando sites que possam ter evoluído sob seleção positiva. A identificação destes sites é relevante para a genética médica pois podem estar relacionados com mutações que possam ser algumas das causas da síndrome de Tourette e/ou autismo. Desta forma, escolhemos sete genes localizados nas regiões pseudoautossomais (PAR) do cromossoma X que estão descritos na literatura como relacionados com estes síndromes: *AKAP17A, NLGN3, NLGN4X, VAMP7, SPRY3, IL9R, ASMTL.*

Utilizando ferramentas bioinformáticas para analise de seleção positiva, os genes *NLGN3* e *IL9R* apresentaram um (NLGN3) ou dois (IL9R) sites com vestígios de seleção positiva no ramo do Homem. Apesar de não haver estudos detalhados sobre a função especifica destes sites, a informação disponível aponta para que possam estar envolvidos em vários processos, tais como melanoma cutâneo da pele (no caso do gene *IL9R*) e até mesmo terem efeitos deletérios (em ambos os genes). Além disso, a existência de mutações em sites próximos associados ao autismo sugere que, no gene *NLGN3*, os sites descobertos também possam ter algum tipo de influência no desenvolvimento do autismo. Os resultados são expectáveis tendo em conta o facto da maioria destes genes têm função no cérebro humano, sendo, portanto, altamente conservados. Em estudos futuros, seria interessante investigar a estrutura 3D da proteína entre indivíduos saudáveis e pacientes com mutações, recorrendo a estudos de proteómica computacional. Seria também interessante fazer estudos de genética populacional. O significado evolutivo de determinadas mutações poderá ser de particular interesse para a saúde humana e investigação farmacêutica.

Abstract

Sex chromosomes have a fundamental role in shaping the evolutionary forces that act on the entirety of the genome, since they have an unusual pattern of transmission. The X chromosome, in particular, has very interesting characteristics. This chromosome is enriched in genes that participate in brain development, several of which are associated with human hereditary pathologies such as Tourette syndrome (TD) and autism (ASD). The aim of this study is to explore the evolutionary patterns of some of these genes, identifying sites that might have evolved under positive selection. The identification of these sites is important for human medical investigation, since they can be related with mutations that cause TD and/or Autism. Therefore, we chose seven genes located in the pseudoautosomal region (PAR) of the X chromosome that have some evidence in the literature connecting them to TD and ASD: *AKAP17A, NLGN3, NLGN4X, VAMP7, SPRY3, IL9R, ASMTL.*

Using bioinformatic tools, we preformed positive selection analysis and obtained significant results for *NLGN3* and *IL9R* in the human lineage, in which both showed one (NLGN3) or two (IL9R) sites with evidence of positive selection. Even though the existing of detail studies about the specific function of these sites are scarce, the available information points to the fact that they might be involved in various processes, such as skin cutaneous melanoma (in the *IL9R* gene) and even have deleterious effects (in both genes). The existence of mutations in near sites associated with autism, also suggests that, in the gene *NLGN3*, the discovered sites can also have some type of influence in the development of autism. The results were expected given that most of the genes are extremely conserved due to their important functions in the human brain development. In subsequent studies it would be interesting to explore computational proteomic analysis to further investigate the 3D protein structure between healthy and sick patients with mutations as well as population genetic studies. The evolutionary meaning of certain changes may be of interest to human health and pharmaceutical research.

Table of contents

List of tables

List of figures

List of Abbreviations

DNA- Deoxyribonucleic acid

PAR- pseudo autosomal region

XCI- X chromosome inactivation

ADHD- Attention deficit hyperactivity disorder

OCD- Obsessive compulsive disorder

ASD- Autism spectrum disorder

TD- Tourette syndrome

DLG4- Disc large homolog 4

bp- base pair

kb- kilobase

MAPK- Mitogen-activated protein kinase

NCBI- National Center for Biotechnology information

FTP- File transfer protocol

CDS- Coding sequence

MUSCLE- Multiple sequence comparison by log-Expectation

MEGA- Molecular evolutionary Genetic Analysis

MAFFT- Multiple alignment using fast fourier transform

DAMBE- Data analysis for molecular biology and evolution

SNP- Single nucleotide polymorphism

FN3- Fibronectin type-III domain

COSMIC- Catalogue of Somatic Mutation in Cancer

MSA- Multiple alignment

gCF- Gene concordance factor

sCF- Site concordance factor

bb- bootstraps

1. Introduction

1.1. Definition of a chromosome

A chromosome is a highly organized structure of a cell that contains the genetic material of an organism. Each chromosome contains a molecule of DNA (deoxyribonucleic acid) comprising numerous genes that define most characteristics of the organism. Humans have 23 pairs of chromosomes, one of which is a pair of sex chromosomes. Theother 22 pairs are called autosomes. The sex chromosomes are responsible for the sex determination, and in humans, an XX pair encodes a female (homogametic sex) while XY determines a male (heterogametic sex). The DNA is organized inside the nucleus aggregated to histones (structural proteins) and this complex is called chromatin [1].

In evolutionary genetics, the importance of studying the sex chromosomes lies, mainly, in two reasons. On one hand, their role in the sex determination, and on the other hand, in the fact that they provide a unique approach for the study of the fundamental evolutionary forces that act on the entirety of the genome, due to their unusual pattern of transmission. This means that any biological difference between the sexes can cause the sex chromosomes to experience distinct evolutionary environments [2].

1.2. The X chromosome and the PAR regions.

The X chromosome in particular, has very interesting characteristics. In mammals, females inherit the X chromosome from both parents while males only inherit the maternal X chromosome. The sex chromosomes have evolved throughout the years from a pair of autosomes. During this process some functional elements on the X chromosome have been conserved. However, the Y chromosome lost almost all of its traces of the ancestral autosome, including genes that were shared with the X chromosome. The inactivation of one of the X chromosomes in females (XCI) is a mechanism of dosage control in mammals because many of the genes in the X chromosome do not have homologues in the Y chromosome. Some genes escape this inactivation, mainly genes that have homologues in the Y chromosome. Because of this process, the male X chromosome fails to recombine its entire length during meiosis, recombining only in short regions at the tips of the X chromosome arms that recombine with its equivalent segments in the Y chromosome. These regions are called the pseudo autosomal regions (PAR1 and PAR2 in humans). The genes outside these regions are mostly X-linked [3].

As mentioned before, in females, the inactivation of one of the X chromosomes occurs to obtain gene dosage equilibrium. The PAR region is the only region that escapes XCI in all species and therefore maintains its activity in both chromosomes. Because of this, pair recombination occurs in both females and males [4]. The human PAR regions contain around 29 genes that have diverse roles such as cell signalling, transcriptional regulation and mitochondrial functions. Around 24 of these genes are in the PAR1 regions and half of them have a known function. The genes in both PAR regions are not inherited in a strictly sex-linked fashion but rather an autosomal fashion [5].

The PAR1 is the most prominent region of the PAR. It is located at the end of the short arm of the human sex chromosomes and is approximately 2.7MB long. This region has a very high rate of recombination when compared to the autosomes. Some of the genes in PAR1 are linked to serious diseases like Klinefelter Syndrome, leukaemia and mental disorders. Deletions in PAR1 can result in failure of pairing and male sterility. Some of the genes that are part of the PAR1 region are *AKAP17A, ASMT, ASMTL, CD99, IL3RA* and *SHOX.* In contrast, the PAR2 is approximately 330kb and is located at the end of the long arm of the sex chromosomes. This region, thus far, seems to be exclusive to humans and has likely arisen due to translocations between the X and the Y chromosomes [4]. The behaviour of PAR 2 is similar to the autosomal regions and the crossovers in this region occur at a rate similar to the genome average [6]. The PAR2 region contains 4 protein coding genes such as *DDX11L16, IL9R, SPRY3* and *VAMP7*, 5 pseudogenes and 1 long noncoding antisense RNA gene [5].

1.3. X-linked diseases.

The random inactivation of one of the X chromosomes (XCI) is very important when it comes to relevant phenotypic changes because certain mutations that could lead to very serious complications in males are compensated by another chromosome in many of the female cells. The occurrence of XCI has several implications in the impact of mutations of the X chromosomes as well as allowing the development of multiple X-linked diseases [7].

These diseases can be caused by mutations of certain genes in the X chromosome or X chromosomes anomalies such as monosomies (lacking a full sex chromosome) and trisomy (having an extra chromosome). These anomalies can also occur when someone is missing a part of the chromosome (deletion) or have more than one extra sex chromosome. These sex chromosomes anomalies can cause syndromes with a wide range of physical and cognitive implications. Generally, the

syndromes caused by a sex chromosome anomaly are less severe than the ones caused by an autosome [8].

Some of the diseases associated with the X chromosome are colour blindness, haemophilia, Turner syndrome, Klinefelter syndrome and fragile X chromosome syndrome.

Colour blindness is the inability to see or differentiate colours. The most common form of colour blindness affects the perception of red and green. The total absence of colour vision is a very severe and rare form of colour blindness where it is only possible to see in black and white. The cones are highly specialized nerve cells in the retina that have the ability to colour code objects by capturing the light in multiple wave lengths. These cells are able to recognize and transmit to the brain the colour of the objects. Colour blindness can be acquired; however, it is more common as a hereditary disease. It is cause by recessive genes (*OPN1LW, OPN1MW*, and *OPN1SW*) in the X chromosome and therefore it affects more the male population than the female [9].

Haemophilia is a genetic disease associated with impaired blood clotting caused by the decreased levels in one of the proteins that contributes to the clotting cascade. There are various types of haemophilia. The haemophilia A type is the most common and corresponds to the coagulation factor VIII deficiency. It is caused by mutation in recessive genes (*F8* and *F9*) in the X chromosome and therefore, like the colour blindness, its more common in males than females because a female is haemophilic only if both X chromosomes have that gene mutation [10].

The Turner syndrome, 45X, is a genetic condition that only affects females and is characterized by the total or partial absence of one of the X chromosomes. It is the only viable monosomy in humans. Females with this syndrome present abnormal physical traits like swelling of the hands and feet, redundant creases at the nape of the neck, webbed neck, broad chest, and inverted nipples. They also do not develop secondary sexual characters. They are more at risk of developing some other diseases like congenital heart disease, diabetes, hyperthyroidism and vision and hearing problems [11].

The Klinefelter syndrome, 47XXY/XXX, occurs because of the non-disjunction of the sex chromosomes resulting in individuals with one or more X chromosomes and it is one of the main reasons for male sterility. In females the presence of the extra chromosome does not lead to any significative anomalies and does not affect the woman's fertility [12].

Fragile X chromosome syndrome is a dominant genetic disease caused by a mutation in the gene *FMR1* located in the X chromosome. This mutation causes a deficit in the FMR1 protein that is essential to the development of the connections between neurons. Its symptoms are light to moderate intellectual deficiency [13].

The XCI can have a protective effect in chromosome aneuploidies such as Turner syndrome and Klinefelter syndrome because it balances the chromosome. Some deletions or additions of the X chromosome can be tolerated because the abnormal X chromosome is preferentially inactivated. However, XCI is more often associated with abnormal phenotypes [14].

A correlation between intellectual disability and skewed XCI was also established. Since the males only have one copy of the X chromosome, intellectual disability in the autism spectrum and other mental diseases are more prevalent in males than in females. One example of a disease that causes a wide variation of symptoms in women is haemophilia, ranging from simply being a carrier to have serious implications. As explained earlier, it is more common in males since it is a recessive disease, and the woman would have to carry the gene mutated in both X chromosomes to be affected. Skewed X chromosomes inactivation can create serious consequences, reducing the clotting factor in the blood in the same way the disease affects males [14].

1.4. Tourette syndrome

The Tourette syndrome is a hereditary neuropsychiatric disorder characterised by motor and phonic tics. It is a very complex syndrome with a heterogeneity of symptoms and comorbidities between individuals and therefore it is very hard to diagnose and treat [15]. It is most common in children and the incidence peaks around preadolescence. It is often associated with other disorders like ADHD (attention deficit hyperactivity disorder) and obsessive-compulsive disorder (OCD), as well as autism. Some gene mutations are related to both Tourette's and ADHD, OCD or autism. The tics are involuntary or semi-voluntary, sudden, intermittent and can be repetitive movements (motor) or sounds (phonic). These can be classified into two categories, simple and complex. The simple motor tics involve a single muscle or a group of muscles and can be brief or more prolonged. These tics are often easy to camouflage as voluntary movements and are frequently unnoticed. The complex motor tics produce a coordinated pattern of movements that involve multiple muscle groups [15].

In patients with Tourette's, the tics can be more or less strong over days, weeks or even months and they can gain new tics or regain old ones. In average, the tics start around eight years old, peak in preadolescence and decline in early adulthood [16]. However, the most severe cases of Tourette's appear in adulthood. These severe cases involve self-injurious motor tics (*i.e.*, biting and hitting), and socially unacceptable behaviours like shouting obscenities, racial slurs, and gestures. The frequency of the disorder varies by age and sex since males are four times more likely to be affected than females [17]. Studies have found that genetic and environmental factors likely play a role in causing Tourette syndrome.

The syndrome is believed to be linked to problems in some areas of the brain and chemical substances. However, some people with Tourette have been found to have genetic mutations involving certain genes. The most common is involving *SLITRK1*, but other genes in the X chromosome can be involved in the cause of this syndrome [18].

When it comes to the treatment of this syndrome, there is no cure. Treatment focuses on improving the quality of life, social functioning, and self-esteem. Treatment should be individually tailored and combined use of different medication may be needed as well as behavioural therapy and counselling [16].

1.5. ADHD, OCD and Autism and their relation to Tourette syndrome

Additionally, many patients with Tourette's syndrome can also experience other symptoms leading to other disorders, specifically hyperkinetic disorder such as Attention-deficit/hyperactivity disorder (ADHD) and obsessive-compulsive disorder (OCD) or both [17].Even though ADHD and OCD are the most common, autism spectrum disorder, anxiety, sleep disorder, depression and self-injuries are also associated [15].

Attention-deficit/hyperactivity disorder is defined as a psychiatric disorder that is characterized by a certain set of symptoms such as inattention, hyperactivity and impulsivity [19]. Obsessive compulsive disorder is defined by the presence of recurring, intrusive and distressing thoughts (obsessions) that cause anxiety or emotional stress, often leading to repetitive, stereotypic behaviours or thoughts (compulsions) aimed at neutralizing these negative feelings (American Psychiatric Association 1994). ADHD and OCD are among the most prevalent neuropsychiatric disorders in paediatric populations. Family studies have demonstrated high heritability in both ADHD and OCD, with some

genetic findings revealing shared variants and common pathogenetic mechanisms, while others indicate distinct differences between the two disorders [20] .

Some studies observed the presence of tics in individuals with autism spectrum disorders (ASDs), including chronic motor tics, vocal tics, and Tourette's Disorder (TD). This association likely stems from shared underlying etiological factors, as the co-occurrence rate is higher than what would be expected by chance [21].

Autism spectrum disorder is a group of neurodevelopmental disorders characterized by impairments in social communication, repetitive behaviours, highly restricted interests, and sensory behaviours that begin early in life. Recent advances have shed light on the causes of ASD, revealing a combination of genetic vulnerabilities and environmental risk factors [22].

It is also important to note that Chronic tic disorder is a condition marked by brief, sudden, uncontrollable, spasm movements or vocal outbursts, but not both. When both physical and vocal outburst are present, the condition is identified as Tourette syndrome [23].

While genetic and familial factors likely influence the development of brain pathways and the manifestation of Tourette syndrome, the precise mechanisms behind these interactions remain unclear [15].

As mentioned before, some genes located in the X chromosome can also be one of the causes of this syndrome. For this work we chose seven of those genes to work with. Some of these genes have little to no bibliography and studies relating them to their role in Tourette's, only being mentioned as possibly related to the syndrome. The genes were, therefore, chosen based on their potential implication in Tourette's and are the neuroligin-3 (*NLGN3*), Neuroligin-4-X-linked (*NLGN4X*), Interleukin-9-receptor (*IL9R*), A-Kinase Anchoring Protein 17 (*AKAP17A*), vesicle-associated membrane protein 7 (*VAMP7*), Protein sprouty homolog 3 (*SPRY3*) and Acetylserotonin O-mehtyltranferase like (*ASMTL*).

1.6. Chosen genes for the study

The *NLGN3* gene, Neuroligin-3, is a gene located in the X chromosome, more specifically Xq13.1 and has 10 exons. This gene encodes a member of the neuroligin family of neural cell surface proteins. The neuroligins mediate the formation and maintenance of synapses between neurons. The lack of expression of specific neuroligins could affect neuronal interactions within the synaptic network, leading to

consequent neuropathology [24]. *NLGN3* may act as splice site-specific ligand for beta-neurexins and may be involved in the formation and remodelling of central nervous system synapses. Mutations in this gene can be associated with autism spectrum disorders and cognitive disorders.

The *NLGN4X*, Neuroligin-4 X-linked, is also a member of the neuroligin family and is expressed in the cerebral cortex. It is also located in the X chromosome, more specifically in the Xp22.32-p22.31 and has 32 exons. It has a similar function as the *NLGN3*, so it encodes a protein that is part of the type-B carboxylesterase/lipase family and functions as a neuronal cell surface protein. Proteins in this family may act as splice site- specific ligands for beta-neurexins and could play a role in the formation and remodelling of synapses in the central nervous system. The encoded protein interacts with discs large homolog 4 (DLG4). Mutations in this gene are also found in people that are in the autism spectrum, more particularly with X-linked autism [25]. There are some mutations in this gene that have been described with the correspondent consequences but the most relevant for this study is a 2bp deletion in this gene causing a frame shift and therefore a premature stop codon, which is found in people on the autism spectrum. It has also been noted that a deletion of the exons 4,5 and 6 in this gene can cause a truncated protein and can be one of the causes of Tourette [24].

The *AKAP17A* gene is a gene that encodes the protein A-kinase anchoring protein 17A and it's located in the PAR1 on both sex chromosomes, Xp22.33 and Yp11.2. The protein is a part of the spliceosome complex, and it is involved in the regulation of alternative splicing in some mRNA precursors. Mutations were described in patients with autism and chronic tic disorders [26].

The *IL9R* gene, Interleukin-9 receptor, is a type I cytokine receptor. The protein encoded by this gene is a cytokine receptor that specifically mediates the biological effects of interleukin 9. This gene is located in the pseudoautosomal regions of the X and Y chromosomes, specifically in the PAR2. On the X chromosome its location is Xq28 and in the Y chromosome is Yq1 [27].

The gene *VAMP7*, also known as *SYBL1*, encodes a transmembrane protein that is a member of the soluble N-ethylmaleimide-sensitive factor attachment protein receptor family. This protein is localized in late endosomes and lysosomes and is involved in the fusion of transport vesicles to their target membranes. The gene is located in the Xq28 and Yq12 and it is part of the pseudoautosomal region 2. A reduction of protein levels of

VAMP7 selectively inhibit spontaneous neurotransmission both in cultured neurons and in hippocampal slices. Interestingly, a polymorphism in the regulatory region of the gene has been associated with bipolar affective disorders and mutations in this gene are also found in patients with autism [28].

The *SPRY3* gene is located in Xq28 and Yq12 in the PAR2 and encodes the protein sprouty homolog 3. It is involved in negative regulation of MAPK cascade and mutations in this gene can be found in patients with autism and schizophrenia. It is highly expressed in central and peripheral nervous system ganglion cells in humans, including cerebellar Purkinje cells and retinal ganglion cells. Tourette's syndrome can occur with a mutation that implies a duplication that includes both *VAMP7* and *SPRY3* genes. This mutation encompasses a 260 kb copy gain in the Xq28, located at the terminal region of the long arm of the X chromosome that includes the complete sequence of both *VAMP7* and *SPRY3* genes as well as a portion at the 5' side of the *IL9R* gene in PAR2 [28].

ASMTL gene encodes the Probable Bifunctional DTTP/UTPPyrophosphatase/Methyltransferase Protein (dTTP_UTP_pyrophosphatse) and is located in the PAR1 region Xp22.3. It has a dual function in the stopping of cell division and in the prevention of the incorporation of nucleotides into cellular nucleic acids [29].

The aim of this study is to identify sites on the selected genes of the X chromosome that have evolved under positive selection in the human lineage, which could be related with mutations causing TD and/or autism in humans. This will assist our knowledge to better understand why some forms of Tourette's and autism can be X-related and hereditary. For this, we performed phylogenetic and selection analyses in a set of X-linked genes connected to autism and Tourette syndrome to better understand the evolutionary genetics functional role of the found mutations.

2. Methods

2.1. Genome selection and retrieval

A total of 24 mammalian species of various families and orders were chosen to provide an array of representative species (Table S1). All the genomes were selected and downloaded from the NCBI database [27] using the FTP directory. The assembly choice was based on a few criteria*: i)* most complete genome (higher number of base pairs); *ii)* lower number of scaffolds (which usually provides a less fragmented assembly) ; *iii)* genome assembled at the chromosome level; *iv)* preference for genomes sequenced using long-read technology.

2.2. Query collection

Information regarding the genes chosen for this study was obtained from the Ensembl database [30] and NCBI (Table S2).

Then, we used the Uniport database [31] to retrieve the protein fasta sequence for each individual gene, used then as queary for the extraction of the nucleotide sequences. Additionally, we used Ensembl to verify the existence of different isoforms. This step was done to ensure that we chose one of the longest isoforms, since Uniprot usually shows first the most common one. Then we used GeneCards [32] to check for highly similar paralogs of each gene that might provide similar hits during the extraction of the coding sequences.

This information is compiled in table S3.

2.3. Nucleotide sequences retrieval and dataset preparation

The protein sequences gathered in the previous step were used as query in the protein-2-genome option of the Exonerate software [33] v2.4.0 to obtain the coding sequences (CDS) from the non-annotated genomes and the hit corresponding to the highest raw score (RS) value was extracted. Each hit was extracted from the initial methionine to the final stop codon, whenever it was possible.

When a sequence presented intra-sequence stop codons or was truncated it was noted as a pseudogene and discarded. Information regarding the scaffold/chromosome in which the hits were located and the respective introns (number of introns and size) was collected (Attachment 4 and Table S5). The fasta files corresponding to each hit were then obtained using a custom phyton script.

2.4. Multiple sequence alignments

After obtaining the CDS, an initial screening was performed in MEGA [34] v10.2.2 using preliminary MUSCLE [35] alignments. In this initial alignment we verified the sequences for any additional intra-sequential stop codons and in case of their existence, those sequences were eliminated.

Then, the Guidance 2 webserver [36] was used to align the sequences by codon using the MAFFT [37] alignment algorithm with the remaining options on default. We used the resulting guidance scores for the removal of the columns with low confidence (default threshold 0.93).

Then, MEGA was used to analyse the resulting alignment and performed a manual refinement. Each column that had more that 20% of missing information was eliminated. We also eliminated nucleotides that were not ATGC and took note of the size of the sequences with and without gaps.

2.5. Extraction of annotated sequences

Given its higher divergence among the mammal species, the extraction of *IL9R* gene from non-annotated genomes proved more difficult, resulting in a high number of false pseudogenes. Therefore, this gene was retrieved from annotated sequences available on the NCBI gene database, using Geneious v11.1.5 (https://www.geneious.com). To search the gene in our group of interest we used the queary “IL9R and mammals” and selected the sequences that corresponded to our target species. Each file had a complete gene sequence with CDS and intro annotations (Table S6).

We extracted every annotated CDS for each species and aligned using the translation alignment option in Geneious, in order to choose the most complete CDS.

We then used these nucleotide sequences, and the *Homo sapiens* sequence extracted with Exonerate (to ensure that the *H. sapiens* sequences from all genes were extracted) to create an alignment following the same methodology described above.

2.6. Saturation analysis

The next step is to check the level of saturation for each of the final nucleotide alignments. When substitution saturation occurs, it reduces the phylogenetic information present in the sequences.

To determine the level of saturation we used DAMBE [38] v7.3.32. We estimated the proportion of invariant sites with the most basal species as an outgroup. Then, we measured the substitution saturation using the Xia's test [39].

2.7. Phylogenetic analysis

All the alignments were then concatenated using Geneious, to enhance the phylogenetic signal for the phylogenetic reconstruction.

We performed a Maximum Likelihood inference on the concatenated dataset using IQtree2 [40] v2.0.7 employing 10000 bootstrap replicates (bb) [41] and 10000 Shimodaira–Hasegawa approximate likelihood ratio tests (SH-aLRT) [42]. Each gene was defined as an independent partition and the best-fit evolutionary model for each partition was automatically calculated using ModelFinder on IQtree2. The resulting phylogenetic tree was then used to calculate the gene (gCF) [43] and site (sCF) [43] concordance factors, with 100 quartets for the calculation of sCF. The phylogenetic tree was edited using Figtree [44] v1.4.4.

2.8. Selection Analysis

We used the concatenated tree and marked the *Homo sapiens* lineage as our test subject (foreground branch) to run positive selection analysis. We chose to run the following tests on Hyphy [45] v2.5.36: *i)* RELAX [46]; *ii)* BUSTED [47]; *iii)* aBSREL [48] and iv) contrast-FEL [49].

RELAX is a branch specific test that assess the intensification or relaxation of the natural selection and therefore is not a test to the positive selection. BUSTED is a branch and site test, that evaluates evidence of episodic positive selection. aBRREL is a branch specific tests, indicating which branch (in our case it is only the *Homo sapiens* branch) have evolved under positive selection. Finally, contrast-FEL is a site-specific test that tells us if a specific site has sines of selection.

When the tests provided significant positively selected sites (PSS), we then converted them to match the human protein used as query.

2.9. Protein structure

Finally, we used the Protter webserver [50] v1.0 to create a 2D structure of the proteins that presented significant results and mark the sites that showed evidence of positive selection. This allows us to pinpoint the PSS in the protein structural domain.

3. Results

3.1. Dataset collection

The MSA (multiple sequence alignment) was performed using 88 functional sequences out of the original 175 (having been removed 87 sequences) corresponding to 7 different genes from the X chromosome (*ASMTL, AKAP17A, NLGN4X, NLGN3, VAMP7, SPRY3, IL9R*) of 24 different species of mammals.

Each MSA was tested for nucleotide substitution and the 7 datasets presented low saturation levels, which indicates that they are suitable for phylogenetic inference (Table S7).

3.2. Phylogenetic analysis

The obtained phylogenetic tree, in general, as it can be seen on figure 1, presents well divided basal groups. The support values (bootstraps and SH-aLRT) in these divisions between the Afrotheria, Euarchontoglires and Laurasiatheria have scores above 85% which means that the tree has good support between the most basal groups. Most divisions inside each group are also well supported, with values above 80% and 95% for SH-aLRT and bb, respectively.

Mainly within Laurasiatheria the support levels are lower compared to the rest of the tree, having some values of 64% and 56%. There are some differences between the topology of our resulting tree and the mammalian phylogenomic tree obtained used as a guide [51], for example in the carnivora group, where the *Panthera leo* and *Tremarctos ornatus* species should be placed together but instead are separated in distinct clades.

The branches that lead to *Homo sapiens* are well supported and in conformity with the general topology [51].

When it comes to the gene concordance factor and the site concordance factor, in the *Homo sapiens*, we have a 100% and 52.9% correspondingly, which are good results that support the topology. Generally, most of the branch divisions have gCF values above 50% which means that more than 50% of the individual gene trees support that topology. On the other hand, some cases have lower values of sCF meaning that, in general, some sites within the alignment do not support those divisions.

Even though the number of species with functional nucleotide sequences differs for the 7 genes, all datasets were maintained since there was still a sufficient number of backgrounds species for each gene to perform the subsequent analyses.

However, it means that for some branches, the gCF and the sCF values could not be calculated.

Given that the branch of *Rattus norvegicus* was longer than the outgroup, which indicates an extremely high number of nucleotide substitutions that might have occur due to sequencing or assembly errors, we have removed this species from the selection analysis.

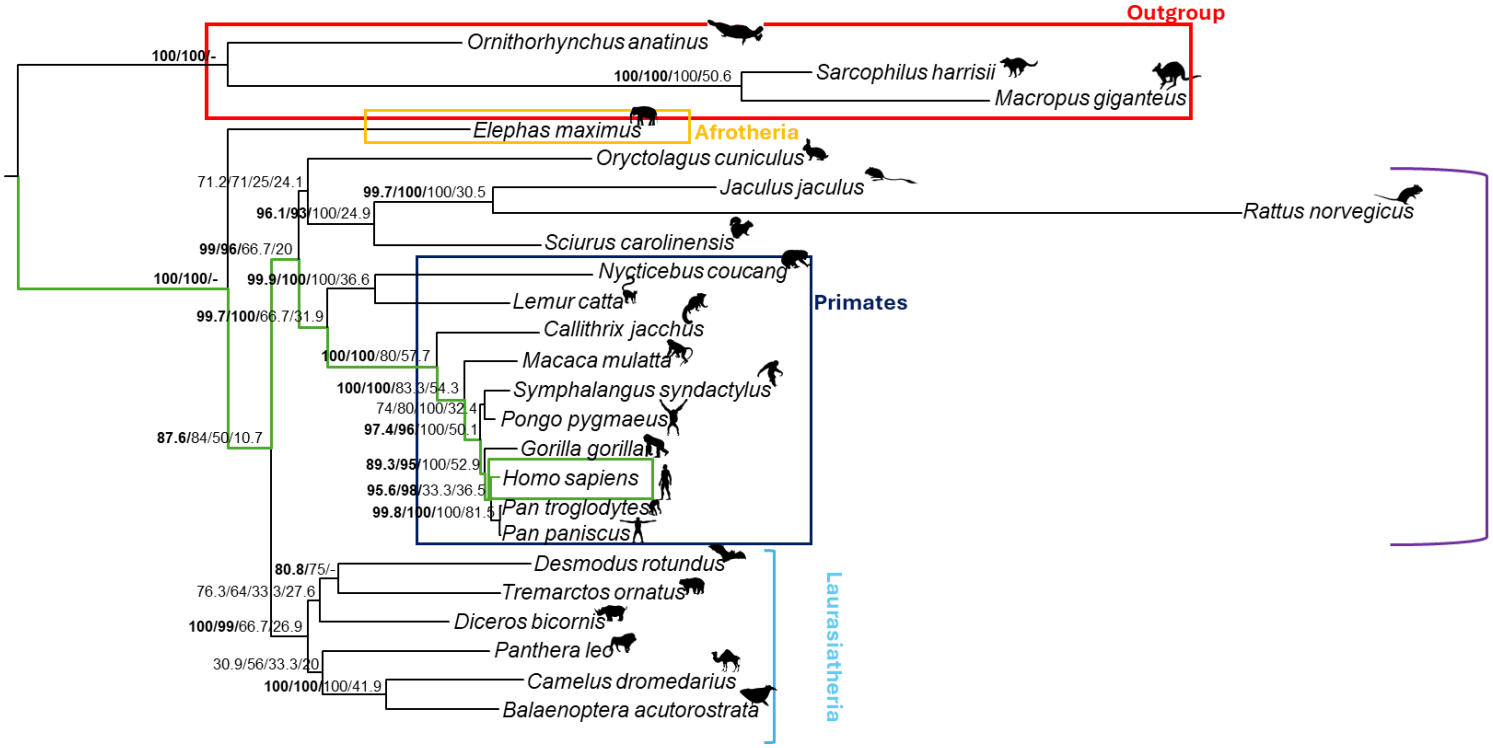


Figure 1. Maximum likelihood phylogenetic tree of the evolution of the genes *AKAP17, NLGN3, NLGN4X, VAMP7, SPRY3, IL9R, ASMTL* in mammals; support values are displayed behind the nodes and show SH-aLRT/bb/gCF/sCF.

3.3. Selection Analysis

3.3.1. Relax test

For this test, only *NLGN3* showed significant signals of relaxation (K<1; *p*< 0,05). For the rest of the genes, the results were not significant (table 1).

Table 1. Results for the selection analysis RELAX using Hyphy for the dataset. k- branch specific relaxation parameter; *p* - *p*-value; LR - likelihood ratio

Gene	k	*p*	LR	Result
AKAP17A	0.53	0.113	2.52	Not significant
ASMTL	7.50	0.926	0.01	Not significant
IL9R	1.00	1.000	0.00	Not significant
NLGN3	**0.00**	**0.024**	**5.13**	**Significant**
NLGN4X	7.09	0.584	0.30	Not significant
SPRY3	1.94	0.610	0.26	Not significant
VAMP7	1.31	0.819	0.05	Not significant

3.3.2. aBSREL test

For the aBSREL test, none of the genes showed significant results (table 2), meaning that there were no signs of episodic selection in any of the genes.

Table 2. Results for the selection analysis aBSREL using Hyphy for the dataset; *p* - *p*-value.

Gene	*p*	Result
AKAP17A	1.00	No selection
ASMTL	1.00	No selection
IL9R	0.432	No selection
NLGN3	0.307	No selection
NLGN4X	1	No selection
SPRY3	1	No selection
VAMP7	1	No selection

3.3.3. BUSTED test

When analysing the BUSTED results, all the 7 genes present not significant results for evidence of positive selection (table 3).

Table 3. Results for the selection analysis BUSTED using Hyphy for the dataset; *p* - *p*-value; Er - Evidence ratio

Gene	p	Er	Result
AKAP17A	0.5000	0 sites	Not significant
ASMTL	0.5000	0 sites	Not significant
IL9R	0.4700	0 sites	Not significant
NLGN3	0.3700	1 site	Not significant
NLGN4X	0.5000	0 sites	Not significant
SPRY3	0.5000	0 sites	Not significant
VAMP7	0.5000	0 sites	Not significant

3.3.4. Contrast-FEL

For this test, the genes *AKAP17A* and *ASMTL* showed positive signs of selection but when the false positive test was performed it showed that there were 0 sites with evidence of positive selection (table 4). Only the genes *IL9R* and *NLGN3* showed sites with signs of positive selection after the false positive test *(q*-value < 0.2) which are showed on table 5.

Table 4. Results for the selection analysis Contrast-FEL using Hyphy for the dataset; *p* - *p*-value; *q* - *q*-value (false positive discovery)

Gene	Sites with $p < 0.05$	Sites with $q < 0.2$
AKAP17A	7	0
ASMTL	6	0
IL9R	**5**	**2**
NLGN3	**1**	**1**
NLGN4X	0	0
SPRY3	0	0
VAMP7	0	0

Table 5. Results of the *q*- value < 0.2 (false discovery) test

Gene	Codon	*q*-value
IL9R	311	0.0416
IL9R	232	0.0650
NLGN3	223	0.0259

3.4. Protter protein structure

Using Protter we obtained the 2D structure of both proteins (*IL9R* and *NLGN3*). Both proteins have an extracellular and a cytoplasmic part. This representation also shows the signal peptide, the disulfide bonds and the variants.

In *IL9R* (figure 2), both residues are located in the extracellular part and on *NLGN3* (figure 3), the residue 223 is also located in the extracellular part.

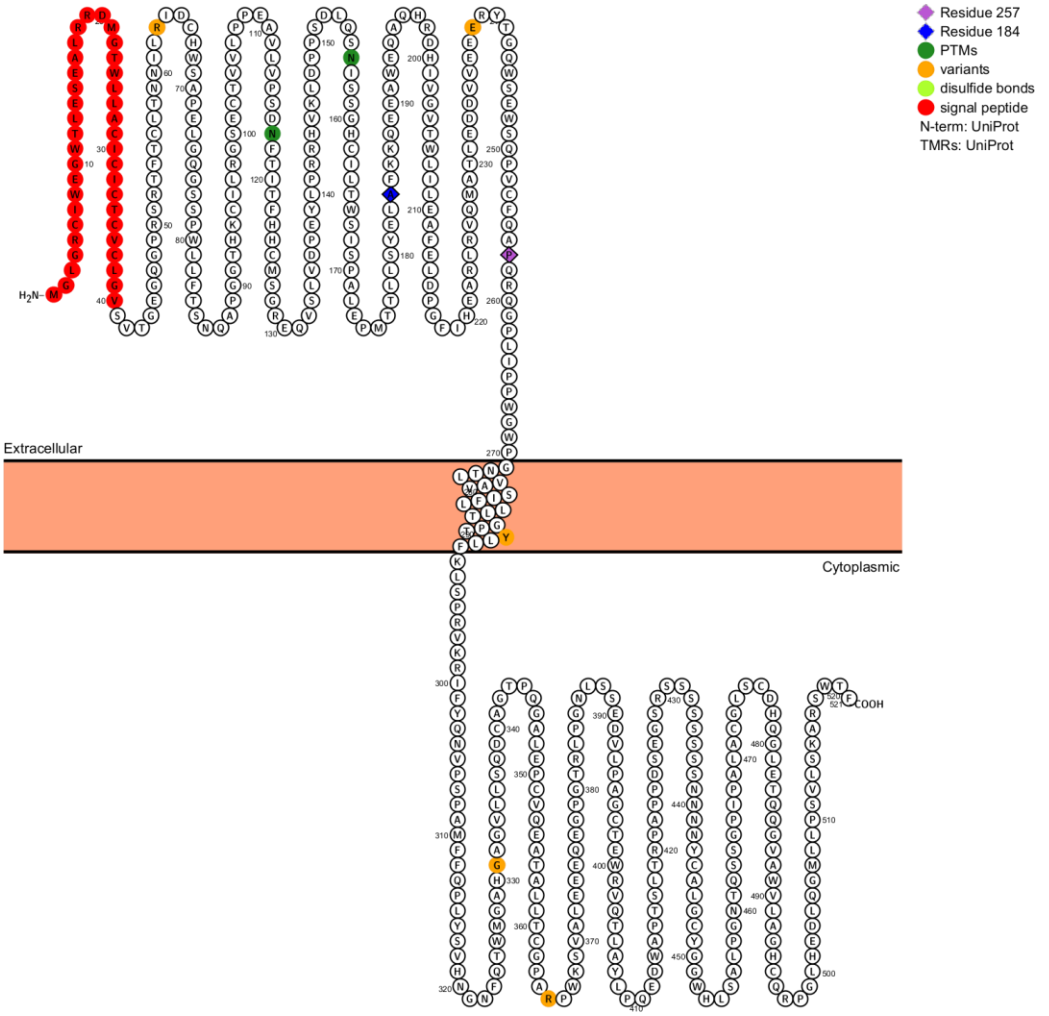


Figure 2. 2D structure of the Interleukin-9 receptor protein with the residues 184 and 257 marked.

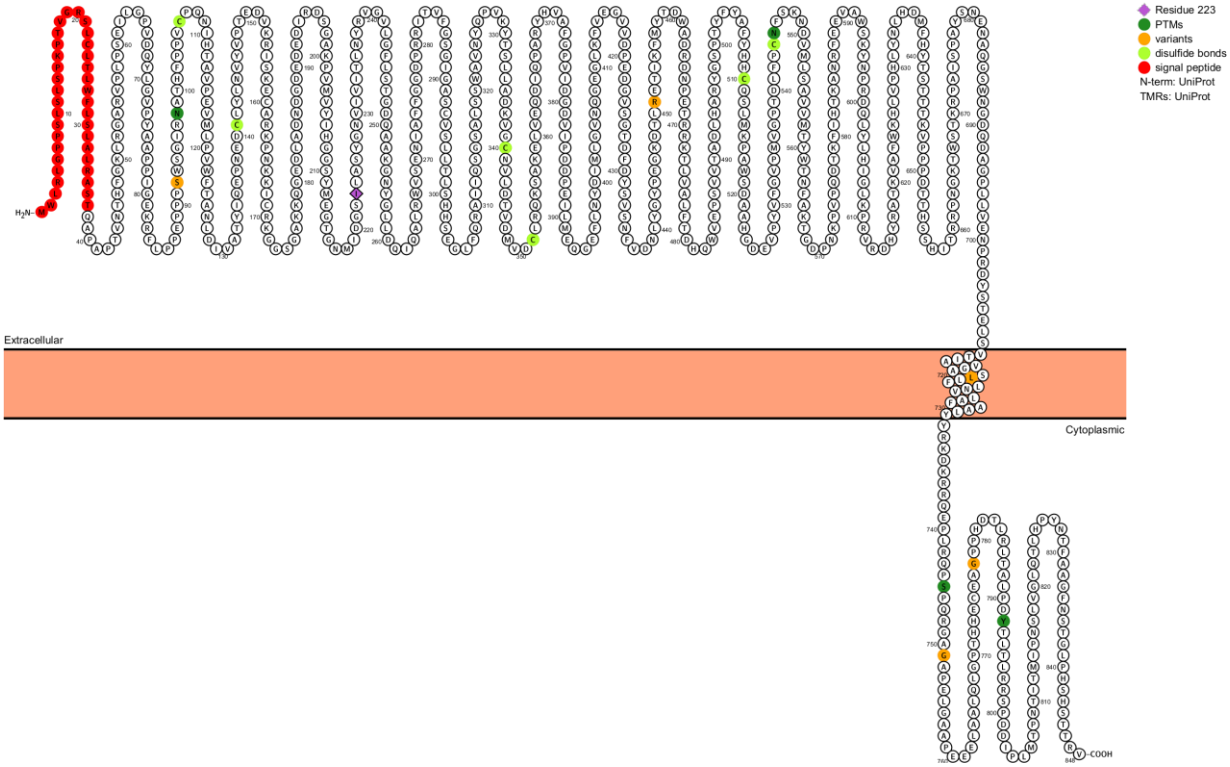


Figure 3. 2D structure of the Neuroligin-3 protein with the residue 223 marked.

4. Discussion

4.1. Phylogenetic framework.

The differences in topology seen in our obtained phylogenetic tree when compared to the guide tree [51] may be due to the fact that our tree was constructed based on a small subgroup of genes. Our tree represents the gene evolution of a set of genes that are connected to a specific characteristic, each of them having their own evolutionary characteristics, and with a small subgroup of genes (7). Considering this, it is expected that these genes will provide a distinct evolutionary reconstruction and show differences in the most apical divisions of the tree, since this number of markers will not have a very strong phylogenetic signal in comparison with a phylogenomic inference (*i.e* gene tree *versus* species tree).

This is particularly noticeable in the *Panthera leo* and the *Tremarctos ornatus* species, which should be grouped together given the fact that they are both carnivores. The phylogenetic location of these two species in separate clades is concordant with low supported values of bb and SH-aLRT showing low confidence in their placement.

The occurrence of a high number of pseudogenes or non-functional sequences has led to a heterogeneous number of sequences within each gene dataset, which means that which species is also represented by a different number of genes. This can influence the topology (for instance, in the carnivore order). Moreover, the genes are located on the X chromosome in most of the species we analysed. The sex chromosomes are usually more difficult to sequence than the autossomes and until recently it existed little information about them. Nowadays, new technologies of third generation sequencing like PacBio [52] and Nanopore [53] make it effective to characterize sex chromosomes, however it is still difficult to sequence some regions given the high number of repetitive elements [54].

For some tree branches, some support values (bb and SH-aLRT) are high which contrast with the low sCF values. However, bootstraps values can sometimes be inflated if there is a high number of base pairs available (for example, in a concatenated alignment). This highlights the importance of having the concordance factors because they provide extra support information based on different statistics.

For some branches, the concordance factors could not be calculated. This can result from having a different number of species for each gene, as well as having little information for some genes in the alignment. This disagreement resulted, in some cases, in the inability of calculating the concordance values.

The branch of the *Rattus norvegicus* was eliminated. Branch lengths explain the percentage of mutations per branch. In our tree, apart from *R. norvegicus*, the marsupial branches in the outgroup are the longest, which was expected given the fact that they belong to a different mammalian class. The *R. norvegicus* branch is unexpectedly longer than the outgroup. Usually, a long branch can indicate that there are a lot of mutations in the markers used to construct the phylogeny, which means that, either our genes are very divergent in that particular species, or that it can be caused by sequencing/assembly problems. Going forward with the selection analysis, we decided to eliminate it since it could bias the background information, skewing the analysis.

Overall, the phylogeny is not the main priority in this study, as it will be used mostly as a basis for subsequent analysis. We obtained a correct and well supported topology for the primates order, which is our target group.

4.2. Selective pressures underlying the evolution of the X-linked genes

RELAX was one of the tests performed and is a hypothesis testing framework, evaluating if our set of test branches are under intensification or relaxation of selection. It does show evidence of positive or negative selection in our test branch. The parameter k indicates if the branch is under intensification or relaxation of selection, where k<1 indicates that selection has been relaxed along the test branch and k>1 indicates that selection has been intensified [46].

aBSREL is a branch and site test that indicates if, on the foreground branches, exists evidence of positive episodic selection in any sites. However, it does not test for selection in specific sites. This test performs the likelihood ratio test at each branch [48].

BUSTED (Branch-Site Unrestricted Statistical Test for Episodic Diversification) is also not a site-specific test, but it tests for positive selection analysing if the gene experienced positive selection at least at one site on at least one branch. It also calculates the Evidence ratio (ER) for each site whether a specific site has likely evolved under positive selection or not. However, it should not be interpreted has definitive evidence for positive selection at individual sites, as other tests like Contrast-FEL are more accurate [47].

The last test that was used in the selection analyses was Contrast-FEL that is a site-specific test. It provides us the site where the test branches are associated with a significantly different selective pressure compared to the rest [49].

4.2.1. Evolution of genes *AKAPA17A, ASMTL, NLGN4, SPRY3, VAMP7*

For these genes, the selection analyses present no significant results, meaning that none of them showed evidence of positive selection or relaxation/intensification of selection. Given that they are genes with functions connected to the neurologic activity in humans, the lack of results in the *Homo sapiens* lineage was expected. Having evidence of positive selection implies that some parts of the gene are being positively selected, potentially resulting in acquiring a new function. Since these genes are very connected with brain activity, any alteration in the protein could have very negative effects, given the fact that these genes have very specific functions.

4.2.2. Evolution of genes *NLGN3* and *IL9R*

For genes *NLGN3* and *IL9R*, we found sites that showed significant evidence of positive selection. However, not a lot of information is available about these sites, their specific function or the consequences of potential mutations on those sites in the human protein. The information available on the Uniprot database had no identification of the functional relevance of the mutation at those sites.

For the IL9R gene we found two sites, one of them being residue 184 and the other one residue 257.

For residue 184, no literature or references were found regarding changes in this site. On UniProt there were 3 variant entrances on this site: *i.)* rs762681741; *ii.)* rs762681741; *iii.)* rs1436555178). In all of these described variants the mutation is missense. The first and the second variant appears as two different variants but is a multiallelic one. The difference between those two is that in the first variant, the amino acid Alanine changed into Aspartic Acid given the base change of a cytosine for an adenine. This is a pathogenic missense mutation and is quite likely deleterious. In the second variant the amino acid change is from an Alanine to a Glycine given the base change of a cytosine to a guanine. The third variant is due to the change of an Alanine to a Threonine given the base change of a guanine to an adenine.

For the residue 257, we also found 3 variants in UniProt: *i)* COSM3913688; COSM3913689; COSM3913690; rs780443490; *ii)* rs756382464; *iii)* rs756382464. Out of these 3 variants, information was only available about the consequences of a mutation on this site on the first variant. This is characterized by the mutation of a Proline to a Leucine given the base change of a cytosine to a thymine. On Uniprot, the variant presents vague and ambiguous description, being assessed as benign but also as potentially deleterious. This variant also has 3 entries in the COSMIC database [55] (Catalogue of Somatic Mutations in Cancer) where it appears that the missense mutation of a Proline to a Leucine in site 257 can cause skin cutaneous melanoma (malignant) and affects IL-9 signalling pathways. There is a contradiction when comparing the information given on these two databases. UniProt identifies this variant as likely benign while COSMIC refers it as malignant. This discrepancy supports the fact that this site needs further investigation to understand the consequences of these mutations as well as its functioning.

The other two variants are multiallelic ones. The second variant is characterized by a mutation of a Proline to a Serine given the base change of a cytosine to a thymine, is a missense benign mutation and potentially deleterious. The final variant occurs due to a mutation of the Proline to a Threonine (base change of a cytosine to an adenine).

According to UniProt, both sites are in the Functional Domain-Fibronectin type-III that ranges from the sites 149 to 259.

Fibronectin is a large protein, crucial for the formation of the extracellular matrix and cell-to-cell interactions, that is composed of multiple repeats of three types of small domains: type-I, type-II and type-III. The Fibronectin type-III domain (FN3) is a compact, self-contained folding unit present in numerous animal's proteins that participate in ligand binding [56]. It also mediates a wide range of cellular interactions in cell adhesion, migration, growth and differentiation [57], and has been shown to be involved in binding substrate. Substrate-binding proteins and substrate binding domains constitute a group of proteins (and protein domains) commonly linked with membrane protein complexes involved in transport or signal transduction [58].

For the gene *NLGN3* we found only one site with evidence of positive selection, the site 223. This site only had one variant listed on Uniprot. This variant is characterized by a substitution of an Isoleucine to a Phenylalanine given the change of an adenine to a thymine.

Even though we did not find relevant information about this site, near it on site 236, a variant is described of an Asparagine to a Serine that could lead to X-linked autism. More studies need to be done to try and understand if variants in site 223 could also affect disorders such as autism and TIC, given the fact that *NLGN3* is a gene with some variants connected to these disorders. The site 451, even if not close to the site 223, it is the most prevalent site connected with autism with a lot more studies performed [59].

According to UniProt, site 223 is inside the Pocket 8 of the *NLGN3* protein. A protein binding pocket is a cavity on or within a protein that accommodates ligand binding. The surrounding amino acids shape the pocket's structure, location, physicochemical properties, and function [60]. Ligand binding is a fundamental part of the protein function, so any selection in this area might potentially signify a new function through modification of the ligands it can bind.

Given the fact that we found almost no information about the sites that showed signals of positive selection for both genes, we can hypothesize that the evidence of positive selection on these sites might suggest that they are important for the human protein function.

5. Concluding remarks and Future work

In this study we identified PSS that can be relevant for the study of X-linked disorders associated with brain function. We were able to identify three sites that showed evidence of positive selection in the genes *IL9R* and *NLGN3*. Even though we could not find prevalent information for this study, we can conclude that such sites might suggest that they are important for the human protein function. Therefore, there are other future analyse that could be done in order to further refine the obtained results.

Building on the previously mentioned idea that the indication of positive selection might suggest that the sites in question are important for the function of the human protein, would be interesting to consider in the future computational proteomic studies to try and shape the protein with multiple mutations on those sites and see how it would affect the 3D conformation, structure and biochemical function.

Clinical and population genetic studies can also be carried out to search for variants on those sites to try to understand if there are any individuals with the disease in study with the amino acid's variations in that residue. Given the fact that this is a basal study of the evolution of the chosen genes, it is important to have genomes of the *Homo sapiens* of multiple populations, including individuals with the syndromes to try to identify mutations that are associated with the syndrome relatively to other variants that may occurr in the general population.

Additionally, we could also have retrieved all the single nucleotide polymorphisms (SNPs) that are connected to the disorders in study from public databases and analyse the consequences of each SNP on the 3D structure of the protein, modelling the normal structure as well the structure of the mutated protein with the mutations that are described and compare and analyse both.

References

1 *Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P.* ***Molecular Biology of the Cell****, 4th ed.; Garland Science: New York, NY, USA, 2002.*

2. Bachtrog, D.; Kirkpatrick, M.; Mank, J.E.; McDaniel, S.F.; Pires, J.C.; Rice, W.R.; Valenzuela, N. Are All Sex Chromosomes Created Equal? *Trends in Genetics* 2011, *27*, 350–357.

3. Ross, M.T.; Grafham, D.V.; Coffey, A.J.; Scherer, S.; McLay, K.; Muzny, D.; Platzer, M.; Howell, G.R.; Burrows, C.; Bird, C.P.; Frankish, A.; Lovell, F.L.; Howe, K.L.; Ashurst, J.L.; Fulton, R.S.; Sudbrak, R.; Wen, G.; Jones, M.C.; Hurles, M.E.; Andrews, T.D.; Scott, C.E.; Searle, S.; Ramser, J.; Whittaker, A.; Deadman, R.; Carter, N.P.; Hunt, S.E.; Chen, R.; Cree, A.; Gunaratne, P.; Havlak, P.; Hodgson, A.; Metzker, M.L.; Richards, S.; Scott, G.; Steffen, D.; Sodergren, E.; Wheeler, D.A.; Worley, K.C.; Ainscough, R.; Ambrose, K.D.; Ansari-Lari, M.A.; Aradhya, S.; Ashwell, R.I.; Babbage, A.K.; Bagguley, C.L.; Ballabio, A.; Banerjee, R.; Barker, G.E.; Barlow, K.F.; Barrett, I.P.; Bates, K.N.; Beare, D.M.; Beasley, H.; Beasley, O.; Beck, A.; Bethel, G.; Blechschmidt, K.; Brady, N.; Bray-Allen, S.; Bridgeman, A.M.; Brown, A.J.; Brown, M.J.; Bonnin, D.; Bruford, E.A.; Buhay, C.; Burch, P.; Burford, D.; Burgess, J.; Burrill, W.; Burton, J.; Bye, J.M.; Carder, C.; Carrel, L.; Chako, J.; Chapman, J.C.; Chavez, D.; Chen, E.; Chen, G.; Chen, Y.; Chen, Z.; Chinault, C.; Ciccodicola, A.; Clark, S.Y.; Clarke, G.; Clee, C.M.; Clegg, S.; Clerc-Blankenburg, K.; Clifford, K.; Cobley, V.; Cole, C.G.; Conquer, J.S.; Corby, N.; Connor, R.E.; David, R.; Davies, J.; Davis, C.; Davis, J.; Delgado, O.; Deshazo, D.; Dhami, P.; Ding, Y.; Dinh, H.; Dodsworth, S.; Draper, H.; Dugan-Rocha, S.; Dunham, A.; Dunn, M.; Durbin, K.J.; Dutta, I.; Eades, T.; Ellwood, M.; Emery-Cohen, A.; Errington, H.; Evans, K.L.; Faulkner, L.; Francis, F.; Frankland, J.; Fraser, A.E.; Galgoczy, P.; Gilbert, J.; Gill, R.; Glöckner, G.; Gregory, S.G.; Gribble, S.; Griffiths, C.; Grocock, R.; Gu, Y.; Gwilliam, R.; Hamilton, C.; Hart, E.A.; Hawes, A.; Heath, P.D.; Heitmann, K.; Hennig, S.; Hernandez, J.; Hinzmann, B.; Ho, S.; Hoffs, M.; Howden, P.J.; Huckle, E.J.; Hume, J.; Hunt, P.J.; Hunt, A.R.; Isherwood, J.; Jacob, L.; Johnson, D.; Jones, S.; de Jong, P.J.; Joseph, S.S.; Keenan, S.; Kelly, S.; Kershaw, J.K.; Khan, Z.; Kioschis, P.; Klages, S.; Knights, A.J.; Kosiura, A.; Kovar-Smith, C.; Laird, G.K.; Langford, C.; Lawlor, S.; Leversha, M.; Lewis, L.; Liu, W.; Lloyd, C.; Lloyd, D.M.; Loulseged, H.; Loveland, J.E.; Lovell,

J.D.; Lozado, R.; Lu, J.; Lyne, R.; Ma, J.; Maheshwari, M.; Matthews, L.H.; McDowall, J.; McLaren, S.; McMurray, A.; Meidl, P.; Meitinger, T.; Milne, S.; Miner, G.; Mistry, S.L.; Morgan, M.; Morris, S.; Müller, I.; Mullikin, J.C.; Nguyen, N.; Nordsiek, G.; Nyakatura, G.; O'Dell, C.N.; Okwuonu, G.; Palmer, S.; Pandian, R.; Parker, D.; Parrish, J.; Pasternak, S.; Patel, D.; Pearce, A.V.; Pearson, D.M.; Pelan, S.E.; Perez, L.; Porter, K.M.; Ramsey, Y.; Reichwald, K.; Rhodes, S.; Ridler, K.A.; Schlessinger, D.; Schueler, M.G.; Sehra, H.K.; Shaw-Smith, C.; Shen, H.; Sheridan, E.M.; Shownkeen, R.; Skuce, C.D.; Smith, M.L.; Sotheran, E.C.; Steingruber, H.E.; Steward, C.A.; Storey, R.; Swann, R.M.; Swarbreck, D.; Tabor, P.E.; Taudien, S.; Taylor, T.; Teague, B.; Thomas, K.; Thorpe, A.; Timms, K.; Tracey, A.; Trevanion, S.; Tromans, A.C.; d'Urso, M.; Verduzco, D.; Villasana, D.; Waldron, L.; Wall, M.; Wang, Q.; Warren, J.; Warry, G.L.; Wei, X.; West, A.; Whitehead, S.L.; Whiteley, M.N.; Wilkinson, J.E.; Willey, D.L.; Williams, G.; Williams, L.; Williamson, A.; Williamson, H.; Wilming, L.; Woodmansey, R.L.; Wray, P.W.; Yen, J.; Zhang, J.; Zhou, J.; Zoghbi, H.; Zorilla, S.; Buck, D.; Reinhardt, R.; Poustka, A.; Rosenthal, A.; Lehrach, H.; Meindl, A.; Minx, P.J.; Hillier, L.W.; Willard, H.F.; Wilson, R.K.; Waterston, R.H.; Rice, C.M.; Vaudin, M.; Coulson, A.; Nelson, D.L.; Weinstock, G.; Sulston, J.E.; Durbin, R.; Hubbard, T.; Gibbs, R.A.; Beck, S.; Rogers, J.; Bentley, D.R. The DNA sequence of the human X chromosome. *Nature* **2005**, *434*, 325-337. DOI: 10.1038/nature03440.

4. dos Santos, C.S.; Mendes, T.; Antunes, A. The Genes from the Pseudoautosomal Region 1 (PAR1) of the Mammalian Sex Chromosomes: Synteny, Phylogeny and Selection. *Genomics* **2022**, *114*, doi:10.1016/j.ygeno.2022.110419.

5. Mangs, A.H.; Morris, B.J. *The Human Pseudoautosomal Region (PAR): Origin, Function and Future*; 2007; Vol. 8;.

6. Hinch, A.G.; Altemose, N.; Noor, N.; Donnelly, P.; Myers, S.R. Recombination in the Human Pseudoautosomal Region PAR1. *PLoS Genet* **2014**, *10*, doi:10.1371/journal.pgen.1004503.

7. Cantone, I.; Fisher, A.G. Human X Chromosome Inactivation and Reactivation: Implications for Cell Reprogramming and Disease. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2017, *372*.

8. Powell-Hamilton, N.N. Overview of Sex Chromosome Abnormalities. MSD Manual. Available online: https://www.msdmanuals.com/home/children-s-health-

issues/chromosome-and-gene-abnormalities/overview-of-sex-chromosome-abnormalities (accessed on July 2024).

9. **Daltonismo: O que precisa de saber**. Lusiadas. Available online: https://www.lusiadas.pt/blog/prevencao-estilo-vida/saude-familia/daltonismo-que-precisa-saber (accessed on 29 July 2024).

10. **Hemofilia: Sintomas e diferentes tipos**. Available online: https://www.lusiadas.pt/blog/doencas/sintomas-tratamentos/hemofilia-sintomas-diferentes-tipos (accessed on 29 July 2024).

11. Hemani, F.; Niaz, S.; Kumar, V.; Khan, S.; Choudry, E.; Ali, S.R. A Case of Early Diagnosis of Turner Syndrome in a Neonate. *Cureus* **2021**, doi:10.7759/cureus.16733.

12. Kim, S.H.; Park, M.J.; Cho, E.H.; Kim, S.; Yoo, S.J. Different Parental Origins of Supernumerary X Chromosomes in Brothers with Klinefelter Syndrome: A Case Report. *Medicine* **2019**, *98*, e17838, doi:10.1097/MD.0000000000017838.

13. **Fragile X Syndrome**. Available online: https://medlineplus.gov/genetics/condition/fragile-x-syndrome (accessed on 29 July 2024).

14. Pereira, G.; Dória, S. X-Chromosome Inactivation: Implications in Human Disease. *J Genet* 2021, *100*.

15. Johnson, K.A.; Worbe, Y.; Foote, K.D.; Butson, C.R.; Gunduz, A.; Okun, M.S. Tourette Syndrome: Clinical Features, Pathophysiology, and Treatment. *Lancet Neurol* 2023, *22*, 147–158.

16. Kenney, C.; Kuo, S.H.; Jimenez-Shahed, J. Tourette's syndrome. *Am. Fam. Physician* **2008**, *77*, 651–658. PMID: 18350763.

17. Leckman, J.F. Tourette's Syndrome. In Proceedings of the Lancet; Elsevier B.V., November 16 2002; Vol. 360, pp. 1577–1586.

18. Lin, W. De; Tsai, F.J.; Chou, I.C. Current Understanding of the Genetics of Tourette Syndrome. *Biomed J* 2022, *45*, 271–279.

19. Pievsky, M.A.; McGrath, R.E. The Neurocognitive Profile of Attention-Deficit/Hyperactivity Disorder: A Review of Meta-Analyses. *Archives of Clinical Neuropsychology* 2018, *33*, 143–157.

20. Brem, S.; Grünblatt, E.; Drechsler, R.; Riederer, P.; Walitza, S. The Neurobiological Link between OCD and ADHD. *ADHD Attention Deficit and Hyperactivity Disorders* 2014, *6*, 175–202.

21. Canitano, R.; Vivanti, G. Tics and Tourette Syndrome in Autism Spectrum Disorders. *Autism* **2007**, *11*, 19–28, doi:10.1177/1362361307070992.

22. Wang, S.; Wang, B.; Drury, V.; Drake, S.; Sun, N.; Alkhairo, H.; Arbelaez, J.; Duhn, C.; Bromberg, Y.; Brown, L.W.; et al. Rare X-Linked Variants Carry Predominantly Male Risk in Autism, Tourette Syndrome, and ADHD. *Nat Commun* **2023**, *14*, doi:10.1038/s41467-023-43776-0.

23. Holland, K. Motor Tics: Understanding Chronic Tic Motor Disorder. Medically reviewed by Moawad, H. *Healthline*. Updated on 25 March 2022. Available online: https://www.healthline.com/health/chronic-motor-tic-disorder (accessed on 29 July 2024).

24. Lawson-Yuen, A.; Saldivar, J.S.; Sommer, S.; Picker, J. Familial Deletion within NLGN4 Associated with Autism and Tourette Syndrome. *European Journal of Human Genetics* **2008**, *16*, 614–618, doi:10.1038/sj.ejhg.5202006.

25. Stelzer, G.; Rosen, N.; Plaschkes, I.; Zimmerman, S.; Twik, M.; Fishilevich, S.; Stein, T.I.; Nudel, R.; Lieder, I.; Mazor, Y.; Kaplan, S.; Dahary, D.; Warshawsky, D.; Guan-Golan, Y.; Kohn, A.; Rappaport, N.; Safran, M.; Lancet, D. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinformatics* **2016**, *54*, 1.30.1–1.30.33. doi: 10.1002/cpbi.5. PMID: 27322403.

26. Skakkebæk, A.; Nielsen, M.M.; Trolle, C.; Vang, S.; Hornshøj, H.; Hedegaard, J.; Wallentin, M.; Bojesen, A.; Hertz, J.M.; Fedder, J.; et al. DNA Hypermethylation and Differential Gene Expression Associated with Klinefelter Syndrome. *Sci Rep* **2018**, *8*, doi:10.1038/s41598-018-31780-0.

27. National Center for Biotechnology Information (NCBI). Available online: https://www.ncbi.nlm.nih.gov (accessed on 29 July 2024). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988].

28. Maccarini, S.; Cipani, A.; Bertini, V.; Skripac, J.; Salvi, A.; Borsani, G.; Marchina, E. Inherited Duplication of the Pseudoautosomal Region Xq28 in a Subject with Gilles de La Tourette Syndrome and Intellectual Disability: A Case Report. *Mol Cytogenet* **2020**, *13*, doi:10.1186/s13039-020-00493-3.

29. Tchigvintsev, A.; Tchigvintsev, D.; Flick, R.; Popovic, A.; Dong, A.; Xu, X.; Brown, G.; Lu, W.; Wu, H.; Cui, H.; et al. Biochemical and Structural Studies of Conserved Maf Proteins Revealed Nucleotide Pyrophosphatases with a Preference for Modified Nucleotides. *Chem Biol* **2013**, *20*, 1386–1398, doi:10.1016/j.chembiol.2013.09.011.

30. Harrison, P.W.; Amode, M.R.; Austine-Orimoloye, O.; Azov, A.G.; Barba, M.; Barnes, I.; Becker, A.; Bennett, R.; Berry, A.; Bhai, J.; Bhurji, S.K.; Boddu, S.; Branco Lins, P.R.; Brooks, L.; Ramaraju, S.B.; Campbell, L.I.; Martinez, M.C.; Charkhchi, M.; Chougule, K.; Cockburn, A.; Davidson, C.; De Silva, N.H.; Dodiya, K.; Donaldson, S.; El Houdaigui, B.; Naboulsi, T.E.; Fatima, R.; Giron, C.G.; Genez, T.; Grigoriadis, D.; Ghattaoraya, G.S.; Martinez, J.G.; Gurbich, T.A.; Hardy, M.; Hollis, Z.; Hourlier, T.; Hunt, T.; Kay, M.; Kaykala, V.; Le, T.; Lemos, D.; Lodha, D.; Marques-Coelho, D.; Maslen, G.; Merino, G.A.; Mirabueno, L.P.; Mushtaq, A.; Hossain, S.N.; Ogeh, D.N.; Sakthivel, M.P.; Parker, A.; Perry, M.; Piližota, I.; Poppleton, D.; Prosovetskaia, I.; Raj, S.; Pérez-Silva, J.G.; Salam, A.I.A.; Saraf, S.; Saraiva-Agostinho, N.; Sheppard, D.; Sinha, S.; Sipos, B.; Sitnik, V.; Stark, W.; Steed, E.; Suner, M.M.; Surapaneni, L.; Sutinen, K.; Tricomi, F.F.; Urbina-Gómez, D.; Veidenberg, A.; Walsh, T.A.; Ware, D.; Wass, E.; Willhoft, N.L.; Allen, J.; Alvarez-Jarreta, J.; Chakiachvili, M.; Flint, B.; Giorgetti, S.; Haggerty, L.; Ilsley, G.R.; Keatley, J.; Loveland, J.E.; Moore, B.; Mudge, J.M.; Naamati, G.; Tate, J.; Trevanion, S.J.; Winterbottom, A.; Frankish, A.; Hunt, S.E.; Cunningham, F.; Dyer, S.; Finn, R.D.; Martin, F.J.; Yates, A.D. Ensembl 2024. *Nucleic Acids Res.* **2024**, *52*(D1), D891–D899. doi: 10.1093/nar/gkad1049. PMID: 37953337; PMCID: PMC10767893.

31. Bateman, A.; Martin, M.J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E.H.; Britto, R.; Bye-A-Jee, H.; Cukura, A.; et al. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **2023**, *51*, D523–D531, doi:10.1093/nar/gkac1052.

32. Rebhan, M.; Chalifa-Caspi, V.; Prilusky, J.; Lancet, D. GeneCards: A Novel Functional Genomics Compendium with Automated Data Mining and Query Reformulation Support. *Bioinformatics* **1998**, *14*, 656–664. https://doi.org/10.1093/bioinformatics/14.8.656.

33. Slater, G.S.C.; Birney, E. Automated Generation of Heuristics for Biological Sequence Comparison. *BMC Bioinformatics* **2005**, *6*, doi:10.1186/1471-2105-6-31.

34. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **2018**, *35*, 1547–1549, doi:10.1093/molbev/msy096.

35. Edgar, R.C. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Res* **2004**, *32*, 1792–1797, doi:10.1093/nar/gkh340.

36. Sela, I.; Ashkenazy, H.; Katoh, K.; Pupko, T. GUIDANCE2: Accurate Detection of Unreliable Alignment Regions Accounting for the Uncertainty of Multiple Parameters. *Nucleic Acids Res* **2015**, *43*, W7–W14, doi:10.1093/nar/gkv318.

37. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **2013**, *30*, 772–780, doi:10.1093/molbev/mst010.

38. Xia, X. DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. *Mol Biol Evol* **2018**, *35*, 1550–1552, doi:10.1093/molbev/msy073.

39. Xia, X.; Xie, Z.; Salemi, M.; Chen, L.; Wang, Y. An Index of Substitution Saturation and Its Application. *Mol Phylogenet Evol* **2003**, *26*, 1–7, doi:10.1016/S1055-7903(02)00326-3.

40. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; Von Haeseler, A.; Lanfear, R.; Teeling, E. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **2020**, *37*, 1530–1534, doi:10.1093/molbev/msaa015.

41. Sharma, S.; Kumar, S. Fast and Accurate Bootstrap Confidence Limits on Genome-Scale Phylogenies Using Little Bootstraps. *Nat Comput Sci* **2021**, *1*, 573–577, doi:10.1038/s43588-021-00129-5.

42. Anisimova, M.; Gascuel, O. Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Syst Biol* **2006**, *55*, 539–552, doi:10.1080/10635150600755453.

43. Minh, B.Q.; Hahn, M.W.; Lanfear, R. New Methods to Calculate Concordance Factors for Phylogenomic Datasets. *Mol Biol Evol* **2020**, *37*, 2727–2733, doi:10.1093/molbev/msaa106.

44. **Figtree**. Version 1.4.4. Available online: http://tree.bio.ed.ac.uk/software/figtree/ (accessed on 29 July 2024).

45. Kosakovsky Pond, S.L.; Poon, A.F.Y.; Velazquez, R.; Weaver, S.; Hepler, N.L.; Murrell, B.; Shank, S.D.; Magalis, B.R.; Bouvier, D.; Nekrutenko, A.; et al. HyPhy 2.5 - A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol Biol Evol* **2020**, *37*, 295–299, doi:10.1093/molbev/msz197.

46. Wertheim, J.O.; Murrell, B.; Smith, M.D.; Pond, S.L.K.; Scheffler, K. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. *Mol Biol Evol* **2015**, *32*, 820–832, doi:10.1093/molbev/msu400.

47. Murrell, B.; Weaver, S.; Smith, M.D.; Wertheim, J.O.; Murrell, S.; Aylward, A.; Eren, K.; Pollner, T.; Martin, D.P.; Smith, D.M.; et al. Gene-Wide Identification of Episodic Selection. *Mol Biol Evol* **2015**, *32*, 1365–1371, doi:10.1093/molbev/msv035.

48. Smith, M.D.; Wertheim, J.O.; Weaver, S.; Murrell, B.; Scheffler, K.; Kosakovsky Pond, S.L. Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. *Mol Biol Evol* **2015**, *32*, 1342–1353, doi:10.1093/molbev/msv022.

49. Kosakovsky Pond, S.L.; Wisotsky, S.R.; Escalante, A.; Magalis, B.R.; Weaver, S. Contrast-FEL - A Test for Differences in Selective Pressures at Individual Sites among Clades and Sets of Branches. *Mol Biol Evol* **2021**, *38*, 1184–1198, doi:10.1093/molbev/msaa263.

50. Omasits, U.; Ahrens, C.H.; Müller, S.; Wollscheid, B. Protter: Interactive Protein Feature Visualization and Integration with Experimental Proteomic Data. *Bioinformatics* **2014**, *30*, 884–886, doi:10.1093/bioinformatics/btt607.

51. Esselstyn, J.A.; Oliveros, C.H.; Swanson, M.T.; Faircloth, B.C. Investigating Difficult Nodes in the Placental Mammal Tree with Expanded Taxon Sampling and Thousands of Ultraconserved Elements. *Genome Biol Evol* **2017**, *9*, 2308–2321, doi:10.1093/gbe/evx168.

52. Rhoads, A.; Au, K.F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* 2015, *13*, 278–289.

53. Wang, Y.; Zhao, Y.; Bollas, A.; Wang, Y.; Au, K.F. Nanopore Sequencing Technology, Bioinformatics and Applications. *Nat Biotechnol* 2021, *39*, 1348–1365.

54. Xiao, T.; Zhou, W. The Third Generation Sequencing: The Advanced Approach to Genetic Diseases. *Transl Pediatr* 2020, *9*, 163–173.

55. Tate, J.G.; Bamford, S.; Jubb, H.C.; Sondka, Z.; Beare, D.M.; Bindal, N.; Boutselakis, H.; Cole, C.G.; Creatore, C.; Dawson, E.; et al. COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **2019**, *47*, D941–D947, doi:10.1093/nar/gky1015.

56. Koide, A.; Bailey, C.W.; Huang, X.; Koide, S. *The Fibronectin Type III Domain as a Scaffold for Novel Binding Proteins*;

57. Zhao, J.; Ren, J.; Wang, N.; Cheng, Z.; Yang, R.; Lin, G.; Guo, Y.; Cai, D.; Xie, Y.; Zhao, X. Crystal Structure of the Second Fibronectin Type III (FN3) Domain from Human Collagen A1 Type XX. *Acta Crystallographica Section F:Structural Biology Communications* **2017**, *73*, 695–700, doi:10.1107/S2053230X1701648X.

58. Berntsson, R.P.A.; Smits, S.H.J.; Schmitt, L.; Slotboom, D.J.; Poolman, B. A Structural Classification of Substrate-Binding Proteins. *FEBS Lett* 2010, *584*, 2606–2617.

59. Wang, L.; Mirabella, V.R.; Dai, R.; Su, X.; Xu, R.; Jadali, A.; Bernabucci, M.; Singh, I.; Chen, Y.; Tian, J.; et al. Analyses of the Autism-Associated Neuroligin-3 R451C Mutation in Human Neurons Reveal a Gain-of-Function Synaptic Mechanism. *Mol Psychiatry* **2022**, doi:10.1038/s41380-022-01834-x.

60. Stank, A.; Kokh, D.B.; Fuller, J.C.; Wade, R.C. Protein Binding Pocket Dynamics. *Acc. Chem. Res.* **2016**, *49*(5), 809–815. doi: 10.1021/acs.accounts.5b00516. Epub 25 April 2016. PMID: 27110726.

Attachments

Attachment 1

Table S1. Information regarding the 24 Mammal Species used in this study.

Order	Family	WGS	Species	Level	Genome assembly	Bioproject	Size
Monotremate	Ornithorhynchidae	RZJT02	*Ornithorhynchus anatinus*	Chromossome	GCA_004115215.4	PRJNA534073	1.859,28
Diprotodontia	Macropodidae	JAQQTA01	*Macropus giganteus*	Chromossome	GCA_028627215.1	PRJNA512907	3.535,62
Dasyuromorphia	Dasyuridae	CACPPN01	*Sarcophilus harrisii*	Chromossome	GCA_902635505.1	PRJEB35073	3.086,67
Proboscidea	Elephantidae	JAMZQU01	*Elephas maximus*	Chromossome	GCA_024166365.1	PRJNA861314	3.401,25
Rodentia	Sciuridae	CACRXI02	*Sciurus carolinensis*	Chromossome	GCA_902686445.2	PRJEB35386	2.815,4
Rodentia	Dipodidae	JAJGSA01	*Jaculus Jaculus*	Chromossome	GCA_020740685.1	PRJNA778790	2.863,87
Rodentia	Muridae	JACYVU01	*Rattus norvegicus*	Chromossome	GCA_015227675.2	PRJNA677964	2.647,92
Lagomorpha	Leporidae	JABWOR01	*Oryctolagus cuniculus*	Chromossome	GCA_013371645.1	PRJNA635554	2.780,37
Chiroptera	Phyllostomidae	JAKFGA01	*Desmodus rotundus*	Chromossome	GCA_022682395.1	PRJNA789527	1.976,15
Carnivora	Ursidae	JAPYXY01	*Tremarctos ornatus*	Chromossome	GCA_028551375.1	PRJNA512907	2.342,01
Carnivora	Felidae	JAEQMX01	*Panthera leo*	Chromossome	GCA_018350215.1	PRJNA751610	2.297,57
Perissodactyla	Rhinocerotidae	JAJIAY01	*Diceros bicornis*	Chromossome	GCA_020826845.1	PRJNA773944	3.005,52
Artiodactyla	Camelidae	JWIN03	*Camelus dromedarius*	Chromossome	GCA_000803125.3	PRJNA565028	2.169,36
Cetacea	Balenopterids	CATLKF01	*Balaenoptera acutorostrata*	Chromossome	GCA_949987535.1	PRJEB61579	2.772,9
Primates	Hominidae	JAQQLK01	*Gorilla gorilla*	Chromossome	GCA_028885475.1	PRJNA916733	3.310,53
Primates	Hominidae	JAQQLO01	*Pan troglodytes*	Chromossome	GCA_028858805.1	PRJNA916737	3.015,38
Primates	Hominidae	JAQQLL01	*Pan paniscus*	Chromossome	GCA_028858845.1	PRJNA916735	3.224,71
Primates	Hominidae	JAQQLT01	*Pongo pygmaeus*	Chromossome	GCA_028885525.1	PRJNA916740	3.038,23
Primates	Hylobatidae	JAQQLH01	*Symphalangus syndactylus*	Chromossome	GCA_028878055.1	PRJNA916728	3,198,446,238
Primates	Lemuridae	JAJGPY01	*Lemur catta*	Chromossome	GCA_020740605.1	PRJNA779062	2.245,6
Primates	Cercopithecidae	VSDM01	*Macaca mulatta*	Chromossome	GCA_008058575.1	PRJNA514196	3.037,16
Primates	Lorisidae	JAPZEE01	*Nycticebus coucang*	Chromossome	GCA_027406575.1	PRJNA927103	2.917,15
Primates	Callitrichidae	JAALXQ01	*Callithrix jacchus*	Chromossome	GCA_011078405.1	PRJNA558086	2.811,15

Primates	Hominidae	-	*Homo sapiens*	Chromossome	GCA_000001405.29	PRJNA31257	3,099,441,038

Attachment 2

Table S2. Selected genes (*ASMTL, AKAP17A, NLGN4X, NLGN3, VAMP7, SPTY3* and *IL9R*) , their location and overall function

Gene	Symbol	Full name	Location	General Function
ASMTL	ASMTL	Acetylserotonin O-mehtyltranferase like	Xp22.3	Nucleoside triphosphate pyrophosphatase that hydrolyzes dTTP and UTP; it Can also hydrolyze CTP and the modified nucleotides pseudo-UTP, 5-methyl-UTP (m5UTP) and 5-methyl-CTP (m5CTP). May have a dual role in cell division arrest and in preventing the incorporation of modified nucleotides into cellular nucleic acids. In addition, it is probable that the presence of the putative catalytic domain of S-adenosyl-L-methionine binding in the C-terminal region argues for a methyltransferase activity. (UniProt)
AKAP17A	AKAP17A	A-kinase anchoring protein 17A	Xp22.33	Splice factor regulating alternative splice site selection for certain mRNA precursors. It also Mediates regulation of pre-mRNA splicing in a PKA-dependent manner and is widely expressed. It can be found in the heart, brain, lung, liver, skeletal muscle, kidney and pancreas. Expressed in activated B-cells and placenta
NLGN4X	NLGN4X	Neuroligin 4 X-linked	Xp22.32	This gene encodes a member of the type-B carboxylesterase/lipase protein family. Members of this family may act as splice site-specific ligands for beta-neurexins and may be involved in the formation and remodelling of central nervous system synapses (NCBI).
NLGN3	NLGN3	Neuroligin 3	Xq13.1	This gene encodes a member of a family of neuronal cell surface proteins. Members of this family may act as splice site-specific ligands for beta-neurexins and may be involved in the formation and remodelling of central nervous system synapses (UniProt). Plays a role in synapse function and synaptic signal transmission and may mediate its effects by clustering other synaptic proteins. May promote the initial formation of synapses but is not essential for this. May also play a role in glia-glia or glia-neuron interactions in the developing peripheral nervous system (By similarity) (NCBI).
				Encodes a transmembrane protein that is a member of the soluble N-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) family. The encoded protein localizes to late

VAMP7	VAMP7	Vesicle associated membrane protein 7	Xq28	endosomes and lysosomes and is involved in the fusion of transport vesicles to their target membranes. Alternate splicing results in multiple transcript variants
SPRY3	SPRY3	Sprouty RTK signaling antogonist 3	Xq28	Involved in negative regulation of MAPK cascade
IL9R	IL9R	Interleukin 9 receptor	Xq28	A The protein encoded by this gene is a cytokine receptor that specifically mediates the biological effects of interleukin 9 (IL9).

Attachment 3

Table S3. Selected genes (*ASMTL, AKAP17A, NLGN4X, NLGN3, VAMP7, SPRY3, IL9R)* with the corresponding protein entries

Gene	Protein	Entry	Entry name
ASMTL	Probable bifunctional dTTP/UTP pyrophosphate/methyl transferase	O95671	ASML_HUMAN
AKAP17A	A-kinase anchor protein 17A	Q02040	AK17A_HUMAN
NLGN4X	Neuroligin-4, X-linked	Q8N0W4	NLGNX_HUMAN
NLGN3	Neuroligin-3	Q9NZ94	NLGN3_HUMAN
VAMP7	Vesicle-associated membrane protein 7	P51809	VAMP7_HUMAN
SPRY3	Protein Sprouty homolog 3	O43610	SPY3_HUMAN
IL9R	Interleukin-9 receptor	Q01113	IL9R_HUMAN

Attachment 4

Table S4. Selected species and its respective introns for *NLGN3*

Order	Family	Species	Target	Introns
Primates	Hominidae	*Homo sapiens*	0 849 . CM004615.1 66501074 66523746 + 4477	6
Primates	Hominidae	*Gorilla gorilla*	194 849 . CM054581.1 30032236 30150128 + 2689	3
Primates	Hominidae	*Pan troglodytes*	-	-
Primates	Hominidae	*Pan paniscus*	0 849 . CM054507.1 70828723 70850984 + 4476	6
Primates	Hominidae	*Pongo pygmaeus*	-	-
Primates	Hylobatidae	*Symphalangus syndactylus*	Haplotypes	-
Primates	Lemuridae	*Lemur catta*	0 849 . CM036499.1 47596155 47578489 - 4465	6
Primates	Cercopithecidae	*Macaca mulatta*	0 849 . CM017758.1 68812173 68834925 + 4475	6
Primates	Lorisidae	*Nycticebus coucang*	0 849 . CM050161.1 79480609 79503167 + 4441	6
Primates	Callitrichidae	*Callithrix jacchus*	0 849 . CM021881.1 65309037 65329709 + 4458	6
Cetacea	Balenopterids	*Balaenoptera acutorostrata*	0 849 . OX465355.1 61559629 61577647 + 4464	6
Artiodactyla	Camelidae	*Camelus dromedarius*	0 849 . CM016663.2 65744576 65726476 - 4472	6
Perissodactyla	Rhinocerotidae	*Diceros bicornis*	0 849 . CM036987.1 64456352 64474350 + 4412	6
Carnivora	Felidae	*Panthera leo*	-	-
Carnivora	Ursidae	*Tremarctos ornatus*	0 849 . CM052623.1 58823984 58805946 - 4458	6
Chiroptera	Phyllostomidae	*Desmodus rotundus*	-	-
Lagomorpha	Leporidae	*Oryctolagus cuniculus*	0 849 . CM023793.1 51123498 51143752 + 4448	6
Rodentia	Muridae	*Rattus norvegicus*	0 849 . CM026994.1 66432310 66451558 + 4428	6
Rodentia	Dipodidae	*Jaculus jaculus*	0 849 . CM036412.1 63013657 63033872 + 4434	6
Rodentia	Sciuridae	*Sciurus carolinensis*	0 849 . LR738601.1 54350981 54369203 + 4445	6
Proboscidea	Elephantidae	*Elephas maximus*	0 849 . CM044047.1 98543724 98525906 - 4453	6
Dasyuromorphia	Dasyuridae	*Sarcophilus harrisii*	Pseudogene	-
Diprotodontia	Macropodidae	*Macropus giganteus*	Pseudogene	-
Monotremate	Ornithorhynchidae	*Ornithorhynchus anatinus*	Pseudogene	-

Table S5. Selected species and its respective intron for *AKAP17A*

Order	Family	Species	Target	Introns
Primates	Hominidae	*Homo sapiens*	0 696 . CM004615.1 1431950 1440036 + 3534	3
Primates	Hominidae	*Gorilla gorilla*	0 696 . CM054581.1 8354709 8363733 + 3487	3
Primates	Hominidae	*Pan troglodytes*	-	-
Primates	Hominidae	*Pan paniscus*	0 696 . CM054507.1 1491967 1500200 + 3482	3
Primates	Hominidae	*Pongo pygmaeus*	-	-
Primates	Hylobatidae	*Symphalangus syndactylus*	-	-
Primates	Lemuridae	*Lemur catta*	0 696 . CM036499.1 72369124 72361514 - 2864	3
Primates	Cercopithecidae	*Macaca mulatta*	-	-
Primates	Lorisidae	*Nycticebus coucang*	0 696 . CM050161.1 1302809 1310292 + 2767	3
Primates	Callitrichidae	*Callithrix jacchus*	Pseudogene	-
Cetacea	Balenopterids	*Balaenoptera acutorostrata*	0 696 . OX465355.1 1136512 1142897 + 2719	3
Artiodactyla	Camelidae	*Camelus dromedarius*	Pseudogene	-
Perissodactyla	Rhinocerotidae	*Diceros bicornis*	Pseudogene	-
Carnivora	Felidae	*Panthera leo*	0 696 . CM031449.1 1118986 1128095 + 2699	4
Carnivora	Ursidae	*Tremarctos ornatus*	0 696 . CM052623.1 107848378 107841565 - 3004	3
Chiroptera	Phyllostomidae	*Desmodus rotundus*	0 696 . CM040287.1 1908980 1915043 + 2851	3
Lagomorpha	Leporidae	*Oryctolagus cuniculus*	-	-
Rodentia	Muridae	*Rattus norvegicus*	-	-
Rodentia	Dipodidae	*Jaculus jaculus*	-	-
Rodentia	Sciuridae	*Sciurus carolinensis*	0 696 . LR738601.1 998054 1003765 + 2543	3
Proboscidea	Elephantidae	*Elephas maximus*	-	-
Dasyuromorphia	Dasyuridae	*Sarcophilus harrisii*	-	-
Diprotodontia	Macropodidae	*Macropus giganteus*	-	
Monotremate	Ornithorhynchidae	*Ornithorhynchus anatinus*	not complete	

Table S6. Selected species and its respective intron for *ASMTL*

Family	Species	Target	Intr
Hominidae	*Homosapiens*	0 622 . CM004615.1 1283070 1229366 - 3070	
Hominidae	*Gorilla gorilla*	-	
Hominidae	*Pan troglodytes*	-	
Hominidae	*Pan paniscus*	0 622 . CM054507.1 1354981 1290510 - 2973	
Hominidae	*Pongo pygmaeus*	-	
Hylobatidae	*Symphalangus syndactylus*	-	
Lemuridae	*Lemur catta*	Pseudogene	
Cercopithecidae	*Macaca mulatta*	0 622 . CM017758.1 2759205 2714824	
Lorisidae	*Nycticebus coucang*	Pseudogene	
Callitrichidae	*Callithrix jacchus*	211 622 . CM021881	
Balenopterids	*Balaenoptera acutorostrata*		
Camelidae	*Camelus dromedarius*		
Rhinocerotidae	*Diceros bicornis*		
Felidae	*Panthera leo*		
Ursidae	*Tremarctos ornatus*		
Phyllostomidae	*Desmodus rotundus*		
Leporidae	*Oryctolagus cuniculus*		
Muridae	*Rattus norvegicus*		
Dipodidae	*Jaculus jaculus*		
Sciuridae	*Sciurus caroline*		
Elephantidae	*Eleph*		
Dasyuridae			
Macropod			
Orn			

Table S7. Selected species and its respective intron for *IL9R*

Family	Species	Target	Introns
Hominidae	*Homo sapiens*	0 522 . CM004615.1 148565534 148578496 + 2726	8
Hominidae	*Gorilla gorilla*	Pseudogene	-
Hominidae	*Pan troglodytes*	Pseudogene	-
Hominidae	*Pan paniscus*	0 522 . CM054507.1 155981120 155994040 + 2608	8
Hominidae	*Pongo pygmaeus*	Pseudogene	-
Hylobatidae	*Symphalangus syndactylus*	Pseudogene	-
Lemuridae	*Lemur catta*	9 433 . CM036473.1 151143 144390 - 1499	7
Cercopithecidae	*Macaca mulatta*	0 434 . CM017758.1 152162130 152172504 + 2047 and 431 522 . CM017758.1 152172474 152172747 + 449	8
Lorisidae	*Nycticebus coucang*	0 427 . CM050148.1 96489254 96477316 - 1456	8
Callitrichidae	*Callithrix jacchus*	0 433 . CM021881.1 146864504 146876760 + 1912 and 421 522 . CM021881.1 146876700 146877006 + 313	8
Balenopterids	*Balaenoptera acutorostrata*	0 431 . OX465365.1 39882997 39893731 + 1392 and 414 497 . OX465365.1 39891268 39893899 + 156	8
Camelidae	*Camelus dromedarius*	0 377 . JWIN03000071.1 4938793 4948672 + 1323 and 390 431 . JWIN03000071.1 4948693 4948816 + 129 and 428 497 . JWIN03000071.1 4948777 4948984 + 152	8
Rhinocerotidae	*Diceros bicornis*	0 431 . CM036971.1 48229588 48238716 + 1415 and 428 506 . CM036971.1 48238677 48238911 + 189	8
Felidae	*Panthera leo*	0 429 . CM031465.1 41338219 41347631 + 1416	8
Ursidae	*Tremarctos ornatus*	0 429 . CM052606.1 63702987 63692930 - 1379	8
Phyllostomidae	*Desmodus rotundus*	27 442 . CM040274.1 113230372 113548428 + 1227	8
Leporidae	*Oryctolagus cuniculus*	18 427 . CM023777.1 145371 139469 - 1326 and 432 493 . CM023777.1 139487 139304 - 158	7
Muridae	*Rattus norvegicus*	-	-
Dipodidae	*Jaculus jaculus*	47 431 . CM036399.1 113061291 113068963 + 1240	6
Sciuridae	*Sciurus carolinensis*	0 296 . LR738608.1 41539301 41545887 + 1128	6
Elephantidae	*Elephas maximus*	48 324 . CM044031.1 57358515 57355483 - 1008 and 386 431 . CM044031.1 57354272 57354137 - 152	5
Dasyuridae	*Sarcophilus harrisii*	-	-
Macropodidae	*Macropus giganteus*	-	-
Ornithorhynchidae	*Ornithorhynchus anatinus*	-	-

Table S8. Selected species and its respective intron for *SPRY3*

Table S9. Selected species and its respective intron for *VAMP7*

Order	Family	Species	Target	Introns
Primates	Hominidae	*Homo sapiens*	0 221 . CM004615.1 148455328 148508670 + 1052	6
Primates	Hominidae	*Gorilla gorilla*	-	-
Primates	Hominidae	*Pan troglodytes*	-	-
Primates	Hominidae	*Pan paniscus*	0 221 . CM054507.1 155873214 155925887 + 1054	6
Primates	Hominidae	*Pongopygmaeus*	-	-
Primates	Hylobatidae	*Symphalangus syndactylus*	-	-
Primates	Lemuridae	*Lemur catta*	0 221 . CM036499.1 8255178 8306839 + 1055	6
Primates	Cercopithecidae	*Macaca mulatta*	0 221 . CM017758.1 152078446 152127836 + 1055	6
Primates	Lorisidae	*Nycticebus coucang*	0 221 . CM050161.1 187151514 187236956 + 1051	6
Primates	Callitrichidae	*Callithrix jacchus*	0 221 . CM021881.1 146797869 146851508 + 1053	6
Cetacea	Balenopterids	*Balaenoptera acutorostrata*	0 221 . OX465355.1 133483698 133533671 + 1054	6
Artiodactyla	Camelidae	*Camelus dromedarius*	0 221 . CM016663.2 152929 113535 -1055	6
Perissodactyla	Rhinocerotidae	*Diceros bicornis*	0 221 . CM036987.1 139869985 139905677 + 1051	6
Carnivora	Felidae	*Panthera leo*	-	-
Carnivora	Ursidae	*Tremarctos ornatus*	0 221 . CM052623.1 69722 18832 - 1051	6
Chiroptera	Phyllostomidae	*Desmodus rotundus*	-	-
Lagomorpha	Leporidae	*Oryctolagus cuniculus*	0 221 . CM023793.1 136661244 136709474 + 1058	6
Rodentia	Muridae	*Rattus norvegicus*	0 221 . CM026985.1 16735316 16762113 + 1005	6
Rodentia	Dipodidae	*Jaculus jaculus*	0 221 . CM036412.1 153964902 154004845 + 1003	6
Rodentia	Sciuridae	*Sciurus carolinensis*	0 167 . LR738601.1 131599397 131579784 - 795	4
Proboscidea	Elephantidae	*Elephas maximus*	0 221 . CM044047.1 104461 64908 - 1043	6
Dasyuromorphi	Dasyuridae	*Sarcophilus harrisii*	0 221 . LR735560.1 7607450 7582251 - 984	6
Diprotodontia	Macropodidae	*Macropus giganteus*	0 114 . CM053636.1 16699531 16697221 - 565 and 167 221 . CM053636.1 16683508 16672198 - 220	3
Monotremate	Ornithorhynchidae	*Ornithorhynchus anatinus*	0 221 . CM014205.1 14578852 14590997 + 996	6

Table S10. Selected species and its respective intron for *NLGN4X*

Order	Family	Species	Target	Introns
Primates	Hominidae	*Homo sapiens*	0 817 . CM004615.1 5827098 5566768 - 4349	4
Primates	Hominidae	*Gorilla gorilla*	0 817 . CM054581.1 29929226 30150128 + 4241	4
Primates	Hominidae	*Pantroglodytes*	-	
Primates	Hominidae	*Panpaniscus*	0 817 . CM054507.1 7699983 7438672 - 4349	4
Primates	Hominidae	*Pongo pygmaeus*	-	-
Primates	Hylobatidae	*Symphalangus syndactylus*	-	-
Primates	Lemuridae	*Lemur catta*	0 817 . CM036499.1 69736667 69962416 + 4303	4
Primates	Cercopithecidae	*Macaca mulatta*	0 817 . CM017758.1 6447027 6189353 - 4339	4
Primates	Lorisidae	*Nycticebus coucang*	0 817 . CM050161.1 4817703 4538810 - 4231	4
Primates	Callitrichidae	*Callithrix jacchus*	0 817 . CM021881.1 4618956 4373116 - 4330	4
Cetacea	Balenopterids	*Balaenoptera acutorostrata*	0 817 . OX465355.1 4170830 3914517 - 4320	4
Artiodactyla	Camelidae	*Camelus dromedarius*	0 817 . CM016663.2 117704300 117930648 + 4153	4
Perissodactyla	Rhinocerotidae	*Diceros bicornis*	0 817 . CM036987.1 4316039 4089381 - 4301	4
Carnivora	Felidae	*Panthera leo*	0 817 . CM031449.1 3732763 3495440 - 4324	4
Carnivora	Ursidae	*Tremarctos ornatus*	0 817 . CM052623.1 105272434 105499128 + 4329	4
Chiroptera	Phyllostomidae	*Desmodusrotundus*	0 817 . CM040287.1 3299952 3142194 - 4217	4
Lagomorpha	Leporidae	*Oryctolagus cuniculus*	-	-
Rodentia	Muridae	*Rattus norvegicus*	39 817 . CM026994.1 66432415 66451558 + 3116	4
Rodentia	Dipodidae	*Jaculus jaculus*	38 817 . CM036412.1 63013759 63033872 + 3134	4
Rodentia	Sciuridae	*Sciurus carolinensis*	0 817 . LR738601.1 3251598 3036991 - 4297	4
Proboscidea	Elephantidae	*Elephas maximus*	0 817 . CM044047.1 174683582 174959252 + 4315	4
Dasyuromorphia	Dasyuridae	*Sarcophilus harrisii*	0 817 . LR735556.1 209940027 210283053 + 4301	4
Diprotodontia	Macropodidae	*Macropus giganteus*	0 817 . CM053633.1 67052383 67364964 + 4296	4
Monotremate	Ornithorhynchidae	*Ornithorhynchus anatinus*	0 817 . CM014214.1 18659645 18550273 - 4274	4

Attachment 5

Table S11. Introns of the genes and its respective sizes

Gene	*AKAP-17A*	*NLGN3*	*NLGN4X*	*VAMP7*	*SPRY3*	*ASMTL*	*IL9R*
Intron 1	Size:1131bp Begin: 255	Size:639bp Begin: 153	Size:122759bp Begin: 152	Size:6010bp Begin: 50	-	Size:10335bp Size: 32	Size:5126bp Begin: 10
Intron 2	Size:3641bp Begin: 304	Size:4609bp Begin: 174	Size:120536bp Begin: 209	Size:2431bp Begin: 68	-	Size:3041bp Size: 76	Size:429bp Begin: 48
Intron 3	Size:1226bp Begin: 385	Size:1677bp Begin: 194	Size:5187bp Begin: 271	Size:249bp Begin: 114	-	Size:3440bp Size: 93	Size:116bp Begin: 85
Intron 4	-	Size:9065bp Begin: 243	Size:9397bp Begin: 543	Size:20020bp Begin: 145	-	Size:610bp Size: 113	Size:564bp Begin: 145
Intron 5	-	Size:2682bp Begin: 305	-	Size:19881bp Begin: 168	-	Size:336bp Size: 134	Size:712bp Begin: 194
Intron 6	-	Size:1453bp Begin: 568	-	Size:2088bp Begin: 198	-	Size:4154bp Size: 170	Size:603bp Begin: 261
Intron 7	-	-	-	-	-	Size:2050bp Size: 300	Size:1331bp Begin: 296
Intron 8	-	-	-	-	-	Size:3635bp Size: 354	Size:2512bp Begin: 324
Intron 9	-	-	-	-	-	Size:2543bp Size: 416	-

Intron 10	-	-	-	-	-	Size:865bp Size: 460	-
Intron 11	-	-	-	-	-	Size:107bp Size: 508	-
Intron 12	-	-	-	-	-	Size:1290bp Size: 549	-

Attachment 6

Table S12. Gene IDs of target species obtained using geneious

Species	Gene ID
Diceros bicornis minor	131422884
Symphalangus syndactylus	129475942
Pan paniscus	129395308
Pongo pygmaeus	129025267
Nycticebus coucang	128560754
Elephas maximus indicus	126087446
Lemur catta	123631382
Panthera leo	122210094
Desmodus rotundus	112298507
Pan troglodytes	112204413
Gorilla gorilla gorilla	109027695
Sarcophilus harrisii	105749532
Camelus dromedarius	105098408
Balaenoptera acutorostrata	103013363
Jaculus jaculus	101595011
Callithrix jacchus	100408034
Oryctolagus cuniculus	100356855
Pan troglodytes	741980
Macaca mulatta	704266
Rattus norvegicus	24500

Attachment 7

Data Analysis in Molecular Biology and Evolution

File Edit Alignment Sequences Seq. Analysis Phylogenetics Structure Bioinfo Genomics Graphics Tools Help

```
Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.
=======================================================
Prop. invar. sites                0,6186
Mean H                            0,7046
Standard Error                    0,0153
Hmax                              1,6173
Iss                               0,4356
Iss.c                             0,8095
T                                24,3733
DF                                745
Prob (Two-tailed)                 0,0000
95% Lower Limit                   0,4055
95% Upper Limit                   0,4657

=======================================================
Part II. For an extreme asymmetrical (and generally very
unlikely) tree.
=======================================================
Iss.c                             0,7020
T                                17,3686
DF                               745
Prob (Two-tailed)                 0,0000

95% Lower Limit                   0,4055
95% Upper Limit                   0,4657

=======================================================

Interpretation of results:
              Significant Difference
              ----------------------
              Yes                 No
-------------------------------------------------------
Iss < Iss.c   Little              Substantial
              saturation          saturation
-------------------------------------------------------
Iss > Iss.c   Useless             Very poor
              sequences           for phylogenetics
-------------------------------------------------------
```

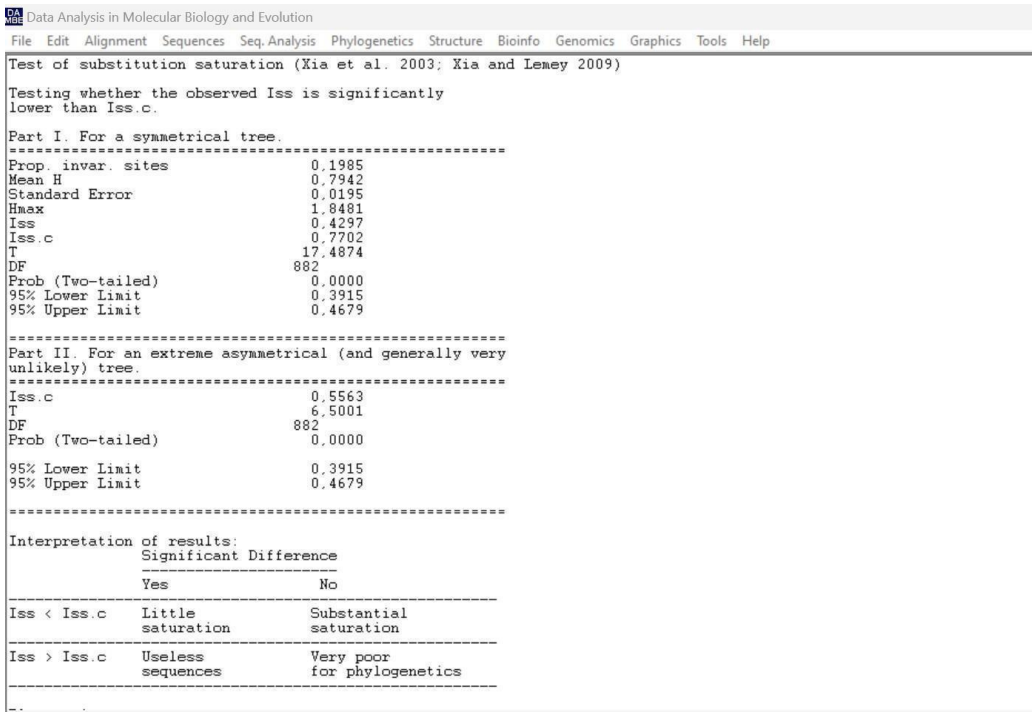
Figure S1. Xia Test Results for *AKAP17A* Gene Based on the DAMBE Analysis

Data Analysis in Molecular Biology and Evolution

File Edit Alignment Sequences Seq. Analysis Phylogenetics Structure Bioinfo Genomics Graphics Tools Help

```
Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.
=======================================================
Prop. invar. sites                0,4453
Mean H                            1,1453
Standard Error                    0,0134
Hmax                              1,6409
Iss                               0,6980
Iss.c                             0,8119
T                                 8,5137
DF                               883
Prob (Two-tailed)                 0,0000
95% Lower Limit                   0,6717
95% Upper Limit                   0,7242

=======================================================
Part II. For an extreme asymmetrical (and generally very
unlikely) tree.
=======================================================
Iss.c                             0,7131
T                                 1,1330
DF                               883
Prob (Two-tailed)                 0,2575

95% Lower Limit                   0,6717
95% Upper Limit                   0,7242

=======================================================

Interpretation of results:
              Significant Difference
              ----------------------
              Yes                 No
-------------------------------------------------------
Iss < Iss.c   Little              Substantial
              saturation          saturation
-------------------------------------------------------
Iss > Iss.c   Useless             Very poor
              sequences           for phylogenetics
-------------------------------------------------------
```

Figure S2. Xia test Results for *ASMTL* gene based on the DAMBE Analysis

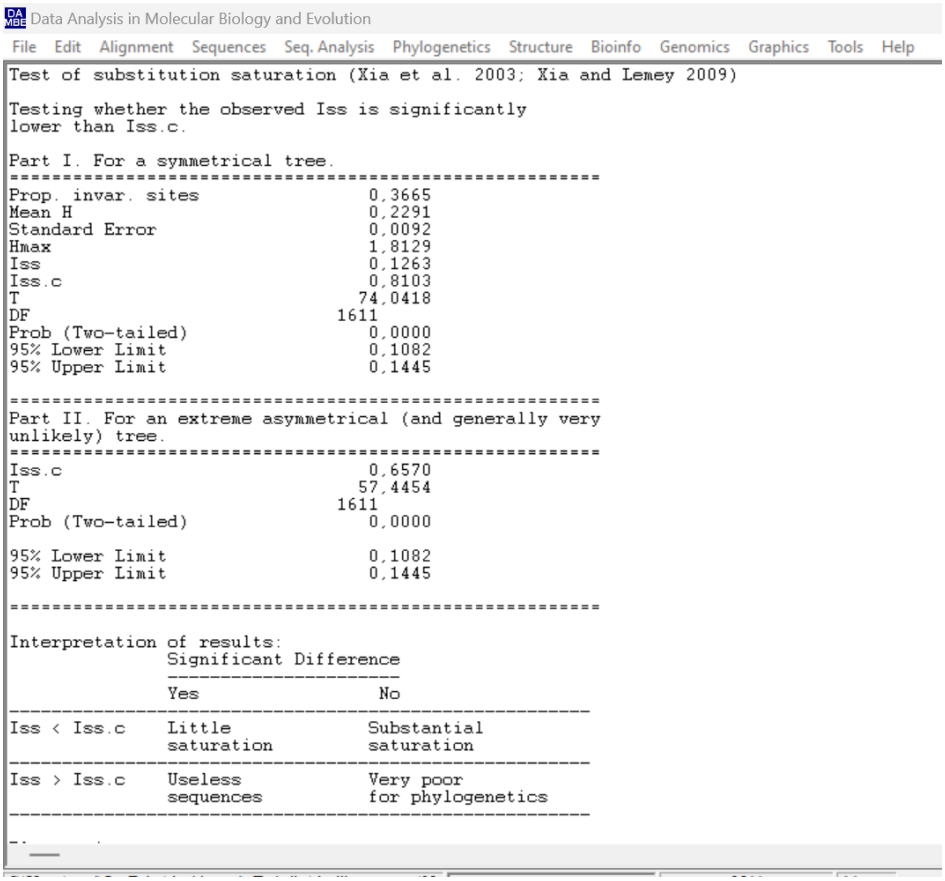


Data Analysis in Molecular Biology and Evolution

File Edit Alignment Sequences Seq. Analysis Phylogenetics Structure Bioinfo Genomics Graphics Tools Help

```
Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.
=====================================================
Prop. invar. sites                0,1985
Mean H                            0,7942
Standard Error                    0,0195
Hmax                              1,8481
Iss                               0,4297
Iss.c                             0,7702
T                                17,4874
DF                               882
Prob (Two-tailed)                 0,0000
95% Lower Limit                   0,3915
95% Upper Limit                   0,4679

=====================================================
Part II. For an extreme asymmetrical (and generally very
unlikely) tree.
=====================================================
Iss.c                             0,5563
T                                 6,5001
DF                               882
Prob (Two-tailed)                 0,0000

95% Lower Limit                   0,3915
95% Upper Limit                   0,4679

=====================================================

Interpretation of results:
                Significant Difference
                ----------------------
                Yes                 No
-----------------------------------------------------
Iss < Iss.c     Little              Substantial
                saturation          saturation
-----------------------------------------------------
Iss > Iss.c     Useless             Very poor
                sequences           for phylogenetics
-----------------------------------------------------
```

Figure S3. Xia test Results for *IL9R* gene based on the DAMBE Analysis

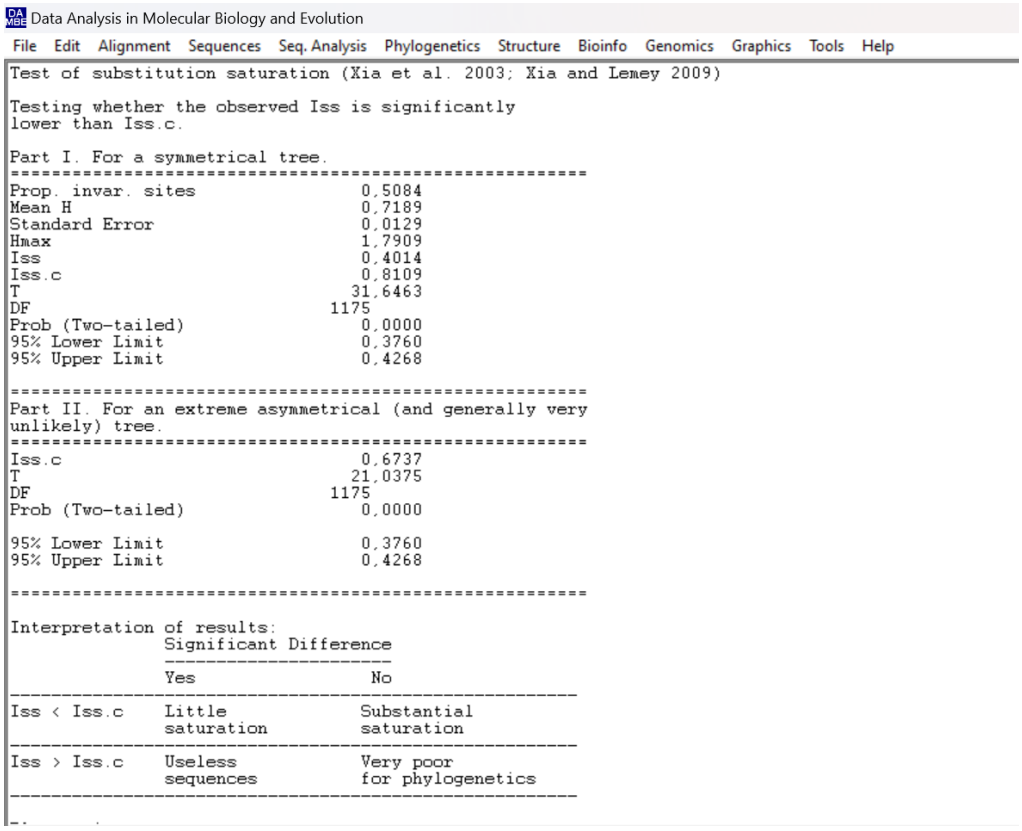


Data Analysis in Molecular Biology and Evolution

File Edit Alignment Sequences Seq. Analysis Phylogenetics Structure Bioinfo Genomics Graphics Tools Help

```
Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.
=====================================================
Prop. invar. sites                0,3665
Mean H                            0,2291
Standard Error                    0,0092
Hmax                              1,8129
Iss                               0,1263
Iss.c                             0,8103
T                                74,0418
DF                              1611
Prob (Two-tailed)                 0,0000
95% Lower Limit                   0,1082
95% Upper Limit                   0,1445

=====================================================
Part II. For an extreme asymmetrical (and generally very
unlikely) tree.
=====================================================
Iss.c                             0,6570
T                                57,4454
DF                              1611
Prob (Two-tailed)                 0,0000

95% Lower Limit                   0,1082
95% Upper Limit                   0,1445

=====================================================

Interpretation of results:
                Significant Difference
                ----------------------
                Yes                 No
-----------------------------------------------------
Iss < Iss.c     Little              Substantial
                saturation          saturation
-----------------------------------------------------
Iss > Iss.c     Useless             Very poor
                sequences           for phylogenetics
-----------------------------------------------------
```

Figure S4. Xia test Results for *NLGN3* gene based on the DAMBE Analysis



Data Analysis in Molecular Biology and Evolution

File Edit Alignment Sequences Seq. Analysis Phylogenetics Structure Bioinfo Genomics Graphics Tools Help

Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.

Prop. invar. sites	0,5084
Mean H	0,7189
Standard Error	0,0129
Hmax	1,7909
Iss	0,4014
Iss.c	0,8109
T	31,6463
DF	1175
Prob (Two-tailed)	0,0000
95% Lower Limit	0,3760
95% Upper Limit	0,4268

Part II. For an extreme asymmetrical (and generally very unlikely) tree.

Iss.c	0,6737
T	21,0375
DF	1175
Prob (Two-tailed)	0,0000
95% Lower Limit	0,3760
95% Upper Limit	0,4268

Interpretation of results:

	Significant Difference	
	Yes	No
Iss < Iss.c	Little saturation	Substantial saturation
Iss > Iss.c	Useless sequences	Very poor for phylogenetics

Figure S5. Xia test Results for *NLGN4X* gene based on the DAMBE Analysis

Data Analysis in Molecular Biology and Evolution

File Edit Alignment Sequences Seq. Analysis Phylogenetics Structure Bioinfo Genomics Graphics Tools Help

```
Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.
=======================================================
Prop. invar. sites                0,4297
Mean H                            0,6876
Standard Error                    0,0153
Hmax                              1,8335
Iss                               0,3750
Iss.c                             0,7543
T                                24,7973
DF                               481
Prob (Two-tailed)                 0,0000
95% Lower Limit                   0,3450
95% Upper Limit                   0,4050

=======================================================
Part II. For an extreme asymmetrical (and generally very
unlikely) tree.
=======================================================
Iss.c                             0,5445
T                                11,0807
DF                               481
Prob (Two-tailed)                 0,0000

95% Lower Limit                   0,3450
95% Upper Limit                   0,4050

=======================================================

Interpretation of results:
                Significant Difference
                ----------------------
                Yes                 No
-------------------------------------------------------
Iss < Iss.c     Little              Substantial
                saturation          saturation
-------------------------------------------------------
Iss > Iss.c     Useless             Very poor
                sequences           for phylogenetics
-------------------------------------------------------
```

Figure S6. Xia test Results for *SPRY3* gene based on the DAMBE Analysis

Data Analysis in Molecular Biology and Evolution

File Edit Alignment Sequences Seq. Analysis Phylogenetics Structure Bioinfo Genomics Graphics Tools Help

```
Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.
=======================================================
Prop. invar. sites                0,3310
Mean H                            0,2700
Standard Error                    0,0171
Hmax                              1,7981
Iss                               0,1501
Iss.c                             0,7387
T                                34,3663
DF                               441
Prob (Two-tailed)                 0,0000
95% Lower Limit                   0,1165
95% Upper Limit                   0,1838

=======================================================
Part II. For an extreme asymmetrical (and generally very
unlikely) tree.
=======================================================
Iss.c                             0,5612
T                                24,0035
DF                               441
Prob (Two-tailed)                 0,0000

95% Lower Limit                   0,1165
95% Upper Limit                   0,1838

=======================================================

Interpretation of results:
                Significant Difference
                ----------------------
                Yes                 No
-------------------------------------------------------
Iss < Iss.c     Little              Substantial
                saturation          saturation
-------------------------------------------------------
Iss > Iss.c     Useless             Very poor
                sequences           for phylogenetics
-------------------------------------------------------
```

Figure S7. Xia test results for *VAMP7* gene based on the DAMBE Analysis