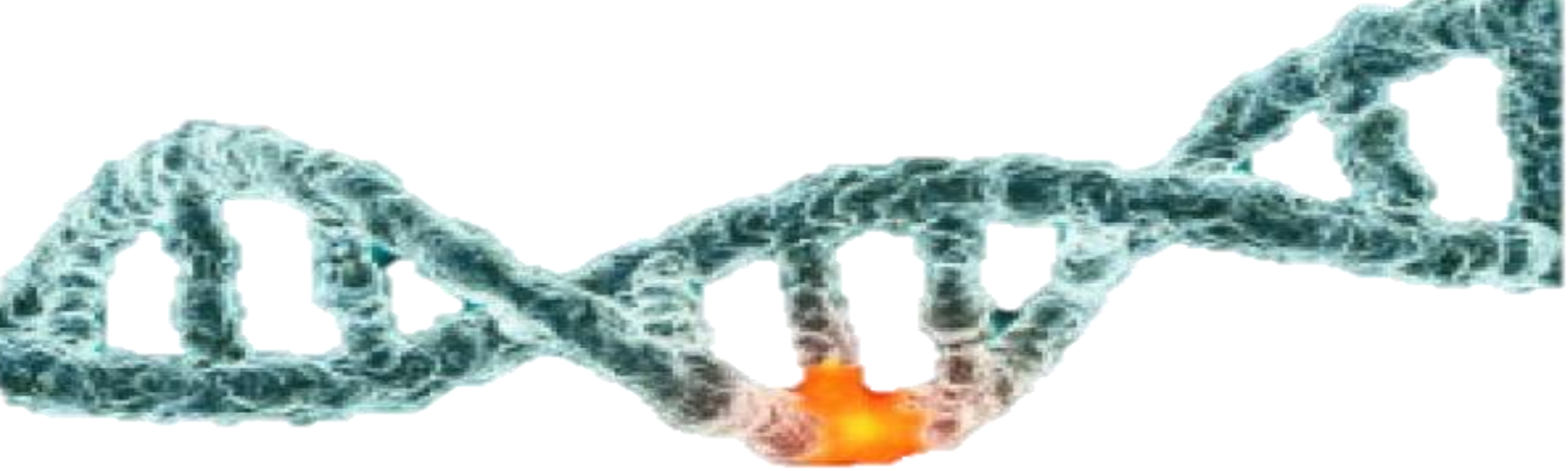


Identification and analysis of concurrent multiple nucleotide substitutions

Catarina Tavares Serrano

Mestrado em Genética Forense

Departamento de Biologia

Faculdade de Ciências da Universidade do Porto

2015

Orientadora / Supervisor:

Doutora Luísa Azevedo

IPATIMUP

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.
O Presidente do Júri,

Porto, ______/______/_________

“Continuous effort - not strength or intelligence - is the key to unlocking our potential.”

Winston Churchill

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Dr. Luisa Azevedo for her extraordinary support in my research and in writing this thesis. I thank her for sharing her knowledge, for helping me to achieve my goals and for always encouraging me to stay motivated. I could not have imagined a better advisor for my master thesis.

Besides my advisor, I would like to thank Dr. João Carneiro for his patience while introducing me to the complex world of programming, for his availability and for his positive comments, which helped me improve. I would also like to thank Professor António Amorim for his guidance and support. My sincere thanks also goes to Dr. David Cooper for the interest demonstrated in my work and all the comments and suggestions made.

My family was very important in supporting me emotionally and mentally when I got insupportable and needed time. I thank them for encouraging me and believe in me more than myself. A special thanks goes to Pedro for the interesting talks about statistic and scientific questions in general, I promise that I will deep into statistic.

Last but not the least, I would like to thank Levi Santos for his dedication, love and for being by my side through good and bad times. Thank you for being proud and not let me fall. I believe in you like you believe in me.

Abstract

The mutational process is nonrandom. Distinct mutations in a given genomic segment can be the result of the same mutational event, rather than being the successive accumulation of mutations through generations, these are known as concurrent mutations. Two mechanisms are able to explain the origin of multiple mutations: the independent generation and successive accumulation of Single Nucleotide Substitution (SNS) and the occurrence of different mutations in a concurrent manner, concurrent Multiple Nucleotide Substitution (MNS).

This thesis focuses on the identification of concurrent MNS from polymorphic variation in five human genes (*BRCA1, LDLR, LRRK2, PSEN1, RET*). Our analyses resulted in a total of 43 clusters with representative examples of distinct concurrent MNS (Tandem Base Mutations, Complex MNS and Noncontiguous MNS). Patterns of double, triple and more than three mutations were detected and analyzed. Other studies were performed such as what concerns the population specificity and the existence of a preference in the type of nucleotide substitution.The major outcome was the observation of a significant difference in the transition/transversion rate ($R_{Ts/Tv}$) between TBS and Noncontiguous MNS, which is independent of distance between mutations.

In the Forensic Genetics field the estimation of the mutation rate is of critical importance. Thus, the simultaneous birth of mutations may impose some bias in current estimations. To date, concurrent mutations were only detected in the mutational spectra of human disease-causing genes. Our work, demonstrate the occurrence of concurrent mutations in the human polymorphic variation, which is of irrefutable relevance in upcoming estimations of the human mutation rate.

Key-words: concurrent Multiple Nucleotide Substitutions, mutation rate, Tandem Base Mutation, Complex MNS, and Noncontiguous MNS.

Resumo

O processo mutacional não é aleatório. Mutações múltiplas numa determinada região genómica podem ter sido geradas no mesmo evento mutacional, ao invés de terem resultado da acumulação sucessiva de mutações ao longo do tempo, sendo estas designadas por mutações *concurrent*. Existem dois mecanismos que são capazes de explicar a origem das mutações *concurrent*: a origem independente e acumulação sucessiva de *Single Nucleotide Substitution* (SNS) e a ocorrência de mutações múltiplas em simultâneo, *concurrent Multiple Nucleotide Substitution* (MNS).

A atenção foi focada na identificação de *concurrent* MNS presentes em cinco genes humanos (*BRCA1, LDLR, LRRK2, PSEN1, RET*). A análise resultou num total de 43 *clusters* constituídos por diferentes *concurrent* MNS (*Tandem Base Mutations,* MNS complexos e MNS não contíguos). Foram detetados padrões em diferentes tipos de MNS: MNS duplos, triplos e MNS com mais de três mutações. Outras análises foram efetuadas, tais como a especificidade populacional das mutações em estudo e a verificação da existência da preferência no tipo de substituição nucleotídica. O principal resultado decorrente desta tese foi a observação de diferenças significativas na taxa de transição/transversão ($R_{Ts/Tv}$) entre *Tandem Base Substitution* e MNS não contíguos, o que se mostrou ser independente da distância entre as mutações.

Na Genética Forense a correta estimação da taxa de mutação é de grande importância. O facto de existirem mutações que ocorrem em simultâneo faz com que existam desvios relativamente às taxas de mutação atualmente existentes. Até à data, mutações *concurrent* foram apenas detetadas em estudos centrados em mutações diretamente relacionadas com doenças. O presente trabalho demonstra a existência de mutações *concurrent* em polimorfismos humanos, o que é de relevância irrefutável em futuras estimativas da taxa de mutação do genome humano..

Palavras-chave: concurrent Multiple Nucleotide Substitutions, taxa de mutação, Tandem Base Mutation, MNS complexos e MNS não contíguos.

Table of contents

Figures Index

Tables Index

Abbreviations list

A – Adenine

bp – Base pair

C – Cytosine

CpG – Cytosine-phospho-guanine

CSMM – Closely spaced multiple mutations

DNA – Deoxyribonucleic acid

DSB – Double strand break

G – Guanine

INDEL – Insertion/deletion

kb – Kilobases

MJD – Machado-Joseph disease

MNS – Multiple Nucleotide Substitution

PCNA – Proliferating cell nuclear antigen

SNP – Single nucleotide polymorphism

SNS – Single nucleotide substitution

SRS – Serial replication slippage

T – Thymine

TBS – Tandem base substitution

TLS – Translesion

Ts – Transition

Tv – Transversion

UV – ultraviolet radiation

Introduction

The concept of mutation: a very brief history

In the 1900's, the evolutionary biologist Hugo de Vries used, for the first time, the word mutation to indicate the sudden appearance of phenotypic variation (reviewed in [1]). His finding was based in the work with *Oenothera lamarckiana.* At the time, De Vries observed that when a plant was self-pollinated, some offspring were different from the progenitors. When F1 was self-pollinated, some plants were still distinct from the parents. The plants that passed the mutation to the progeny as an inheritable characteristic were considered a new species. To De Vries new species were not formed by continuous variation as Darwin has defended, but by the sudden appearance of variation.

In 1904 De Vries stated, "*Natural selection may explain the survival of the fittest, but it cannot explain the arrival of the fittest*" [2]. To De Vries and his contemporaneous colleagues the causes of mutation were not known, but in an attempt to explain it, he suggested the existence of "pangenes" which, when mutated, could immediately create a new species [3].

Nowadays the concept of mutation is different: a mutation is not the observed phenotypic difference between parents and offspring, it is a change in the DNA sequence. This change can result from distinct mechanisms and may or may not led to phenotypic differences between individuals carrying distinct alleles.

Types of mutations on the DNA sequence

In theory, a mutation can occur with similar probabilities in exonic, intronic or even regulatory regions. When a nucleotide is changed in the DNA sequence we are dealing with a Point Mutation. These type of mutations can be further divided into transitions (when the replacement involves a purine-to-purine or a pyrimidine-to-pyrimidine change), and transversions (substitution of a purine to a pyrimidine or vice versa). When these mutations occur within exonic regions, they may alter the amino acid resulting in an incorrect protein sequence and these are known as missense mutations. The impact of the missense mutations is strongly dependent on the target protein where it occurs, that is, its degree of essentiality, the location on the protein, and on the physical and chemical similarity between the newly arisen amino acid and the original one.

Substitutions that replace an amino acid with another with similar properties are conservative substitutions. It is expected that conservative replacements lead to less dramatic changes on protein function than non-conservative changes. Whenever the mutation introduces a premature STOP codon, the resulting protein is expectedly shorter (truncated protein) and, often, non-functional. Whenever these mutations involve genes involved in critical biological processes, serious implications are expected at the phenotypic level, and consequently in health [4].

There are mutations that result from insertions (IN) and/or deletions (DEL) of one or few bases, they are called INDELS. As expected, the reading frame is altered and an altered protein is produced. This kind of mutation may provoke diseases, an example is the genetic disorder: cystic fibrosis [5]. This disorder is mostly caused by an INDEL in *CFTR* gene, which eliminates a single amino acid.

Duplication is a type of mutation characterized by a piece of DNA that is copied and is present in the genome more than one time. The duplication may occur at the entire gene or in just a part of it. Duplications of oncogenes are related with a number of cancers, for example, a high prevalence of cases with the oncogene *KRAS* duplicated was found in patients with colorectal cancer [6].

Another type of mutations can occur at repetitive stretches of DNA. Mutations that cause repeat expansions are, for instance, causative of the Machado-Joseph Disease (MJD’s), an late-onset autosomal dominant neurodegenerative disorder [7].

The exon 10 of the *ATXN3* gene, responsible for encode ataxin-3 in various tissues, has an unstable CAG repeat that when expanded may cause MJD.

The human mutation rate

Mutation is the ultimate source of variation and, hence, is responsible for evolution and genetic diseases. The estimation of the rate of *de novo* mutations and their properties is an important subject in genetics, having implications in the understanding of the evolutionary process and the molecular basis of human diseases.

The first attempts to estimate the human mutation rate were based on the observation of certain phenotypes or on the direct comparison of homologous sequences among closely related species [8-13]. J. B. S. Haldane was a pioneer in this field [10, 13]. He assumed the importance of accounting the rate at which a new mutation appears and the selective force to eliminate mutations that confers a reduction in fitness to estimate the human mutation rate. In the 1930s and 1940s, Haldane used the frequency of hemophiliac men to calculate the frequency of the disease-causing allele. A novel mutation causative of hemophilia was estimated to occur at a rate of approximately 10^{-5} per generation [10, 13].

Another approach used the direct counts of affected offspring with a dominant disease or a recessive X-linked disease born from healthy parents [11]. Vogel and Motulsky summarized a list of diseases that fit these patterns of inheritance and the respective mutation rate was estimated as to vary from 10^{-6} to 10^{-4} per generation [14]. In passing, it should be mentioned that a critical failure of the methods that focus on the observation of phenotypes to calculate a mutation rate is the obvious underestimation of the true mutation rate of the gene since some mutations may have only slightly effect on phenotypes which may not be of clinical attention or even having no effect at all (neutral mutations).

There are indirect methodologies to estimate the mutation rate based in the assumption that the mutation rate can be estimated through the rate of fixation of neutral mutations in a species [12]. Therefore, stretches of noncoding DNA (which is assumed to evolve neutrally) of two closely related species, such as chimpanzee and humans, are compared. Since the generation time and divergence from the last common ancestor of the two species are known, the mutation rate can be estimated. Using this approach, the human mutation rate was estimated to be 10^{-8} per generation [9, 15].

In 2003, A. Kondrashov and collaborators estimated the mutation rate at 20 loci associated with Mendelian diseases [8]. The estimated mutation rate, taking into account all types of mutations, resulted similar to previous studies: 10^{-8}, per generation.

The work of A. Kondrashov included also an important dimension to the preceding estimates: the analyses of mutation rate by different types of mutation. The mutation rate of INDELS, for example, has been reported as 10^{-9} per site per generation, which corresponds to a lower value than the obtained for single nucleotide substitutions (SNP). It was also demonstrated that mutational hotspots such as the CpG dinucleotides have high mutation rates compared to non-CpG dinucleotides, an issue that is going to be explored in more detail below.

With the next-generation sequencing and the possibility of study more genomic information at once emerged genome-wide studies that have more accuracy, are more direct and have more comprehensive tactics (reviewed in [16]). Whole-genome estimates based on pedigrees resulted in a human mutation rate of roughly 10^{-8} per base pair per generation.

The human mutation rate may vary with innumerous factors. In this regard, Kong and collaborators demonstrated the influence of the father’s age in the rate of *de novo* mutations, where they found that the mutation rate was directed related with the father’s age at the time of conception [17]. On other hand, F.D. Conrad and collaborators observed that in one family the paternal germline was responsible for the most part of the mutation, but in another family it was the maternal germline. This demonstrate variation of the mutation rate between human pedigrees [18].

Source of Mutation

For the purpose of this thesis some factors related with the heterogeneity of the mutational process will be explored in more detail. These are: (a) the DNA polymerases, (b) Non-B DNA conformations and (c) CpG dinucleotides.

(a) DNA polymerases

DNA polymerases are involved in all processes involved in DNA synthesis. Mutations in DNA polymerases or changes in their expression have a critical role in mutagenesis [19]. The humans have more than a dozen of different DNA polymerases, which only five of them are well understood and well stated as essential (Pol α, Pol β, Pol γ, Pol ε and Pol δ) (reviewed in [19]). The key function of these five polymerases is summarized in Table 1.

Tab.1- Main function of the classical DNA polymerases.

Type	Function	Reference
Pol α	Initiation of DNA replication	[20]
Pol β	Base-excision repair	[21]
Pol γ	Replication in mitochondrial DNA	[22]
Pol ε	Synthesis of the leading strand of nuclear DNA	[23]
Pol δ	Lagging strand synthesis and exonuclease activity	[24]

The aforementioned DNA polymerases are interrupted when DNA is damaged, but in order to bypass the innumerous mutations that happen in a cell there are DNA polymerases that are capable of get through precedent lesions: Translesion synthesis (TLS) DNA polymerases (reviewed in [25]). There are irrefutable differences between the classical DNA polymerases and TLS DNA polymerases: the first have higher replication taxes, higher processivity and higher fidelity (reviewed in [25, 26]).

When DNA is damaged the replication process is blocked and, in order to continue, the switch of the DNA polymerase by a TLS polymerase is required. The non-occurrence of this switch implies the cell dead. The switch implies the recruitment of the Proliferating Cell Nuclear Antigen (PCNA), a protein that binds DNA polymerase and prevents the dissociation with the template DNA. The ubiquitination of PCNA allows interactions between the PCNA and TLS DNA polymerases [27, 28].

(b) Non-B DNA Structure

Since 1953, the DNA molecule is known to have a right-handed double helical structure with Watson–Crick base pairing, B-DNA conformation. However, some DNA regions enriched in repetitive sequences can adopt other forms known as non-B DNA conformations. There are about ten types of non-B DNA conformations, namely cruciform/harpin, triplex, slipped, tetraplex and left-handed Z-form (Table 2). By their nature, some sequences are more likely to form non-B DNA conformations but they are mainly formed when certain conditions result in negative supercoiling, encourage the separation of the DNA and histones and/or the separation of the two strands in processes such as the replication and transcription [29].

The cruciform and hairpin conformations require an inverted sequence repeat of at least seven nucleotides. When both DNA strands are involved, the conformation is known as a cruciform, whereas when just one strand is involved it creates a hairpin [30]. The inverted repeats are complementary and pair with each other forming a structure with a hairpin and a loop arm. The energy needed for the formation of the aforementioned structures is believed to be provided by the negative supercoiling [31].

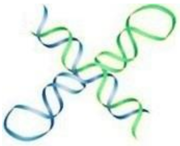
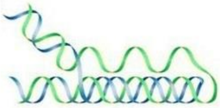
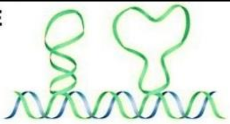
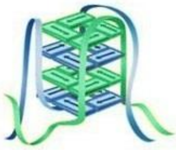
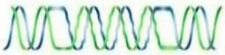
Several homopurine:homopyrimidine sequences with mirror symmetry have the potential to form triplex DNA, this structure is characterized by the wind of a third strand with the usual duplexes [32].

Slipped strand DNA conformations are formed when complementary strands with simple repeats are misaligned, forming a slipped structure with hairpins or lopped-out bases. This kind of structures were only described in short repeats, but they may arise in every repeat [33].

Sequences enriched in guanines are able to form a special conformation: a tetraplex [34]. In a tetraplex structure four guanines are bonded and form a square co-planar of guanines.

When an alternating purine-pyrimidine sequence occurs, the double helix of the DNA can wind to the left in a zig-zag configuration, forming a non-B DNA named Z-DNA [35].

Tab.2- Summary of the sequence requirements by the most common non-canonical DNA structures.

Name	Conformation	Sequence Requirements	Example
Cruciforme/Hairpin		Inverted repeats	TCGGTACCGA AGCCATGGCT
Triplex		Mirror repeat symmetries in polypurine/polypirimidine regions	AAGAGG GGAGAA TTCTCC CCTCTT
Slipped strand DNA		Misalignment of simple repeats	TCGGTTCGGT AGCCAAGCCA
Tetraplex		Sequence rich in Guanine	$AG_3(T_2AG_3)_3$ single strand
Z-DNA		Sequence with alternating pyrimidine and purine	CGCGTGCGTGTG GCGCACGCACAC

It has been shown that non-B DNA conformations have influence in some important biological processes such as chromatin architecture, gene expression regulation, DNA replication, DNA damage and repair and genetic instability [29, 36-38].

Various studies demonstrate the influence of the occurrence of non-B DNA on promoting the genetic instability, increasing the number of mutations as large deletions or even single nucleotide substitution [38]. It has been showed this secondary structures being related with the increase of genetic anomalies such as deletions, rearrangements and chromosomal translocations (reviewed in [29]). ,A previous study analyzed the impact of introducing a sequence capable of form non-B DNA in a transgenic mice and the results showed that approximately 7% of the transgenic mice were carriers of large deletions and chromosomal translocation [39]. Some literature reported the association between the formations of non-B DNA structures at breakpoints of rearrangements and inherited diseases as, for instances, mental retardation, idiopathic pancreatitis and ornithine transcarbamylase deficiency [40-42].

The occurrence of genetic instability by the formation of non-canonical DNA structures can result from two mechanisms: replication-independent mechanism of non-B DNA-induced genetic instability and replication-dependent mechanism of non-B DNA-induced genetic instability.

Some intrinsic characteristics of the replication process facilitate the emergence of non-B DNA conformation and genetic instability. The replication process itself may promote the transition for secondary structures by separating the two strands of DNA and providing the energy necessary to the occurrence of negative supercoiling [29]. An well, the replication process can impact the mutagenesis of the pre-existing non-B DNA structure. During replication the non-B DNA is likely to influence the fidelity of DNA polymerases and inducing replication fork stalling which prolongs the time where exists single strand DNA and prompt genetic instability [43, 44]. If the replication fork stalling is longer than the supported by the organism it may cause the cut off of both the strands of DNA (Double Strand Breaks-DSB) [29]. It was also observed that simple repeats occurring at non-B DNA conformations can promote the DNA polymerase slippage which results in the expansion or contraction of the number of repeats [45].

Apart from the replication process, other biological processes can enable the formation of non-B DNA such are the transcription, damaged DNA and the mechanism of DNA repair itself. When DNA is damaged it may alters the DNA conformations, recruit proteins that interact with DNA or modify the chromatin proteins which can facilitate the formation of this structures [29]. The transition of a B to a non-B conformation can be stimulated by repairing mechanisms [29]. For example, the repair of an DBS in a short inverted repeat lead to an hairpin structure at the extremity of the break [46]. During transcription the two DNA strands are separated and the process generates negative supercoiling inducing the formation of non-B DNA [29]. Studies in the promoter region of the *c-MYC* gene revealed that the transcription is responsible for the formation of a non-B secondary structure [47].

Replication-independent mechanisms like the aforementioned are able to result in genetic instability in non-B DNA structures [29]. Occasionally the non-canonic DNA structure can be confused with damaged DNA due to the chromatin distortions and repair mechanisms are activated which may lead to the cleavage of the DNA near the region of non-B DNA [29]. Non-B DNA conformations also my alters the exposure of the DNA to damaging agents or/and affect the accessibility of repair proteins [29].

(c) CpG dinucleotide

The CpG dinucleotide is a mutational hotspot. The first evidence of the hypermutability of the CpG dinucleotide was provided by Youssoufian et al. [48]. During the attempt to understand the molecular basis of hemophilia they observed the existence of recurrent mutations in the *F7* gene, resulting from a CG→TG transition. This observation suggested the possibility that CpG dinucleotides could be a mutational hotspot. The validation arose from the study of Cooper and Youssoufian [49] in a number of disease-causing mutations, confirming the hypermutability of the CpG dinucleotide and the capacity to cause human genetic disease.

In the human genome, unmethylated cytosine suffers spontaneous deamination to uracil. Because uracil is not a DNA base it is removed by the uracil-DNA glycosylase with high fidelity and the mutation is frequently repaired. However, when methylated cytosine deaminate to thymidine the repair machinery is less efficient because both are DNA bases (reviewed in [38]). It is important to mention that CpG transitions can occur due to mechanisms other than the more common spontaneous deamination of the 5mC [50]. The methylation of CpGs is pointed to be influenced by non-B DNA conformations. A previous study demonstrated that the non-canonical DNA structures quadruplex had significantly low methylation of CpGs, revealing a inverse correlation between the number of CpGs in a sequence and the probability to form non-B structures [51].

Mutational Clusters

To estimate the mutation rate, the mutation process is assumed to be random and mutations to arise in an independent fashion. However, the existence of more mutants with multiple mutations than expected by chance alone showed that this is not always the case: *“although mutations are poorly predictable, they have been observed to be nonrandom”* [52].

Most of the work in this field has been done using the Big Blue transgenic mice as a model [53, 54]. Assuming that the mutation were independent it was expected a frequency of doubles of 5,3x10^{-10}, but, the observed frequency was much larger: 3,5x10^{-7} [53]. In a similar work, Colgin and his collaborators while studying the *HPRT* gene noticed that mutations in a human epithetical cell line had four of their 12 mutations separated by only six bases, this mutations were more closely spaced than the predicted by a random distribution [55].

When multiple mutations occurs in a particular space region of the genome (Closely Spaced Multiple Mutations, CSMM) they are defined as mutational clusters [56, 57]. A mutational cluster can be explain by two mechanisms: (a) the successive accumulation of independent single nucleotide substitution (SNS) during evolution, or, (b) by a unique mutational event that originates different mutations in the same cell cycle, conducting to Concurrent Multiple Nucleotide Substitution (MNS) [56] (Figure 1).

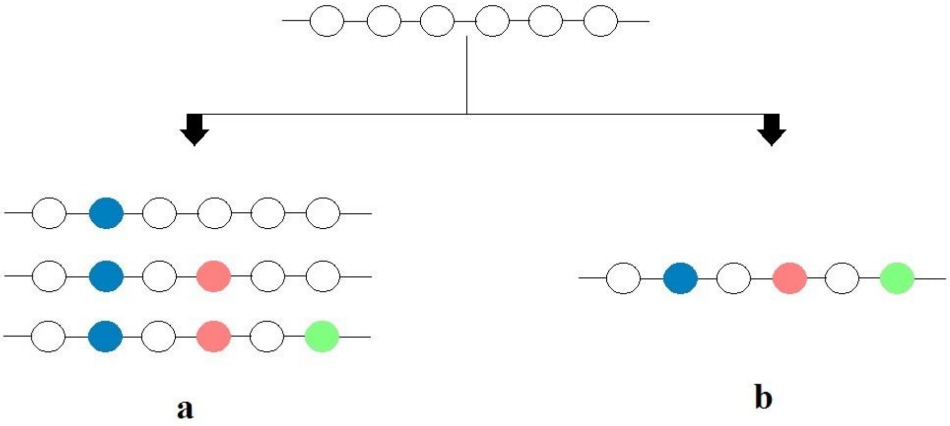


Fig.1- Birth of multiple mutations. **a)** This example shows three independent mutational events that creates a mutation cluster, so this cluster mutation does not arise in the same cell cycle. **b)** Multiple mutations occurring at the same mutational event.

Clustered mutations are not confined to small DNA fragments. In fact, the mutations may be within few bases or even have several kilobases separating them, giving rise to a mutation shower [52, 58]. A previous study add a limit of 30 Kb as the maximum distance of mutations in a shower [58] (Figure 2).

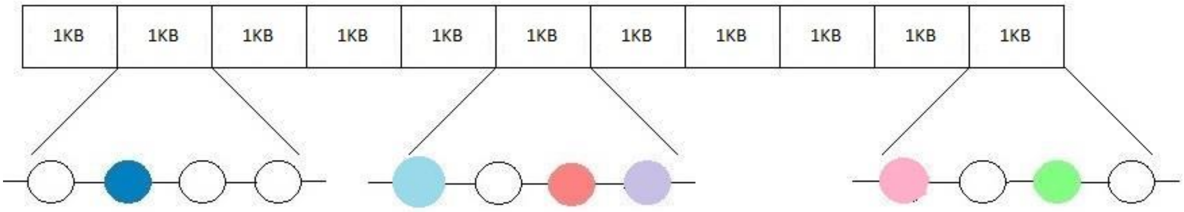


Fig.2- Visualization of a mutation shower where colored circles represent each mutation of a cluster.

Concurrent MNS are able to create compensatory mutations, where mutations that separately where deleterious together are advantage or neutral, and simultaneously generate pathogenic mutations. Like Zhu and his collaborators stated: *"concurrent MNS mutations represent two sides of the same coin"* [59]. Related evidence showed that the majority of the compensatory mutations is caused by mutational showers [52, 60].

There are three types of MNS that differ in the distance between different mutations: **(a)** Tandem Base Substitution (TBS), **(b)** noncontiguous MNS and **(c)** complex MNS (Figure 3) [59].

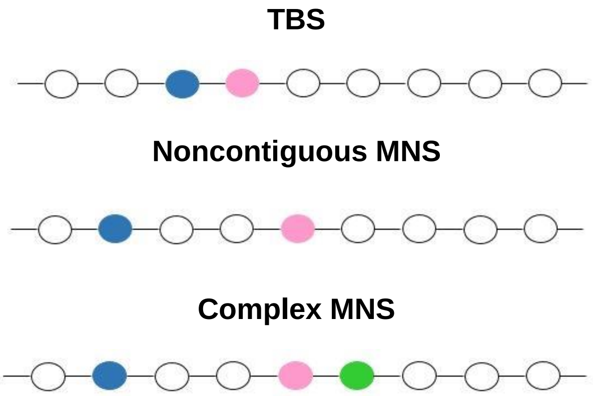


Fig.3- Classification of MNS into three types: Tandem Base Substitution, Noncontigous MSN and Complex MNS.

The origin of concurrent MNS results mainly from two mechanisms: gene conversion and transient hipermutability [57].

Gene conversion involves the transference of a portion of DNA from a donor sequence to an acceptor sequence with a minimum of 90% of homology [61] and it is a molecular mechanism used to repair two-ended DSBs, stabilizing chromosomal mispairing [62]. Gene conversion events tend to occur in regions enriched in CpG and regions that form stable hairpin structures [63].

Transient hypermutability results from processes that implies the newly synthesis of DNA and the misincorporation of bases during DNA replication and repair: (a) problems in the replicative DNA polymerase and proteins implicated in the replication fidelity, (b) the disruption of the balance of the nucleotide pool or (c) the recruitment of an error-prone DNA polymerase [58, 61, 64, 65]. Cases of transient hypermutability were describe in somatic and germline cells [56]. The proportion of CpG is used a crude indicator to discriminate between MNS generated by transient hypermutability or just by the gradual accumulation of mutation in a mutagenic region, so the lower the CpG content the higher is the likelihood that the MNS have occurred by transient hypermutability [56] [57]. An association exists between the enrichment in transversions and concurrent MNS formed via transient hypermutability. Previous evidence has shown that the transitions/transversions ratio (Ts/Tv) is lower in concurrent MNS than in SNS [59].

As shown in a previous section non-canonical DNA conformations have a great impact in biological mechanism including the replication process and the repair itself. Because these mechanisms are required to have a high fidelity to avoid mutation, when they are compromised by non-B DNA tend to cause instability which may promote the mutagenesis of the affected region. Repetitive segments with high mutability rates are expected to be more prone to generate concurrent MNS [38, 56].

DNA polymerases can bypass mutations and are able to replicate damaged DNA at low speed, low processivity and low fidelity through the switch of a classic DNA polymerase to a TLS DNA polymerases [19, 25, 26, 66]. There are studies that point out the role of TLS DNA polymerase in generating concurrent MNS [56, 67-69]. Chen and collaborators emphases the importance of this error-prone DNA polymerases in generating TBS and hypothesized that the major part of the double and triple TBS arise from errors introduced by TLS polymerases [67]. Most of the triples resulting from incorporation of incorrect bases during bypass of DNA lesions by the TLS polymerase.

The mechanism upon which TLS polymerase originates TBS is different for TBS with equal or less than three bases and those of more than three bases. In the first case the contiguous mutations are caused by the bypass of endogenous DNA lesions (incorporation of incorrect bases against the damage bases) and the second one by serial replication slippage (SRS).

TBS are originated mostly by endogenous mechanisms and not by DNA lesions introduced by the exposure to exogenous agents like the UV radiation [67]. Concurrently generated quadruplet to octuplet TBS are generally related with some signatures: the perfect or almost perfect reverse complements (*in situ* inversion) or the perfect or almost perfect matches near upstream sequences (neighboring sequence duplication) [67]. The same study showed the existence of signatures of the combined action of SRS and TLS polymerase in the birth of quadruplet to octuplet TBS. Most of them have microhomologies near the region with the MNS, consistent with the recruitment of the TLS polymerase during the SRS. In accordance, the action of the TLS polymerase in generating complex MNS was recently documented [70].

The mutation rate estimations are based on the assumption that mutations arise independently and randomly, which is now clear not to be the case. Thus, the non-independence of mutations has a major role in the estimation of the human mutation rate because the presence of concurrent MNS mutations not considered in the estimation causes the overestimation of the human mutation rate. Now, it is important to recognize the specificity of the DNA regions and the occurrence of simultaneous mutations to account in futures mutation rate, in order to achieve a more accurate estimation of the mutation rate. Some work is this field has been done, for example, the mutational rate of double and triple TBS was estimated to be 0,2-1,2x10^{-10} and 0,8-4,8x10^{-12} per base per generation, respectively. The mutation rate of double TBS was estimated to be 0,2-1% of the single-base substitution [67].

Cancer genetics is complex and may involve multiple mutations where the most likely scenario is the simultaneous appearance of them, although only a few examples are available in related literature. An example is provided by Chen *et al* [71] showed the oncogene *EGFR* and the tumor suppressor *TP53*, genes related with lung cancer, had some doubles and multiples mutations with a pattern that suggest the simultaneous occurrence of the observed mutations during the same mutational event.

Concurrent mutations are not exclusive in cancer related genes, they can occur in the entire genome. An example is found in mammalian mitochondrial OXPHOS proteins, more specifically in the COI protein [72]. In the COI protein the Leu195-Leu196 is the wild-type in humans and his very conserved among mammals. The deleterious Leu196Ile mutation is associated with an epileptic phenotype in *humans* but in Fat Dormouse (*Myoxus glis*) is the wild-type allele and was predicted to be tolerated due to the compensatory effect of a neighboring residue (Ile195). The simultaneous occurrence of a compensatory mutation allowed the fixation of Ile196 in the rodent lineage.

Aims

This work was based on the analyses of SNPs in five genes related with a disease phenotype. The primary effort was to understand better the occurrence and the characteristics of concurrent mutations, a topic which was not extensively studied previously. The main goals were:

- Evaluation the non-random distribution of the mutations by statistic methods
- Identification of concurrent mutations
- Investigate the possible mechanisms of formation of concurrent mutations
- Identifying patterns in the different concurrent MNS
- Establish general rules related to sequence specificities that may result in concurrent MNS.

Material and Methods

Data mining

Genotypic data was retrieved from phased 3 of the integrated variant call set of 1000 Genomes Browser based on Ensembl v76 GRCh37 which contains genotypes from 2504 individuals across 26 human populations [73]. Here, we focused on five genes related with common diseases: the low density lipoprotein receptor (*LDLR*, chr19:11200038-11244492) that regulates the level of cholesterol in the blood [74], the tumor suppressor gene breast cancer 1 (*BRCA1*, chr17:41196312-41277500) [75-78], the presenilin 1 (*PSEN1*, chr14:73603126-73690399) which is associated with an early-onset form of Alzheimer disease [79-83], the ret proto-oncogene (*RET*, 10:43572475-43625799), a transmembrane receptor tyrosine kinase involved in a large number of diseases [84], and the leucine-rich repeat kinase 2 (*LRRK2*, chr12:40590546-40763087) that encodes the Dardarin protein which has been associated with a late-onset form of Parkinson disease [85-88]. Further explanation about the genes addressed here are present in the Appendix I.

For each of the five genomic regions that encompass each of the five genes, the corresponding VCF file was retrieved and only single nucleotide polymorphisms were maintained for further analyses, that is, no INDELS were considered in this study. This strategy resulted in a total of 11419 SNPs distributed as follows: 1416 in *LDLR*, 1957 in *BRCA1*, 2095 in *PSEN1*, 1569 in *RET* and 4382 in *LRRK2*.

Data analyses

Each SNP dataset was initially analyzed using a sliding window of 100 nucleotides and steps of 30 nucleotides in order to assess the genomic regions that contain more polymorphisms than it would be expected by chance. The neutral expectation is that the mutational process occurs randomly, therefore, the expected distribution follows a Poisson's distribution where the expected value of the number of mutations per site equals the average value. The comparison between the average and the expected value were performed with a Chi-square test using $\alpha = 0.05$.

The average value of the number of mutations per 100 bases was proximally 2.8. We used this value to search each gene-specific SNP dataset in order to select the intervals where the number of mutations differs of two mutations from the immediately flanking interval (See Supplementary File I). This approach resulted in a total of 726

Clusters (*RET*: 112, *PSEN1*: 114, *LRRK2*: 269, *LDLR*: 94, *BRCA1*: 137), which were numerated for each gene in CL1, CL2 and pursuing this logic.

The frequency of each variation was obtained from the data of 1000 Genomes Project and analyzed using python scripts (Python v.2.7.6.0, www.python.org/). Our interest was particularly the concurrent MSN, so it was established some criteria to improve the chance of finding the MNS generated simultaneously: in the previous clusters SNPs with an overall frequency of less than 10% (accepting a difference of 10% to account possible errors) were chosen to ensure that mutations are recent events. After the filtering process, manual inspection of the clusters was performed to verify if the SNPs appear in the same individuals in order to increase the possibility of the haplotype that defines each cluster being generated by concurrent mutations (Supplementary File II).

Results and discussion

According to the neutral expectation, if mutations accumulate randomly and independently across the genome, it is expected their distribution to follow a Poisson distribution. We started to use all polymorphic sites to evaluate the distribution of polymorphic sites obtained from the 1000 Genomes Project for each of the five genes here studied, and obtained for an $\alpha = 0.05$, all five genes showed a significant *p* value (*LDLR* $p = 2.21171 \text{x} 10^{-26}$; *PSEN1* $p = 1.86026 \text{x} 10^{-12}$; *BRCA1* $p = 3.13311 \text{x} 10^{-07}$; *RET* $p = 0.000337$; *LRRK2* $p = 4.2233 \text{x} 10^{-08}$).This result was in accordance with a non-random distribution of mutations across these genes.

We next tried to detect mutations that appear to be linked on the same haplotype but not observed alone in a distinct haplotype, which could indicate their birth by the same mutational event. Polymorphism data extracted from the 1000 Genomes Project for the five genes here studied (*BRCA1, LDLR, LRRK2, PSEN1, RET*) were used to define the clusters (haplotypic backgrounds defined by putative concurrent mutations). This approach resulted in a total of 43 clusters: 1 cluster for *BRCA1*, 13 clusters for *LDLR*, 14 clusters for *LRRK2*, 7 clusters for *PSEN*1 and 8 clusters for *RET*. The 43 clusters comprises a total of 103 distinct mutations. Among all these mutations, 15.5% are localized at CpGs sites.

Concerning to the type two complex MNS (BRCA1-CL1 and RET-CL7) and 13 TBS (*LDLR*: CL1, CL4, CL11; *LRRK2*: CL1, CL2, CL6, CL12, CL13, CL15; *PSEN1*: CL5; *RET*: CL2, CL3, CL4, CL5) were detected. The remaining concurrent MNS found were noncontiguous MNS (BRCA1:CL1; RET: CL7) (Table 3).

Tab.3- Mutation clusters detected through the analysis of standing variation based on SNPs from the 1000 Genomes Project. The number of mutations per cluster is indicated as so their location in each gene. The type of concurrent MNS was defined according to the distance between mutations.

Gene	Cluster	Nº SNP's	CpGs	Intron or/and Exon	Type of concurrent MNS
BRCA1	1	3	0	Intron 19	Complex MNS
LDLR	1	2	0	Intron 1	TBS
	2	2	2/2	Intron 2	Noncontigous MNS
	3	2	0	Intron 2	Noncontigous MNS
	4	4	0	Intron 3	TBS
	5	2	1/2	Intron 3	Noncontigous MNS
	6	2	0	Exon 4	Noncontigous MNS
	7	2	0	Intron 5 and Exon 6	Noncontigous MNS
	8	2	1/2	Intron 6	Noncontigous MNS
	9	2	1/2	Intron 7	Noncontigous MNS
	10	2	0	Intron 8	Noncontigous MNS
	11	2	0	Intron 9	TBS
	12	2	0	Intron 9	Noncontigous MNS
	13	2	0	Intron 15	Noncontigous MNS
LRRK2	1	2	0	Intron 1	TBS
	2	2	0	Intron 1	TBS
	3	3	0	Intron 1	Noncontigous MNS
	4	3	0	Intron 7	Noncontigous MNS
	5	2	0	Intron 9	Noncontigous MNS
	6	2	0	Intron 16	TBS
	7	2	0	Intron 20	Noncontigous MNS
	8	2	0	Intron 20	Noncontigous MNS
	9	2	0	Intron 24	Noncontigous MNS
	10	2	0	Intron 44	Noncontigous MNS
	11	4	2/2	Intron 45 and 49	Noncontigous MNS
	12	2	0	Intron 46	TBS
	13	2	0	Intron 49	TBS
	14	2	0	Exon 54	Noncontigous MNS
PSEN1	1	2	2/2	Intron 4	Noncontigous MNS
	2	4	1/4	Intron 4 and Exon 9	Noncontigous MNS
	3	4	1/4	Intron 7 and 15	Noncontigous MNS
	4	2	1/2	Intron 7	Noncontigous MNS
	5	2	0	Intron 13	TBS
	6	3	0	Intron 14	Noncontigous MNS
	7	6	3/6	Intron 15, 16 and 18	Noncontigous MNS
RET	1	2	1/2	Intron 1	Noncontigous MNS
	2	2	0	Intron 1	TBS
	3	2	0	Intron 1	TBS
	4	2	0	Intron 1	TBS
	5	2	0	Intron 1	TBS
	6	2	0	Intron 3	Noncontigous MNS
	7	3	0	Intron 16	Complex MNS
	8	2	0	Exon 20	Noncontigous MNS

A previous study focused in concurrent double TBS indicates the GC dinucleotide as the most frequent mutated dinucleotide (28.2%) and the GC>AA/TT as the most frequent TBS alteration [67], with a frequency of 7.92% and 9.56%, respectively. Our research is not motivated by concurrent TBS in particular, but by concurrent MNS in general. Therefore, our dataset only presents 13 TBS concurrent MNS, wherein only one is not a double TBS (LDLR-CL4) (Table 4). As shown in Table 4, two double TBS have the GC dinucleotide as the wild-type allele (16.7%), which is lower than the value obtained previously (28.2%) [67], although it is likely to result from the bias introduced by the number of double TBS of our study. The dinucleotide change GC>TT was observed in PSEN-CL5 (8.33%), which is consistent with the conclusions of a previous studies where the authors conclude that this replacement is frequently associated with GC dinucleotides [67].

An interesting and previously unreported observation could be made by a close look into the Table 4: there are two most common mutant alleles: AT (n=4) and TT (n=3), together they account for 58% of all the concurrent double TBS. This may suggest some kind of preference or even a signature of the mechanism underlying the generation of concurrent double TBS.

Tab.4- Distribution of the 13 concurrent TBS

TBS	Gene Cluster	Replacement
1	LDLR-CL1	TC>CA
2	LDLR-CL4	TGCC>ACTG
3	LDLR-CL11	TG>CA
4	LRRK2-CL1	GC>AT
5	LRRK2-CL2	TC>AT
6	LRRK2-CL6	GA>TT
7	LRRK2-CL12	CC>AT
8	LRRK2-CL13	CA>AT
9	PSEN-CL5	GC>TT
10	RET-CL2	GA>TG
11	RET-CL3	TG>GA
12	RET-CL4	GG>TT
13	RET-CL5	GA>TG

The mutation rate and the type of mutation is dependent of the sequence context [38]. A very interesting study about the relation of dependency between the mutation rate and the nucleotides flanking the mutated allele in *Pseudomonas fluorescens* ATCC948 was made by H. Long [89]. This paper demonstrated a higher mutation rate for T bases with a G at 3' and particularly an increase of T mutated when it was flanked by a G at 5' and 3'. In line with this, our data show that of the total of mutations that occur at a T bases 62.5% (n=10) have a G at the 3' flanking base.

As shown in Figure 4, the clusters of mutations were grouped as doubles, when two polymorphic positions are contiguous, triple, whenever three mutations contribute to define the cluster and >3, when more than 3 mutations define the cluster. To a better understanding of how mutations at clusters with three or more than three mutations are distributed along the sequence the distance between the mutations is also shown in this figure.

Pairs of putative concurrent mutations (doubles) were observed in all genes with the exception of *BRCA1* for which a single cluster with three mutations were defined. Double mutations were the most frequent category observed among the set of genes here studied and the distance between the mutations at each pair that define the cluster is highly variable. For instance, in the case of *LDLR*, 12 clusters of double mutations were defined according to the criteria mentioned in the methodology section. The distance between the mutations at those clusters is very heterogeneous with only two clusters showing contiguous mutations and the remaining with mutations spaced by longer distances. Similar patterns of the distribution of double mutations were also observed at the *LRRK2*, *PSEN1* and *RET* genes.

Clusters of three mutations were observed at the *BRCA1*, *LRRK2*, *PSEN1* and *RET*. In these cases, it was possible to establish three patterns for the co-occurrence of three mutations based on their distance on the sequence: (1) two contiguous mutations and a third at a very short distance (one or two nucleotides) as is the case observed at the *BRCA1* and *RET* and (2) three mutations separated by one or few nucleotides. Clusters of triple mutations are rare events and only five of these clusters could be observed in total.

Rare are also clusters defined by more than three mutations. We observed five haplotypes defined by more than three mutations, one at the *LRRK2* and *LDLR* and three at the *PSEN1* gene. The cluster at the *LDLR* is defined by a set of four contiguous mutations, a pattern that clearly differs from that observed at the *LRRK2* and *PSEN1*.Taken together, the four clusters with more than three mutations at the *LRRK2* and *PSEN1* revealed an interesting distribution: pairs of closely linked mutations are separated by larger distances. This pattern is in accordance with the observations made for clusters of double mutations where the distance between any two mutations is never higher than 150 (*PSEN1* CL4).

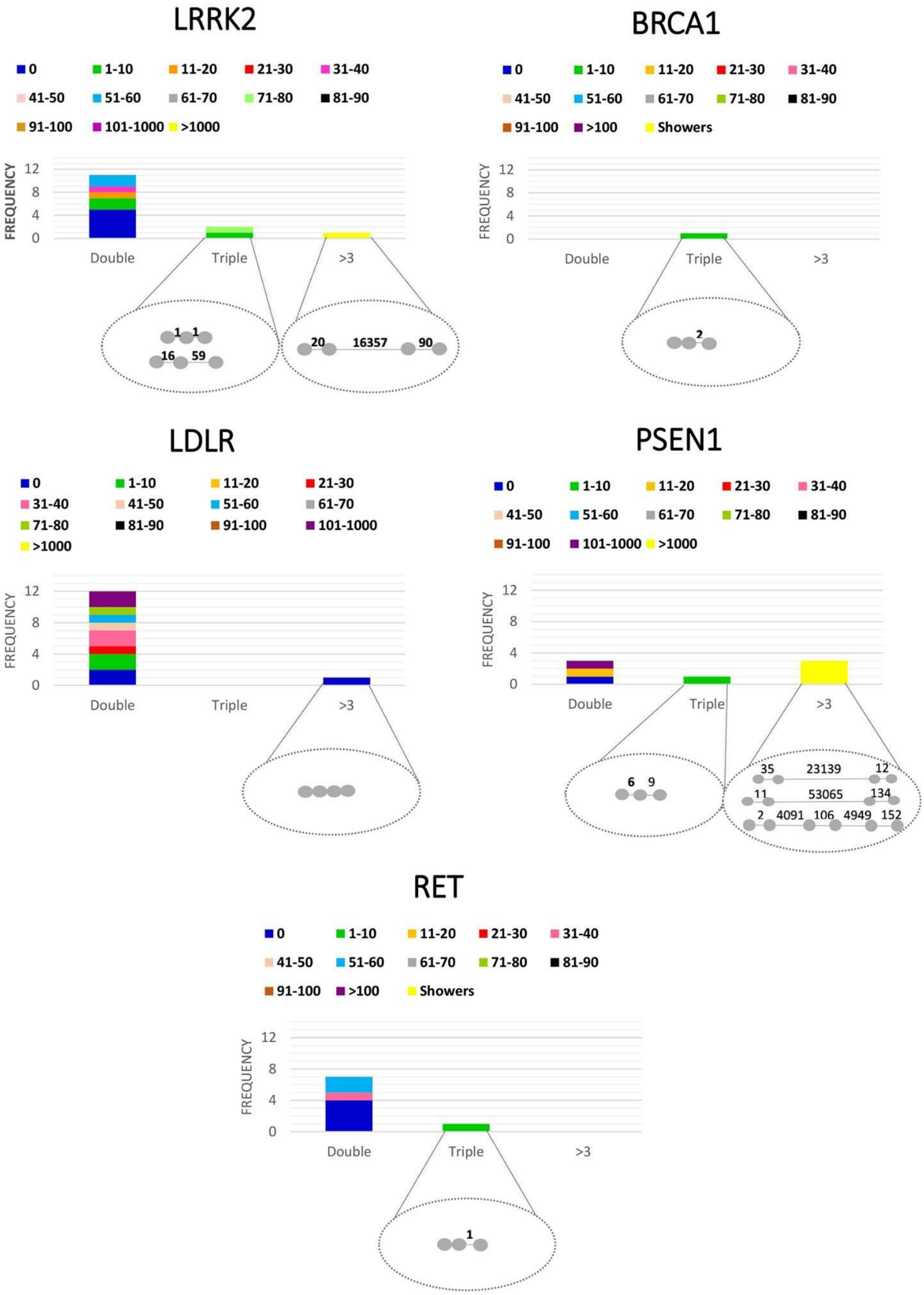


Fig.4- Cluster distribution according to the distance (in nucleotides) between concurrent MNS. For clusters with 3 (triple) or more mutations we use the average value of the distance between all mutations was used to include the cluster into one of the classes.

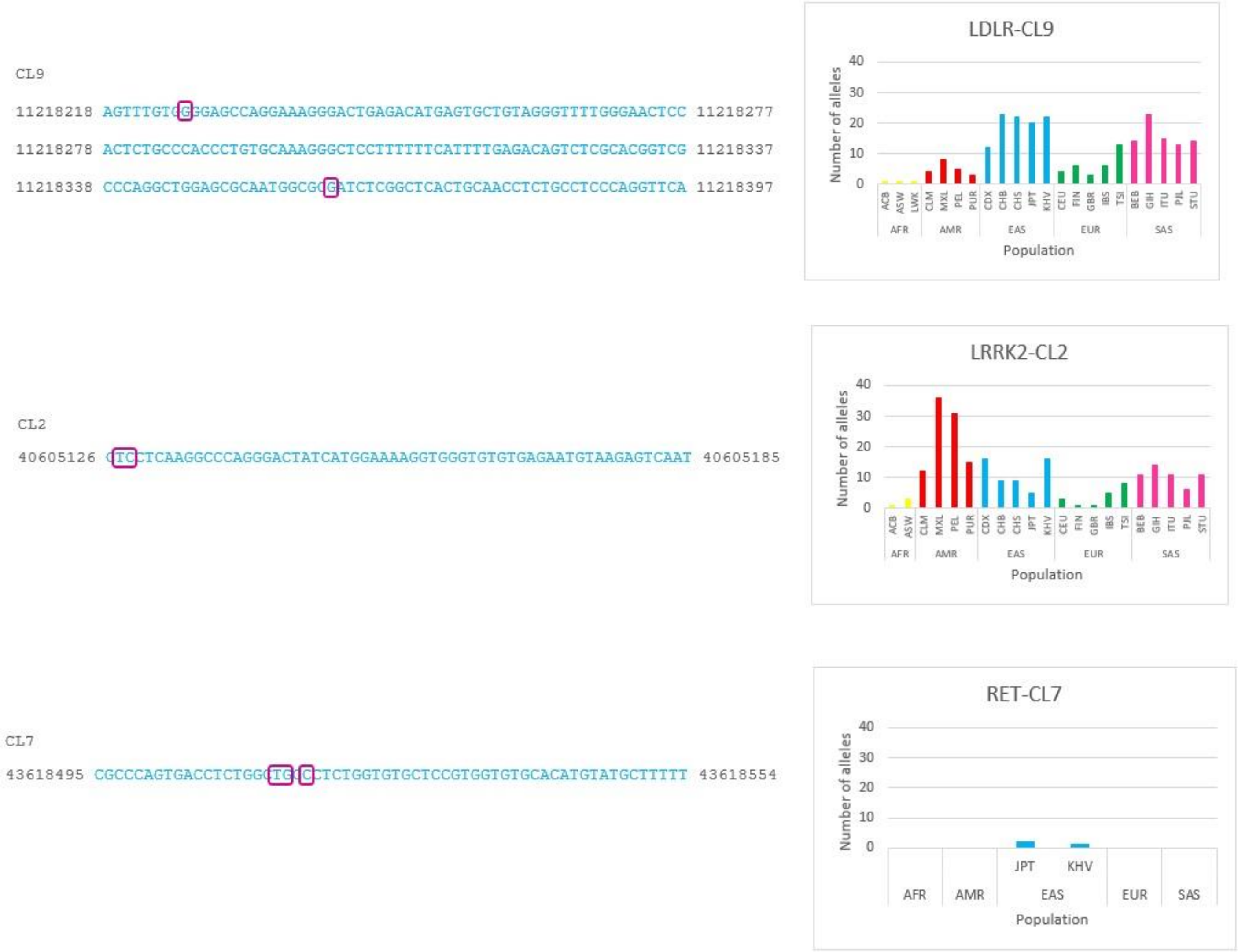
We next analysed the population distribution of concurrent. mutations. The most frequent clusters are distributed across all human populations covered by the 1000

Genomes (AFR, AMR, EAS, EUR and SAS): LDLR-CL9, LRRK2-CL2, LRRK2-CL8, PSEN-CL3, RET-CL3 (Appendix II). For these five common clusters an intesting observation can be made: there is a low frequency of the concurrent MNS in the African population. The CL9 from the gene *LDLR* shows two concurrent MNS. This is an example that demonstrates high general frequency concurrent MNS in humans, although showing to be rare in Africa (Figure 5).

In some of the less frequent clusters the opossite is verified, where the concurrent MNS only apperas in the African population (BRCA1-CL1, LDLR-CL1, LDLR-CL2, LDLR-CL5, LDLR-CL6, LDLR-CL8, LRRK2-CL1, LRRK2-CL4, RET-CL4, RET-CL6, RET-CL8) (Appendix II).

TBS are the most well studied cases of concurrent MNS and the underlying mechanism known. Using data from the human pathologic variation (disease-associated mutations), some studies claim that a key role of TLS polymerases as responsible for most of the TBS [67]. It is important to clarify that TBS distribution ahead from the pathologic mutational spectrum was acessed. In this study, we used the polymorphic variation present in the entire sequence of five genes to demostrate the occurrence and frequency of TBS across human populations (Figure 5).

The complex MNS (mutations that have noncontigous and contigous mutations simultaneously) are the less explored type of concurrent MNS, perhaps because they are uncommon among pathologic data. In our study we report two examples of complex MNS: BRCA1-CLA and RET-CL7. The two complex MNS have a low frequency and are confined to the African and East Asian populations, respectively (Figure 5).



CL9

11218218 AGTTTGTGGGGAGCCAGGAAAGGGACTGAGACATGAGTGCTGTAGGGTTTTGGGAACTCC 11218277

11218278 ACTCTGCCCACCCTGTGCAAAGGGCTCCTTTTTTCATTTTGAGACAGTCTCGCACGGTCG 11218337

11218338 CCCAGGCTGGAGCGCAATGGCGCGATCTCGGCTCACTGCAACCTCTGCCTCCCAGGTTCA 11218397

CL2

40605126 CTCCTCAAGGCCCAGGGACTATCATGGAAAAGGTGGGTGTGTGAGAATGTAAGAGTCAAT 40605185

CL7

43618495 CGCCCAGTGACCTCTGGCTGCCCTCTGGTGTGCTCCGTGGTGTGCACATGTATGCTTTTT 43618554

Fig.5- DNA sequences from LDLR-CL9, LRRK2-CL2 and RET-CL7 and their respective distribution in human populations. The LDLR-CL9 represents a double noncontiguous MNS, the LRRK2-CL2 exemplifies a double TBS and the RET-CL7 is an example of a triple complex MNS.

Once established that co-occurring mutations are not random we ask next whether or not a signature of preferred types of nucleotide substitutions may exist. Figure 6 shows the frequency of replacements from a total of 103 mutations that define the 43 clusters here analyzed. About 53% of all the concurrent mutations are transitions (A→G, T→C, C→T and G→A). There is evident the existence of two more common alterations: C→T and G→A (33% of all the concurrent MNS). The C→T transition is about 64% of the mutations that happen in the nucleotide C and the G→A accounts for 50% of all the mutations in the G nucleotide. Half of the C→T transitions and about 40% of the G→A transitions are located at the mutational hotspot of CpG dinucleotides representing the deamination of cytosine to thymine and guanine to adenine in the other strand [38].

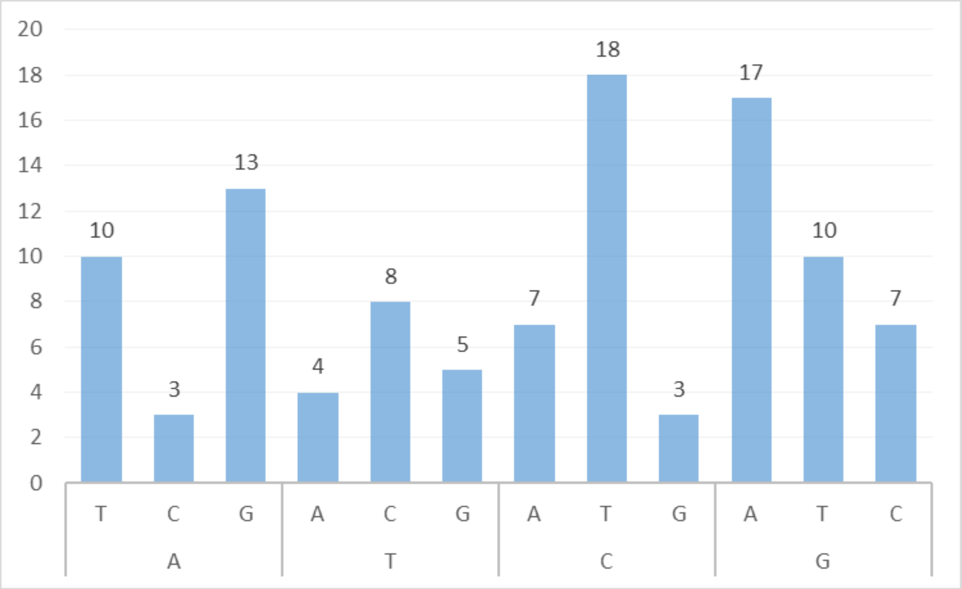


Fig.6- Counting of the type of base change. In the bottom there is the wild-type mutation and above the alternative base.

Although the data from genome-wide de novo SNS mutations is large (N=4933), our dataset for concurrent MNS is a small sample size (N=103; Table 3).

In genome-wide studies of *de novo* SNS mutations, the $R_{Ts/Tv}$ was estimated as to be 2.1 (3344/1589). The number of transitions in concurrent MNS was still higher than the transversions. However, the $R_{Ts/Tv}$ was significantly lower to concurrent MNS (1.10) when compared to SNS (2.10), (p = 0.0009) (Table 5). Our data provides an example that reinforce the study made by Zhu et al. [59], where they concluded a decreased $R_{Ts/Tv}$ in concurrent MNS.

To evaluate the influence of the type of concurrent MNS in the transition/transversion ratio we used the number of transitions and transversions in TBS mutations (N= 28, Table 3) and Noncontigous mutations (N=69, Table 3). Complex MNS were not included because of the small sample size (N=2, Table 3). In TBS mutations more transversions (N=16) than transitions (N=12) were observed leading to a $R_{Ts/Tv}$ <1 (0.75), whereas Noncontigous mutations reveal an enrichment in transitions to transversions ($R_{Ts/Tv}$=1.76). The differences between the $R_{Ts/Tv}$ of TBS and noncontiguous MNS showed to be significant (p = 0.0486) (Table 5).

Recently, Zhu and collaborators [59] demonstrated that the higher $R_{Ts/Tv}$ values are related with a higher distance between concurrent mutations. Our dataset embraces most of the concurrent MNS in a closely spaced region (≤100 nucleotides, N=79) which is in concordance with the decrease of concurrent MNS with the increase of their distance [56].

For concurrent MNS with a maximum distance of 100 nucleotides the $R_{Ts/Tv}$ value is 1.03 (40/39), which indicates a similar content of transitions and transversions. For concurrent MNS with compound mutations separated by more than 100 nucleotides the $R_{Ts/Tv}$ value (1.4) shows an increment in transitions over transversions (14/10). The Fisher's exact test shows no significant differences between the two categories (*p* = 0.3352) (Table 5).

Given the aforementioned results we can corroborate the results of Zhou et al. [56] when they affirm a decreased $R_{Ts/Tv}$ in concurrent MNS over SNS. However, we are able to conclude by our dataset that according to the type of concurrent MNS there are an increase of transition in the case of the noncontiguous MNS and an increase of transversions in TBS. The same does not apply to the distance of the component mutations, where it was not shown a significant difference of $R_{Ts/Tv}$ values in the two different classes. The genome is characterized by a bias in transition mutations [17], but we denote the increment in transversion in the TBS witch is consistent with the signature of TLS polymerase [67]. As stated before, some studies demonstrate that mutagenesis caused by TLS polymerases is common in TBS mutations [67].

Tab.5- Comparison between the $R_{Ts/Tv}$ values for 133 Concurrent MNS and 4933 Genome-wide de novo SNS mutations, 28 TBS and 69 Noncontigous MNS and 79 concurrent MNS with a separation ≤100 nucleotides and 24 concurrent MNS separated by more than 100 nucleotides. The $R_{Ts/Tv}$ for genome-wide de novo SNS mutations was provided by [59]. A Fisher's exact test was used to obtain the level of significance between different classes.

	N	$R_{Ts/Tv}$	*P* value
Genome-wide de novo SNS mutations	4933	2.1 (3344/1589)	0.0009
Concurrent MNS (All)	103	1.1 (54/49)	
TBS	28	0.75 (12/16)	0.0486
Noncontigous MNS	69	1.76 (44/25)	
≤100 nucleotides	79	1.03 (40/39)	0.3352
>100 nucleotides	24	1.4 (14/10)	

Conclusions

The main aim of this work was to detect concurrent MNS in five genes (used as case studies). We were able not only to demonstrate that concurrent mutations are present among the polymorphism spectra of the human genome but also that not all the mutations are independent and some arise simultaneously from the same mutational event. In passing, it shall be noted that this is the first study that identifies concurrent mutations from non-pathologic data. We also demonstrated that the mechanism of origin of different SNS is also different. In this regard, TBS mutations show a clear signature of being originated by TLS polymerases.

In forensic genetics the mutation process itself can lead into misinterpretation of the results. Therefore, a good estimation of a mutation rate is very useful to forensic genetics. The thematic of concurrent MNS is of major importance for forensic applications because the occurrence of simultaneous mutations prevents an accurate estimation of the human mutation rate. The mutation rate is an important reference for forensic genetics. Nowadays it is more robust to calculate a regional mutational rate instead of a general mutation rate that is assumed to be equal to every region, every nucleotide and every circumstance, although it starts to be obvious that this is a simplification of the mutational process.

Future perspectives

During the study of concurrent MNS I was aware of the existence of non-B DNA conformations that may confer some genetic instability to some genetic regions. This genetic instability sometimes conduce to very mutable regions or sites, potentiating sometimes the emerging of simultaneous and dependent multiple mutations. As an attempt to visualize the influence of such biological mechanism it was performed the prediction of the non-B DNA conformation for our sequences. The results of the stable and probable to exist non-B DNA conformations are in Appendix III and an example is provided by Figure 7, where it is visible that the three complex MNS are inside the loop of the non-B DNA conformation, which may explain their birth. In my point of view it would be extremely interesting to do further studies in the specific implications of this non-canonical DNA structures in the promotion of phenomenon that lead to the origin of concurrent MNS.

As well, it is important to expand the pilot study here presented to the entire genome in order to establish general rules that may be important to have in consideration in further estimations of the human mutation rate.

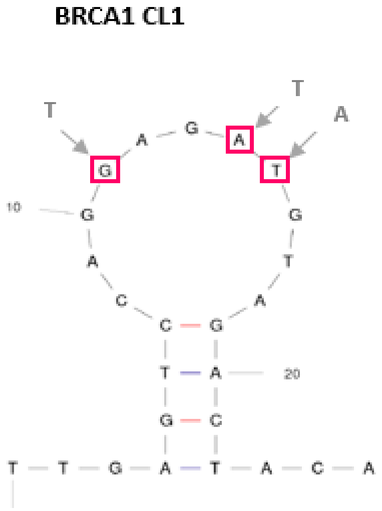

BRCA1 CL1

Fig.7- Example of a prediction of the non-B DNA structure for BRCA1 gene CL1 with a concurrent MNS in the loop. The software UNAFold 3.8 [90] was used to predict the non-B structure of each sequence of the previous selected clusters, being the sequence ten bases after the first mutation of the cluster and ten bases before the last one. The program uses principals of energy minimization, some algorithms and stochastic sampling to predict the folding of sequence of DNA. The number of folding were reduced for two and then it was used the default parameters. In the cases where there was more than one predicted structure it was chosen the one with the lowest free energy (ΔG) because a lower value of ΔG implies a more stable structure.

Appendix

Supplementary Files (provided in electronic format):

Supplementary File I: Selection of clusters by analyze of flanking regions with a difference of 2 mutations.

Supplementary File II: Compilation of the concurrent mutations with the respective allelic frequency and identification of individuals carrying the mutations.

Appendix I: Detailed description of the studied genes

BRCA1

BRCA1 (Breast cancer 1, early onset) is a tumor suppressor gene and some variants potentiate the risk of having breast, ovarian and prostatic cancers and they are linked to a hereditability predisposition to this types of cancers [75-78] . *BRCA1* has an important ro*le in the DNA repair during cellular processes such as cell divisions [91-93].* BRCA1 is also involved in another cellular process: the transcriptional regulation. The interference of its transcriptional activity can lead to tumorigenesis [94].

RET 1

RET (RET proto-oncogene) is a oncogene that encodes a transmembrane receptor tyrosine kinase which is responsible for signaling within cells and it is involved in various cancers and diseases [84]. This signaling process is extremely important to the development of some types of nerve cells, such as the enteric neurons [95]. Studies revealed *RET* as a gene with influence in the normal development of the kidney [95, 96].

LDLR

The expression of low density lipoprotein receptor is carried out by the *LDLR* (low density lipoprotein receptor) gene. The receptor binds two low density lipoprotein that are responsible for transport the cholesterol in the blood and they regulate the extent of cholesterol in the blood [74]. There are variants in the *LDLR* gene causing a disease named familial hypercholesterolemia [97-100]. Mutations in the *LDLR* gene can reduce the number of receptors in cells and/or they may reduce the receptor's capability to remove low-density lipoproteins form the blood resulting in an excess of cholesterol and a higher risk of cardiac problems [74].

PSEN1

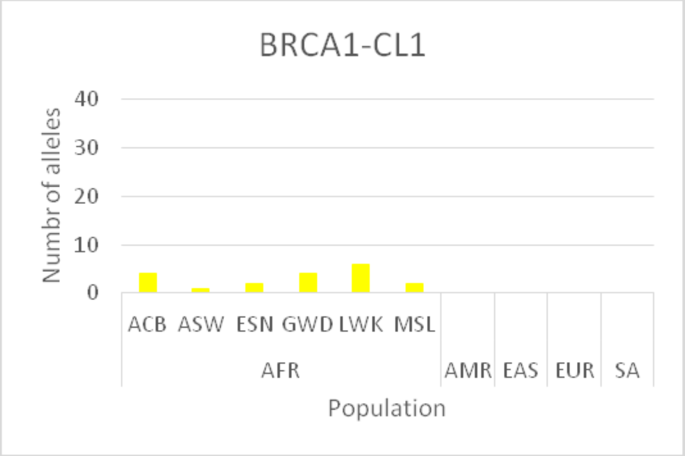
PSEN1 (Presenilin 1) encodes a protein called Presenilin which is involved in proteolytic processes by making part of the complex γ-secretase localized in the membrane, this protein important to cleave proteins and transmit signals from outside of the cell to inside [101]. The best known function of Presenilin is its implication in the metabolism of a protein with impact in the formation of nerve cells [102, 103]. Innumerous variants in *PSEN1* are known to be responsible for early-onset Alzheimer disease [79-83]. Hidradenitis suppurativa, a skin disease, can be caused by one deletion in this gene (725delC) [104].

LRRK2

LRRK2 (Leucine-rich repeat kinase 2) express the Dardarin protein. The previous mentioned protein has various functions such as interact with another proteins, in order to transmit signals or to help cytoskeleton reorganization [105]. Dardarin has another known function: kinase activity, critical for numerous cell activity [106]. Several mutations in this gene are responsible for Parkinson disease later-onset or increase the risk of get the disease [85-88].

Appendix II: Sequence of the genes BRCA1, LDLR, LRRK2, PSEN1 and RET with the annotation of the concurrent MNS and the respective population distribution.

BRCA1



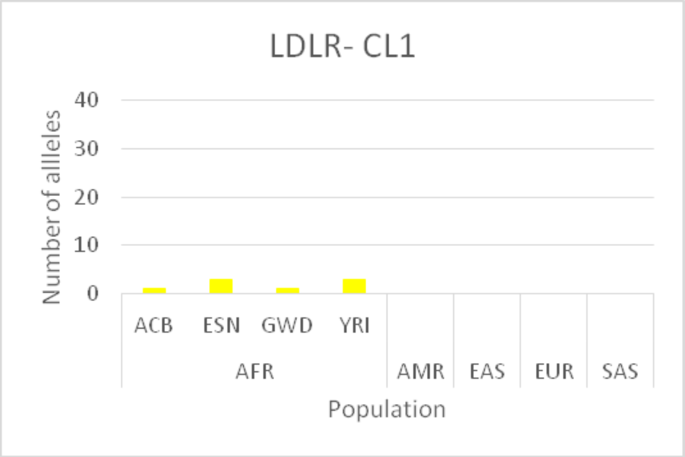
CL1

41254760 GAGATGTAGACTACAGTGAGCTATGATCACTCCACTGCACTTCAGCGTGGGCGGCAAAGC 41254701

LDLR

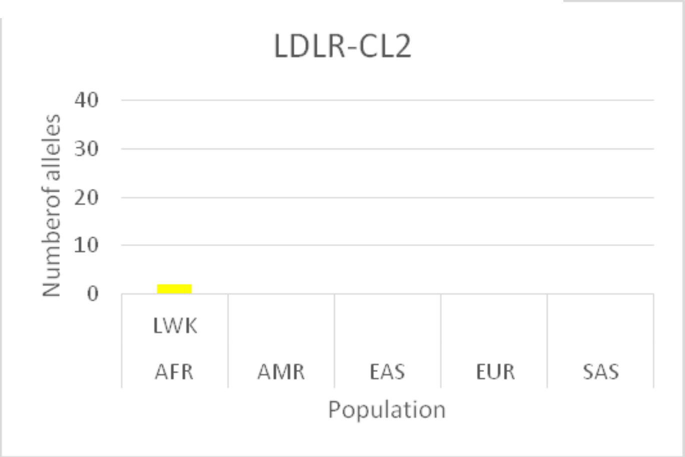
CL1

```
11200818 CGCCCGGCCGGGACCCTCTCTTCTAACTCGGAGCTGGGTGTGGGGACCTCCAGTCCTAAA 11200877
```



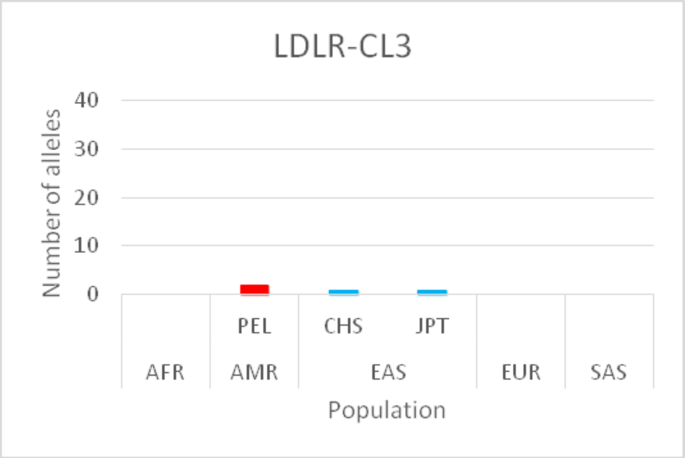
CL2

```
11204838 TGAGATGGAGTTTTGCTCTTGCTGCCCAGGCTGGAGTGCAATGGCGCGATCTCGGCTCAC 11204897

11204898 CGCAACCTCCACCTCCTGGTTCAAGCGATTCTCCTGCCTCAGCCTCCTGAGTAGCTGGGA 11204957
```



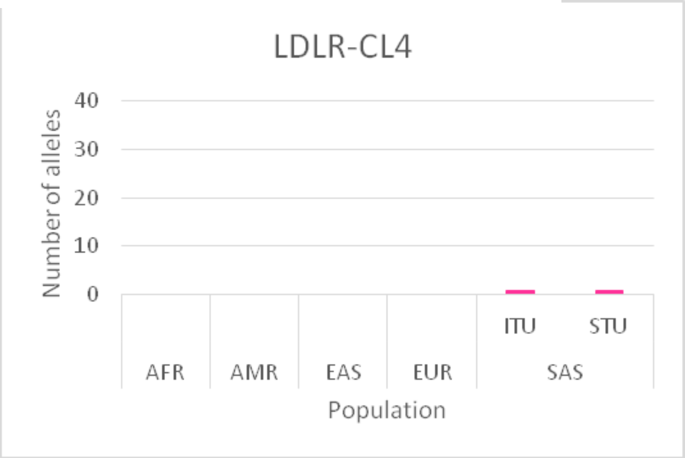
CL3

11204898 CGCAACCTCCACCTCCTGGTTCAAGCGATTCTCCTGCCTCAGCCTCCTGAGTAGCTGGGA 11204957



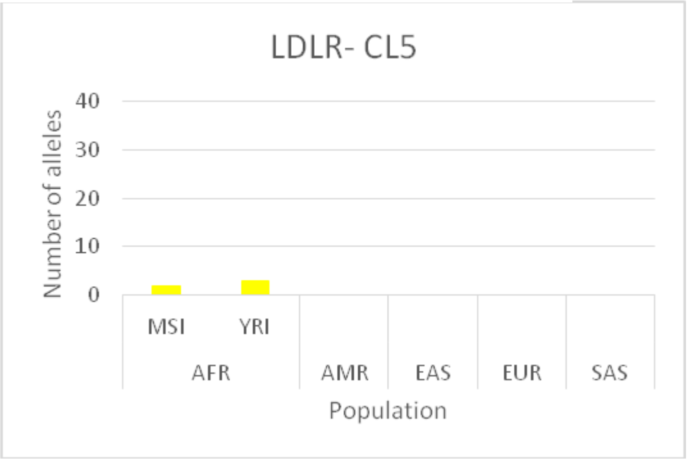
CL4

11212218 CACTGCAAGCTGTGCCTCCTGGGTTCATGCCATTCTCCTGCCTCAGCCTCCCGAGTAGCT 11212277



CL5

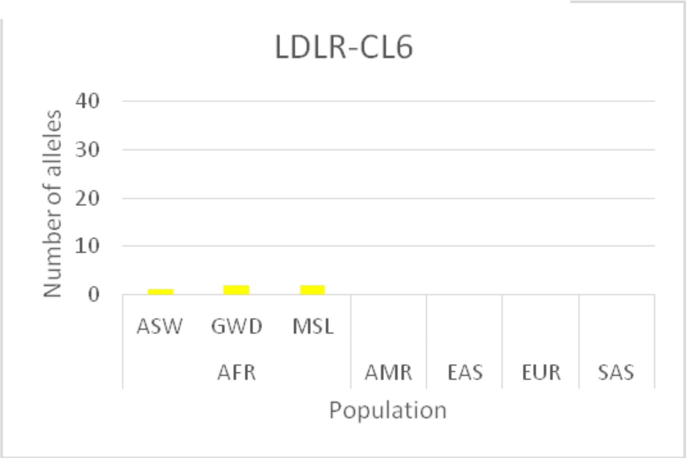
11212818 GCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCAAGATCGTGCCACTGTACTCCAGCCT 11212877



CL6

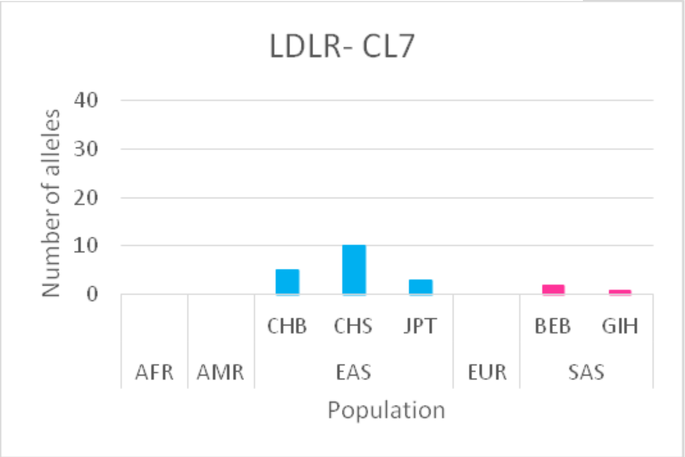
11214558 CTGATAGAAAATTGCAAGTTAGCTCACTGCAACCTCGGCATTATAAGTACTGCACAAAGC 11214617

11214618 CCTCTTCAGCGCACAGCACAAGCACCATTCTATAAAATCTCCAGCAAGCGGCCAGGTGCA 11214677



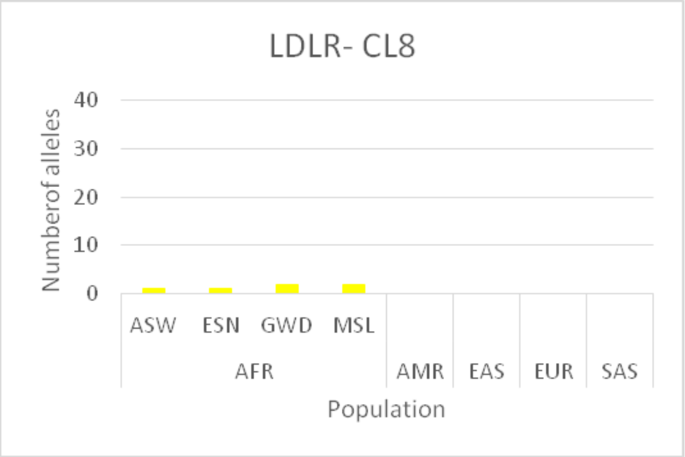
CL7

```
11216958 ACGATTATGGCTCACTGCGGCCTTGATCTCCTTGGCTCAAGCGATCCTCTCACTTCAGCC 11217017
11217018 TCTCAAGCAGTTGGAACCACAGGCTGTACCACCAAGCCTGGCCAATTTTTTTGTACAGAC 11217077
```



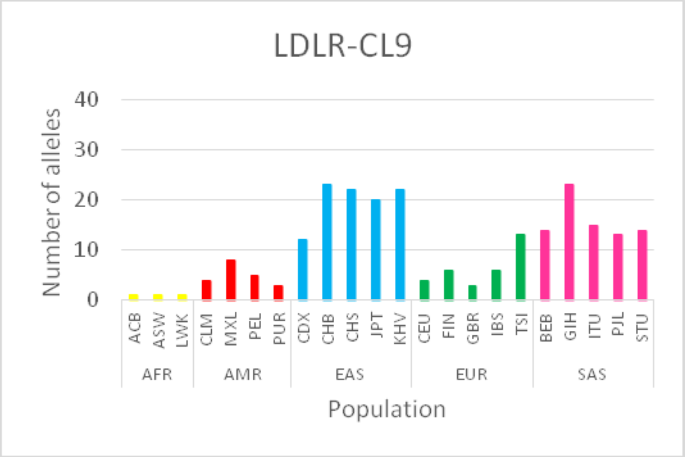
CL8

```
11217918 ACAGGCACAAACCACCGTGCCCGACGCGTTTTCTTAATGAATCCATTTGCATGCGTTCTT 11217977
11217978 ATGTGAATAAACTATTATATGAATGAGTGCCAAGCAAACTGAGGCTCAGACACACCTGAC 11218037
```



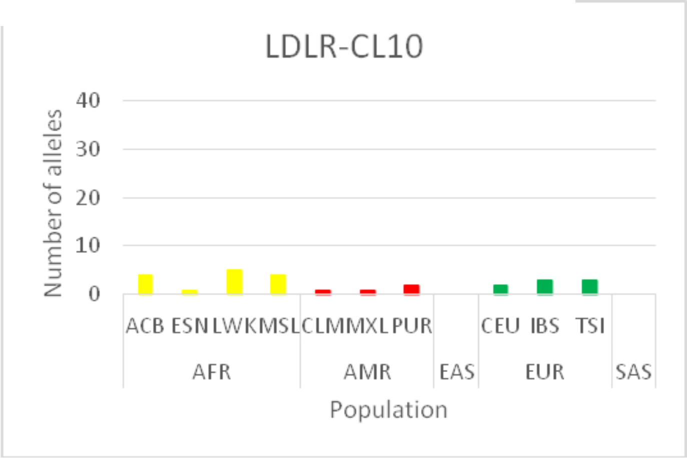
CL9

```
11218218 AGTTTGTGGGGAGCCAGGAAAGGGACTGAGACATGAGTGCTGTAGGGTTTTGGGAACTCC 11218277

11218278 ACTCTGCCCACCCTGTGCAAAGGGCTCCTTTTTTCATTTTGAGACAGTCTCGCACGGTCG 11218337

11218338 CCCAGGCTGGAGCGCAATGGCGCGATCTCGGCTCACTGCAACCTCTGCCTCCCAGGTTCA 11218397
```



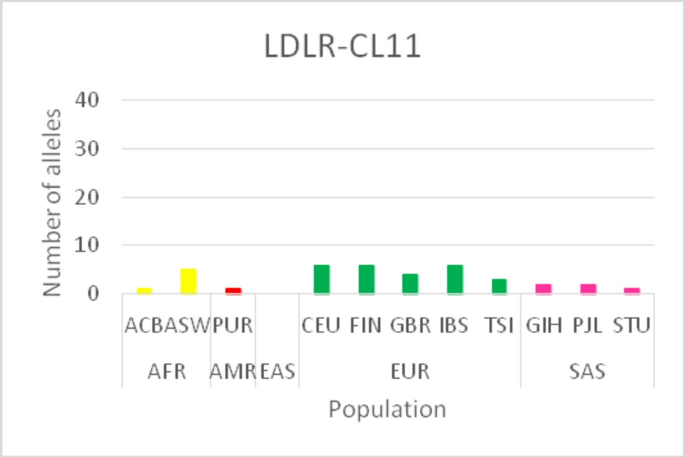
CL10

```
11222058 ATTACATCTCCCGAGAGGCTGGGCTGTCTCCTGGCTGCCTTCGAAGGTGTGGGTTTTGGC 11222117

11222118 CTGGGCCCCATCGCTCCGTCTCTAGCCATTGGGGAAGAGCCTCCCCACCAAGCCTCTTTC 11222177

11222178 TCTCTCTTCCAGATATCGATGAGTGTCAGGATCCCGACACCTGCAGCCAGCTCTGCGTGA 11222237
```



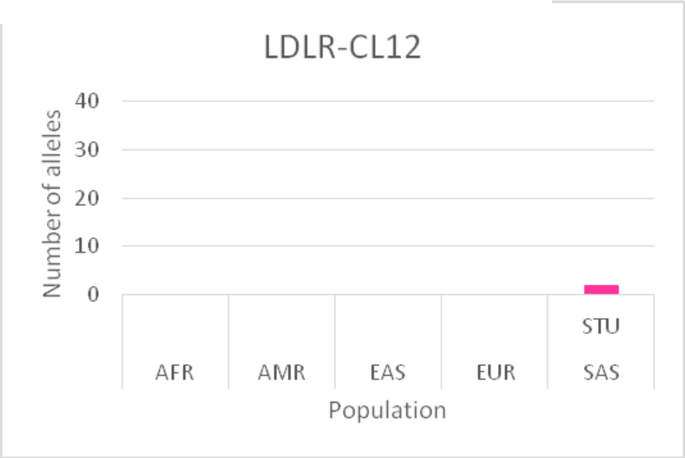
CL11

11222958 TGTATTTTTAGTAGAGACGGGGTTTCACTGTGTTAGCCAGGATGGTCTCGATCTCCTGAC 11223017

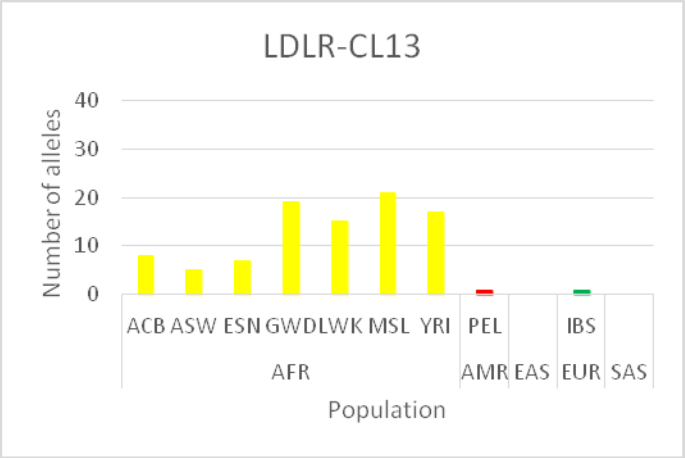


CL12

11223798 CCTGTTCCCGTTGGGAGGTCTTTTCCACCCTCTTTTTCTGGGTGCCTCCTCTGGCTCAGC 11223857



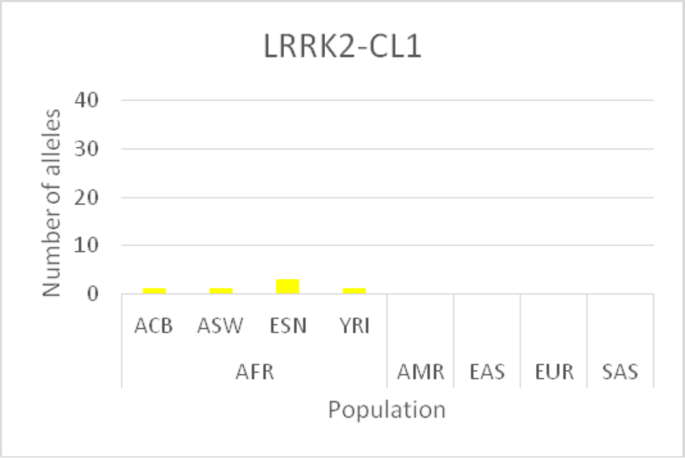
```
CL13

11232918 CGCCTGCCACTACGCCCGGCTACTTTTTTGTATATTTAGTAGAGATGGAGTTTCACTGTG 11232977

11232978 TTAGCCAGGATGGTCTCGATCTCCTGACTTTGTGATCCGCCCGCCTCGGCCTCCCAAAGT 11233037

11233038 GCTGGGATTACAGGCGTGAGCCACCATGCCAGGCTTTTTTTTTTTTTTTTTTTTTTGAGA 11233097
```



LRRK2

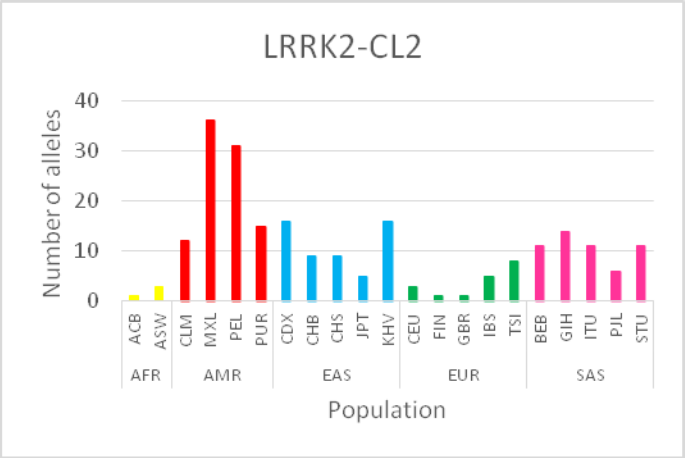
CL1

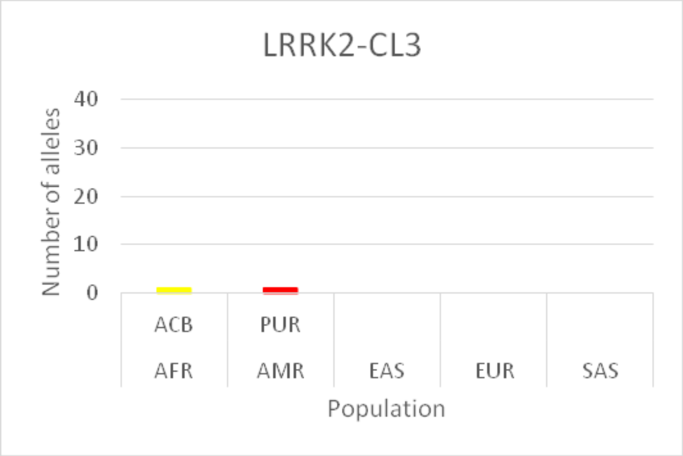
40604826 ACCCACCTGTGACCTGGAAGCCCCCTCCCTGCTCCAAGTTGCTTAGGGAAAAAAAATAAA 40604885



CL2

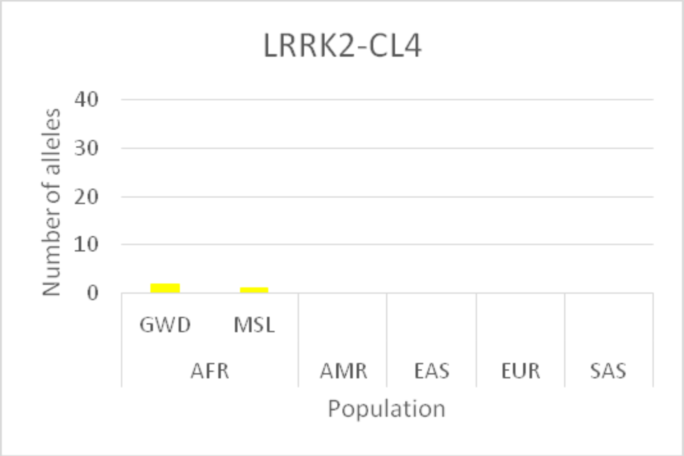
40605126 CTCCTCAAGGCCCAGGGACTATCATGGAAAAGGTGGGTGTGTGAGAATGTAAGAGTCAAT 40605185





CL3

```
40606626 TGTGAGTGTTCTTAACTTACAATAACACAGTTATTTGTATAAGTGCAGTGAAAATCTGTT 40606685
```

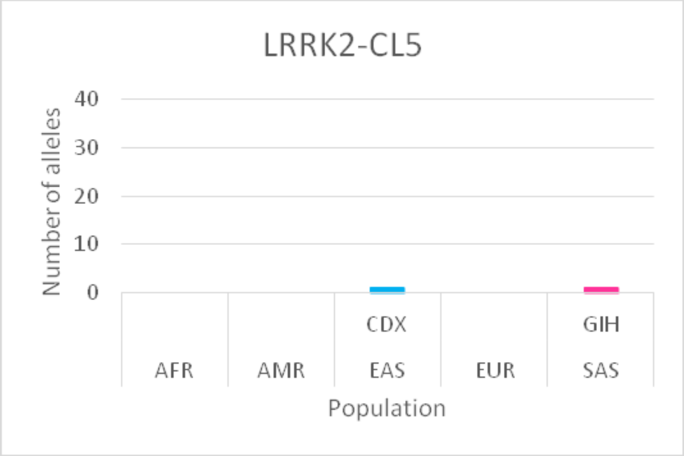


CL4

```
40634106 TAGCTTGTTTTCTCATTATAACATTCTTAGGAAGGGCTGCTTCACAGAAATATATTTTTT 40634165
40634166 ATTTAAGGAGATTACACTTGATGTATCTCACACAACTATAATGAATATTGTAATTTTTGA 40634225
```

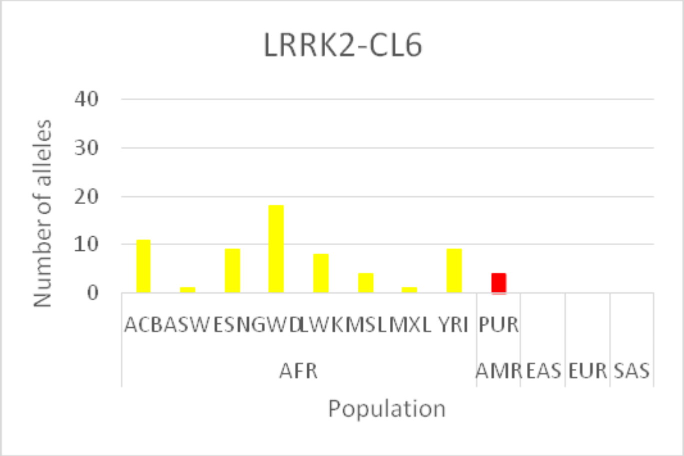
CL5

40641246 TATATAAAAATATGTATATAAATATATACACATTGTATATAAATGTGTATATATATTTAC 40641305



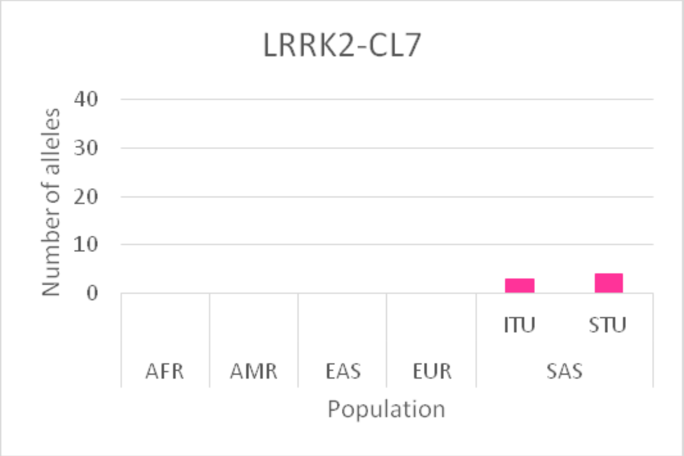
CL6

40667166 AGATGATAAATGAAATGATGTCCAAGCTGAGCAATTAAAGTGTGAAGTAGAAGGACACAG 40667225



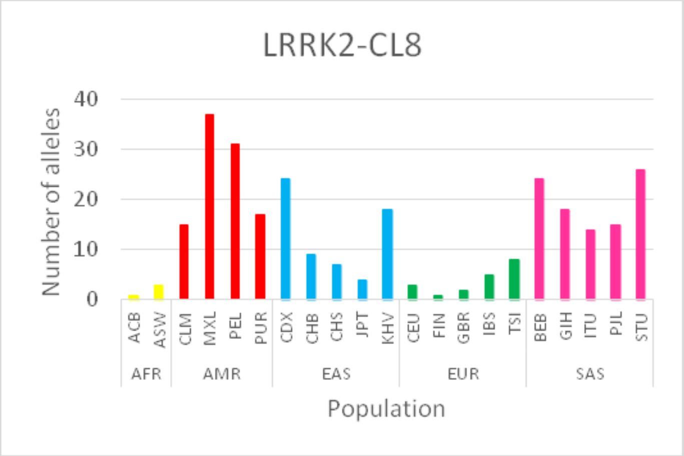
CL7

```
40675266 GATGATCAGAGAAAGATTGCAGGGATAAGAAATTATGCTTTTGATAATCTTTAGTTATAT 40675325
40675326 TCTTAATTTTCTTCATTATTATTTAAATGTAAAAATAAATATCTGTGAGCAGTAGTATTT 40675385
```

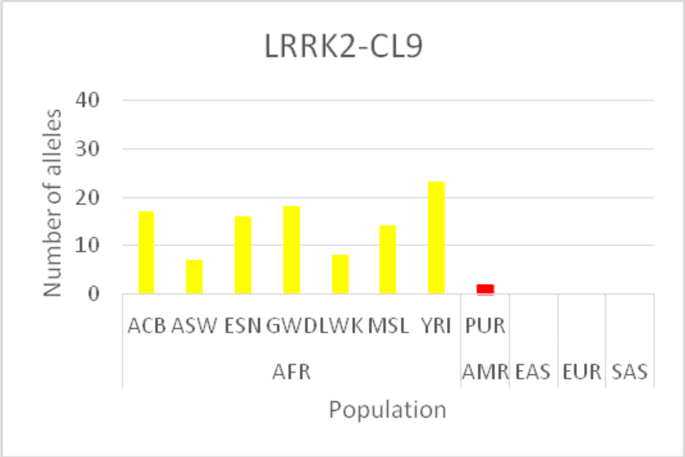


CL8

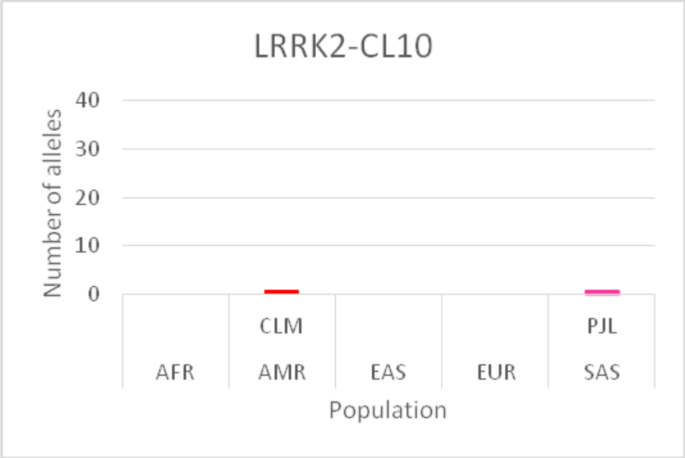
```
40677546 TTGCCTTATTTTATTTTGTTTCATTCCAAATTGGAGATGTAGAGAAAAATCACATGAAGT 40677605
40677606 TTGATTTGCCAGTCTCCTAAAAGGAAGAAAAATGTAGATTTTTAATATACTTAATTTTTT 40677665
```



```
CL9

40688166 TGAAAGACCTGTTCTAACCTATTCTCCAATTTTGATTATAGCTGAGTACTAAAAATATGA 40688225

40688226 GGGTTGTTTTGTGTTAATTCTAGATCTTAAGATGGGTGAAATGAATGACTGTAGTTGAAT 40688285
```

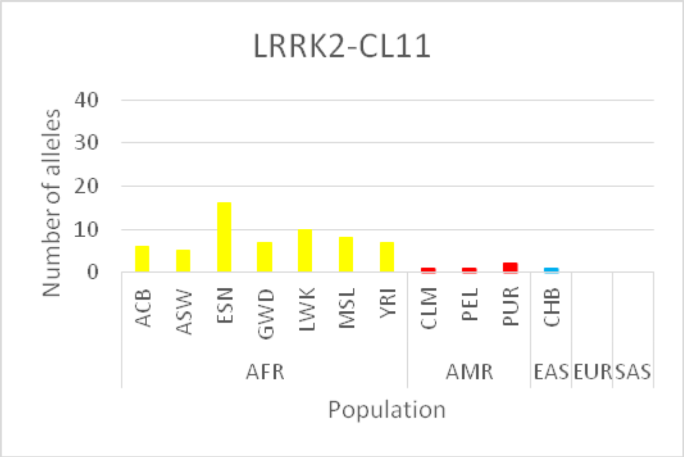


```
CL10

40723146 TGCTATGTGAGATGAGGAAAATTAACGCTATTCTTTCTCCTTTTCCCATCACCTTCTCAA 40723205

40723206 GTTCTTTAATTTATTCTATTATTTTTATGTAGTGAAAGTTTATAACATTTATATTCTGGT 40723265
```



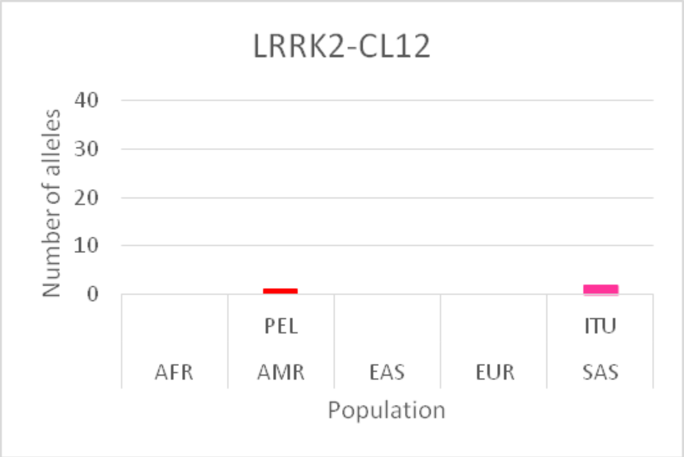
CL11

```
40730406 CTTCTTTCTCTTCTGAAATGCTATGAATATGCCTTTTAGGTAGTATCCAGAAATGTTCCT 40730465
40730466 TCCTGAAAGGGTCCAGAAACTACTGAAAACTGTACAGATTATGAAATGAAACAGGGTGCA 40730525
40730526 ............................................................ 40746785
40746786 CAGACGGGGTGGCAGCCGGGCAGAGGGGCTCCTCACTTCCCAGAAGGGGCGGCCGGGCAG 40746845
40746846 AGGCGCCCCCCACCTCCCAGACGGGGCGGCGGCCGGGCGGGGGCTGCCCCCCACCTCCCG 40746905
40746906 GATGGGGTGGCTGCCCAGCGGAGACGCTCCTCACTTCCCAGACGGGGCGGCTGCTGGGCG 4074696
```



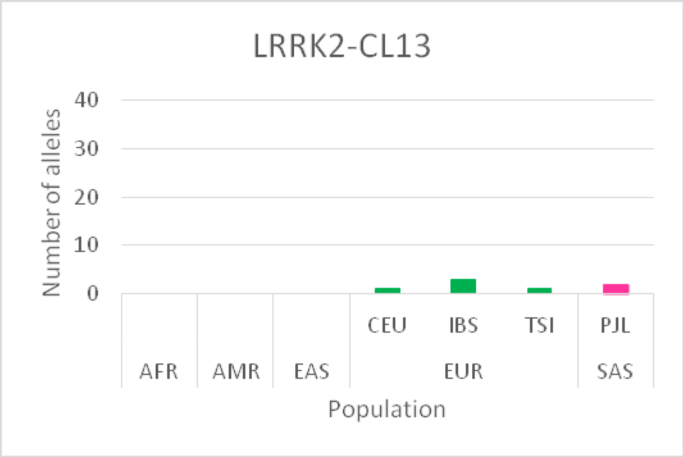
CL12

```
40738746 GTGTAGTAGGGGTTATCATACTCAAATTCGATGTCTCCATCCTTCCAACTCTTCATGCTT 40738805
```



CL13

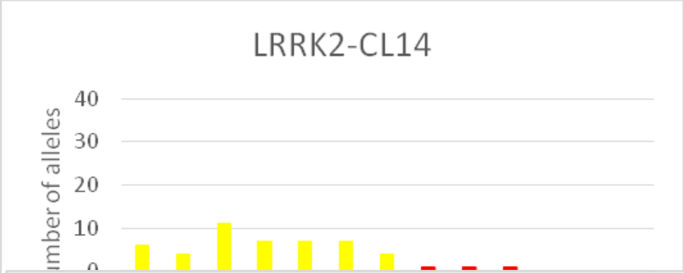
40746306 CGCAGAGGGGGATTTGGCAGGGTCATAGGACAATAATGGAGGGAAGGTCAGCAGATAAAC 40746365



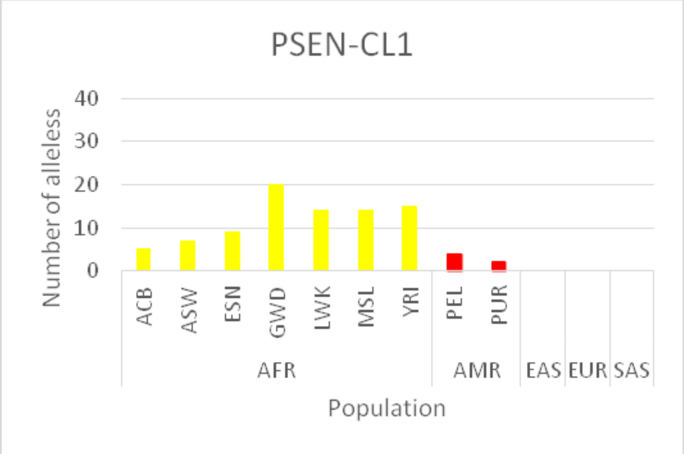
CL14

40762986 TTCATTGTTACTTTGTATTTGCAATTTTTTTTACCAAAGACAAATTAAAAAAATGAATAC 40763045

40763046 CATATTTAAATGGAATAATAAAGGTTTTTTAAAAACTTTAAA 40763087



PSEN1

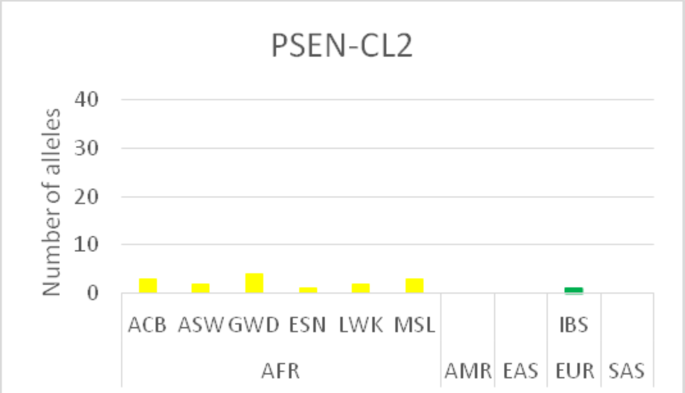


CL1

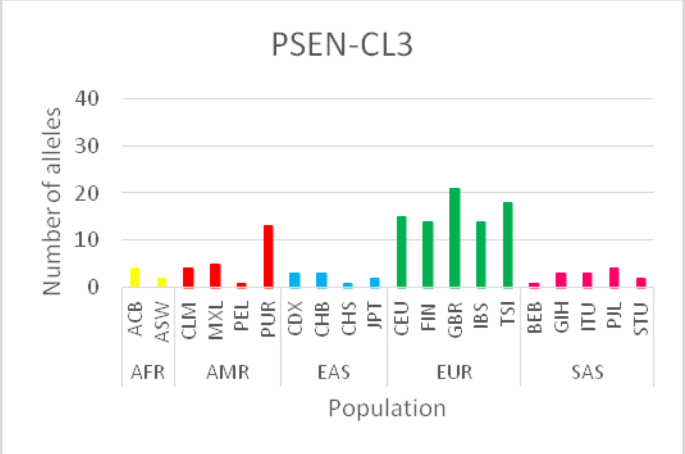
```
73608706 CGGAGCTTGCAGTGAGCTGAGATTGTGCCACTGCACTCCAGCCTGGGCGACAGAGGGAGA 73608765
73608766 CTCCGTCTCAAAAAAAAAAAAAAGAAAAGAAACTACATCTCAATAATAATAATAATTTCA 73608825
```

CL2

```
73612006 CTATGGGTTTTTCCATAATGGAGTGTACTTTTATTATTTTATTATTTGTTTTTTTGAGAC 73612065
73612066 AGGGTCTCACTCTGTCACTCAGGCTGGAGTGTAGTGATGCAGTCACTGCTCACTGCAGCT 73612125
73612126 ............................................................ 73635165
73635166 CTGATTGTCGGAAATACAAGCCACTGAGTGTTGGTGATACAAGTGGCTGAGCAGCGAGCA 73635225
```



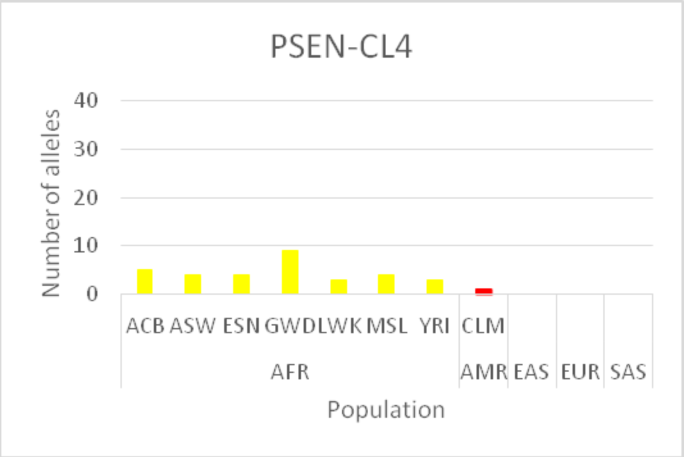
CL3

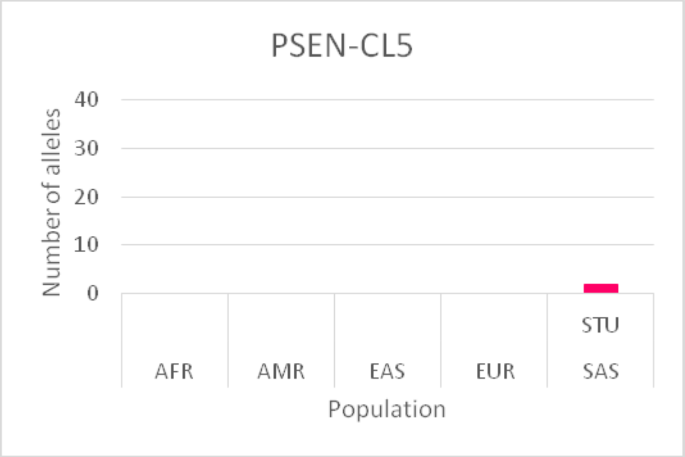


```
73620886 TAAACACAACTATGTATACCTGTGTGTATATATTGAAAAACGTTAGCATTTGGGGTGTCG 73620945
73620944 ............................................................ 73671945
73671946 AGGCTGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACACAGTGA 73672005
73672006 AACCCCACCTCTACTAAAAAAAATACAAAAAATTAGCCAGGCATGGTGGCGGGCGCCTGT 73672065
73672066 AGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCAGGAGACGGAGCTT 73672125
```

CL4

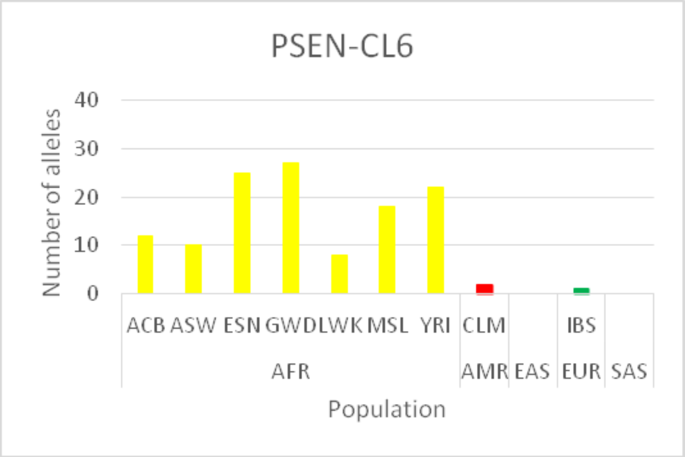
```
73622626 ATGTAGGTTCATCTGTTGTAACGGTTTACCACAGTGGTGCAGGATGTTGATAGAGGCTGT 73622685
73622686 TGGAGGGGTTATGTGGGAACTCTGTGTACTTTCTGCTCAGTTTTGCTGTGGATCTAAGTT 73622745
73622746 TGCATTTAAAAATATTTTATGGCTTGGTGGCTCATACTTGTAATCCTAGCACTTTGGAAG 73622805
73622806 GCTGAGGCGGGTGGATCACTTGAGTTCAGGAGTTTGAGACCAGCCTGGGCAACATGGTGA 73622865
```





CL5

73657966 CTACTCTTCTCTCACACTACTTTGCATAAATGGGATCATAAATGGGATCCTATATTTTTT 73658025



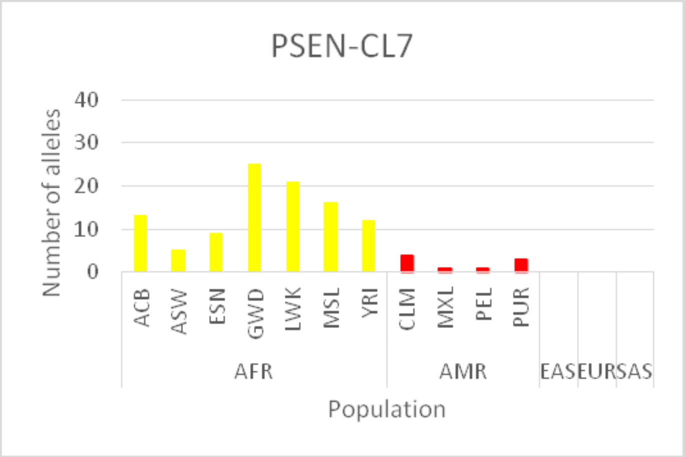
CL6

73663786 ACGCGCCACCACACCCAGCTAATTTTTTGTATTTTTTAGTAGAGACAGGGTTTCACCGTG 73663845

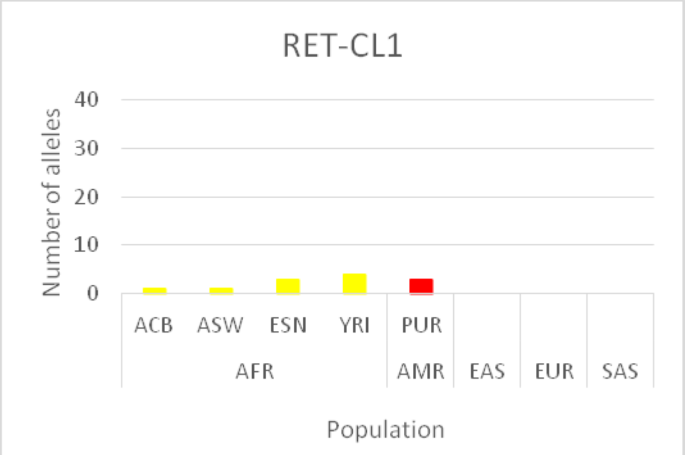
73663846 TTAGCCAGGATGGTCTTGATCTCCTGACCTTGTGATCCGCCCACCTCGGCCTCCCAAAGT 73663905

CL7

```
73670746 GCCCAGGCTGGAGTGCAATGGCACGATCTCGGCTCACCACAGCCTCCGCCTCCTGGATTC 73670805
73670806 AAGCAATTCTCCTGCCTCAGCCTCCCGAGTCGCCGGGATTATAGACATGCACCACCACAC 73670865
73670866 ............................................................ 73674885
73674886 TTCTCTCCTTCTCGCCACCTGCAACGTGGCAAGTGGGGGGCACGTTTCAGCCCTGTTTGT 73674945
73674946 CTTACGGTTCTTTCAATCCTGCCATTCAATGGGTCCCAAGTTCTTGTCCTGCATCCAGGA 73675005
73675006 AGAATGAAGTACGTGGACAGCTGGAGGGAGAGCAAGATGAAAAGGTGCTTTATTGAGCAA 73675065
73675066 ............................................................ 73679925
73679926 CTCAGTTACCACAGTAATTAGGTTGCCTCTTCTACTTTCCTCTTTTCTCACAGGCACCAG 73679985
73679986 GAGCCAGAGGAAATAACATAATAGTTGTTGACCAGAGCAGCAGCATAATTCTTTCATGAC 73680045
73680046 TGCCTTTTCTAATTTGACGATTCCCTCTCCTGAGAGGGCTCTTTGTGTCCTCCTCCTCTT 73680105
73680106 CGTCTCCAACTTTTAAAAAAAAAAAAAGTGAAACTATCAAGTATTGCTCCTGCTAACTTCA 73680165
```



RET

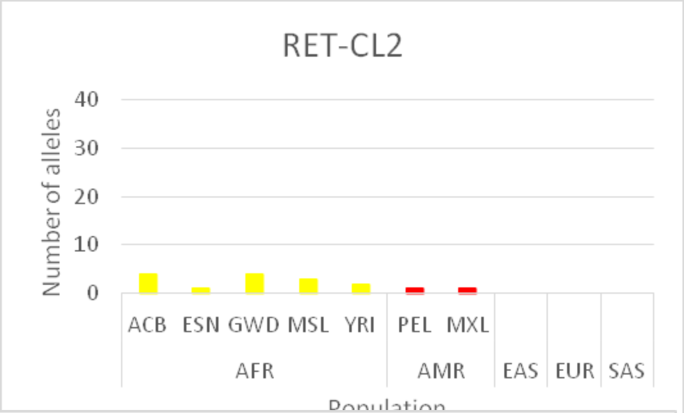


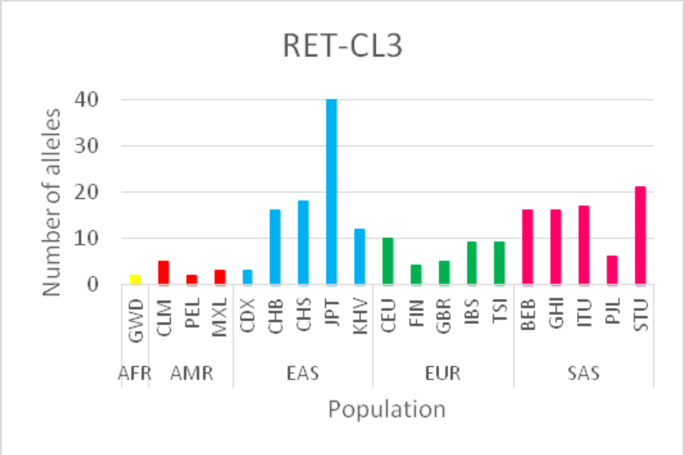
CL1

```
43583575 TGATGGTGGTGGAGGCAGTTGTGGTGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGG 43583634

43583635 TGGTGATGGTGAAGACGGTGGTGGGAATGGTGGTGGTGGAGGCAATGGTGGTGGTGATGG 43583694
```

CL2

```
43583635 TGGTGATGGTGAAGACGGTGGTGGGAATGGTGGTGGTGGAGGCAATGGTGGTGGTGATGG 43583694
```



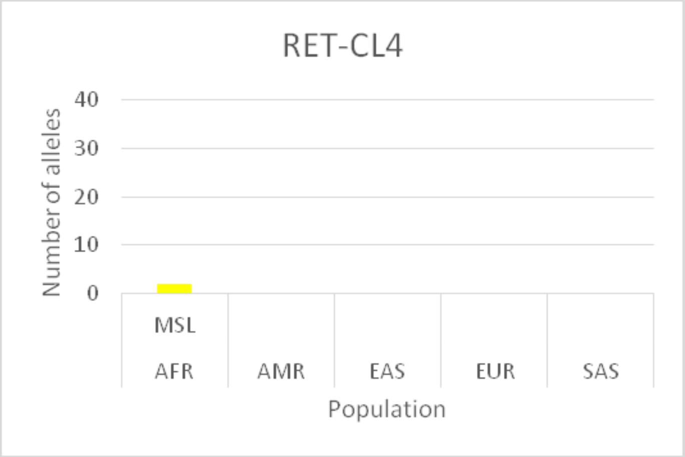


CL3

43583635 TGGTGATGGTGAAGACGGTGGTGGGAATGGTGGTGGTGGAGGCAATGGTGGTGGTGATGG 43583694

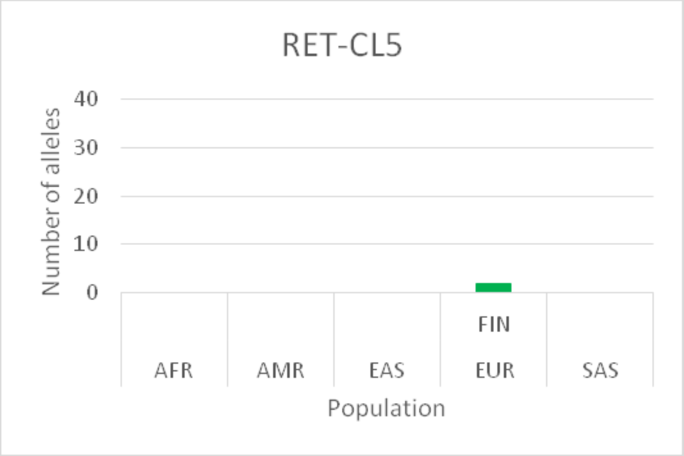
CL4

43590235 GCTTGTTAGAGCCCTATTGGCAGTGTGAGAATGTGTGAGCATATGGGAGAACAAGTGTAT 43590294



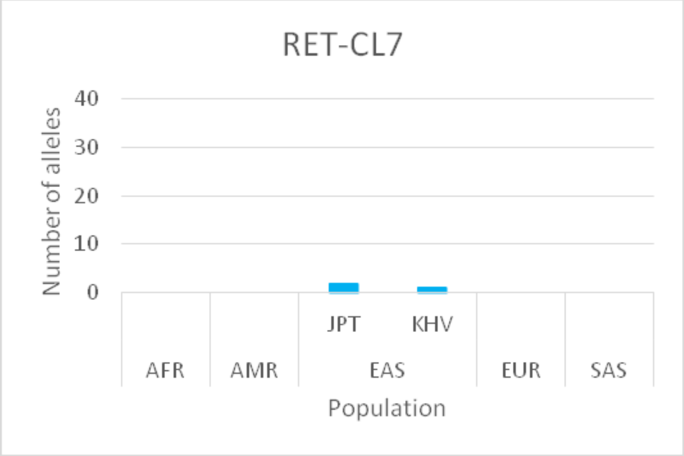
CL5

```
43590895 AGGCCCGCCTGGGCATGAGAACATGCACACCTGGACGGAGTGTGTGGATTGACAAGGTGT 43590954
```



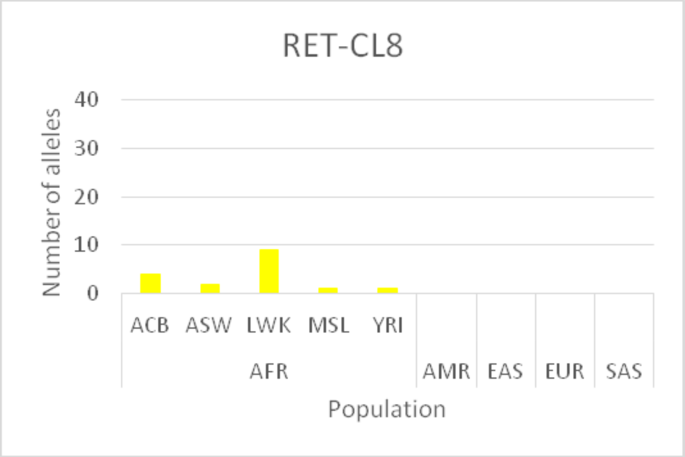
CL6

```
43599295 AGGCATGGTTCTGCCTGGGAGGCTTGTGGAGTTGCCCCCGCCTCCTTGGGGTGGGTTGTG 43599354
43599355 GGGTCTTCAGCGGTGACCTTCCGCCTTCAGCATGCCCCCTTTTCTGGATTATTTCTGGAG 43599414
```



CL7

43618495 CGCCCAGTGACCTCTGGCTGCCCTCTGGTGTGCTCCGTGGTGTGCACATGTATGCTTTTT 43618554



CL8

43625275 ACGTAACCTGGCTCTAATTTGGGCTGTTTTTCAGATACACTGTGATAAGTTCTTTTACAA 43625334

Subtitle:

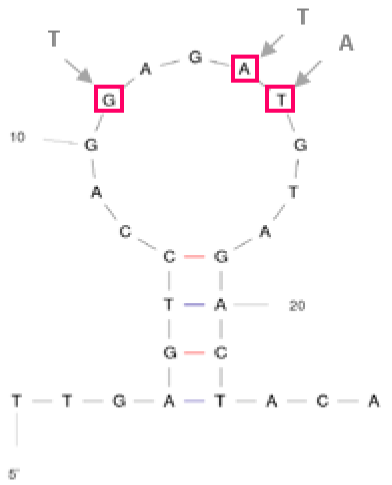
🟦 Intron

🟩 Exon

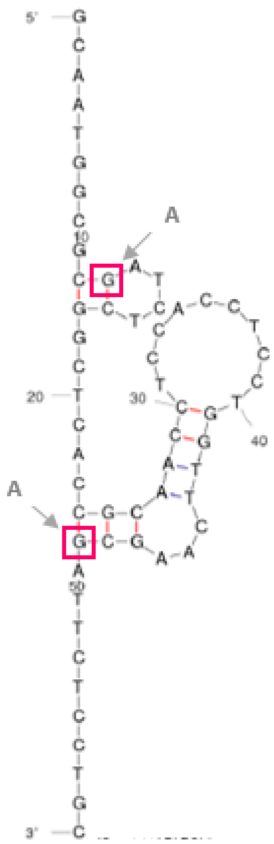
□ SNP

Appendix III: Predicted structures

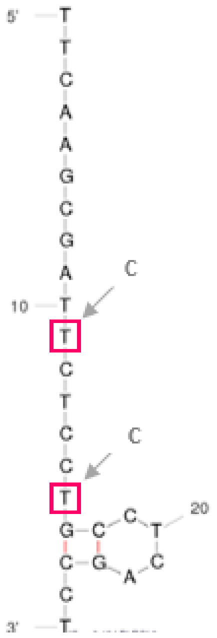
BRCA1 CL1



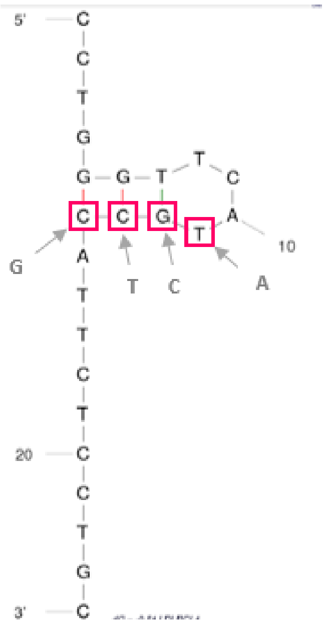
LDLR CL2



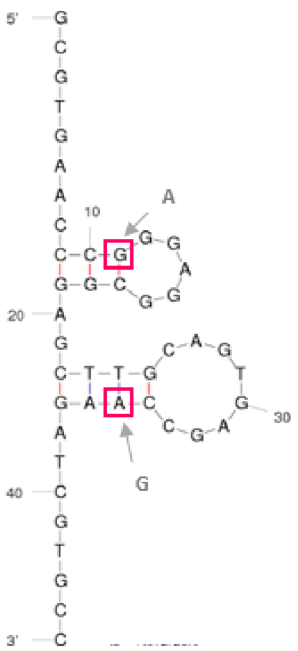
LDLR CL3



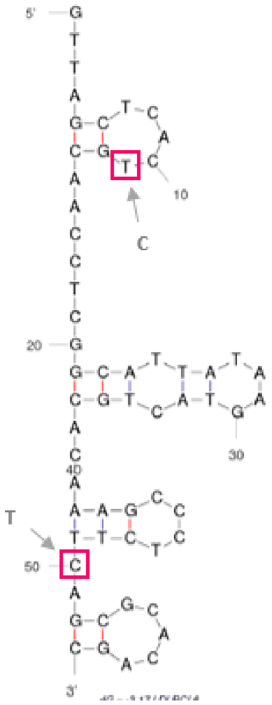
LDLR CL4



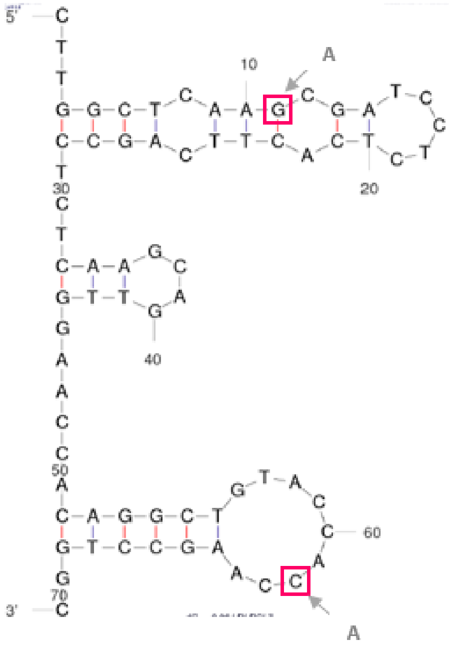
LDLR CL5



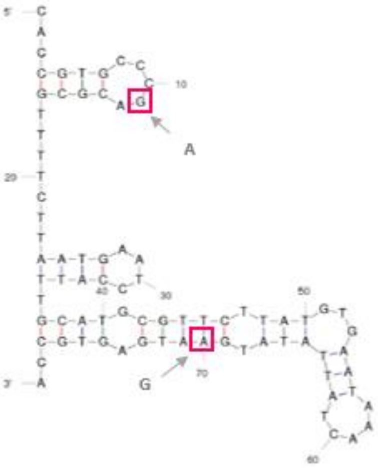
LDLR CL6



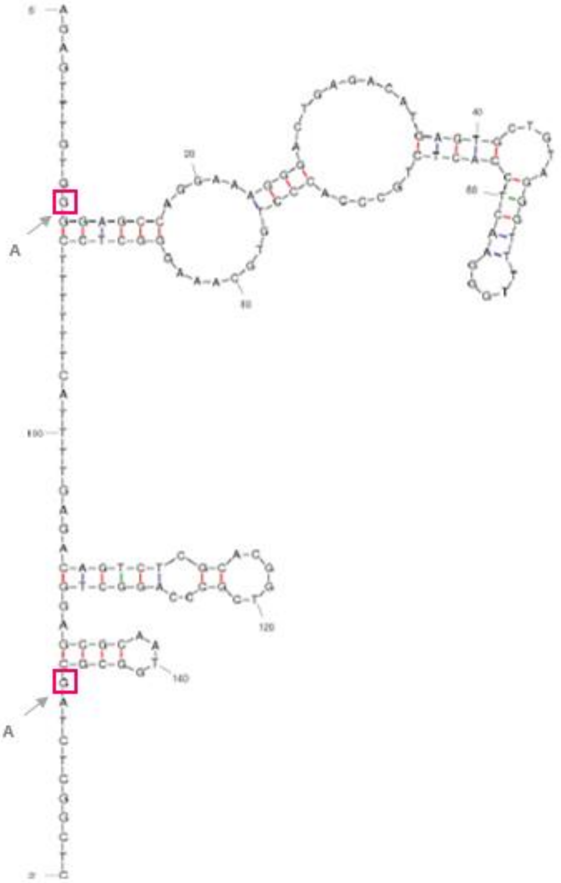
LDLR CL7



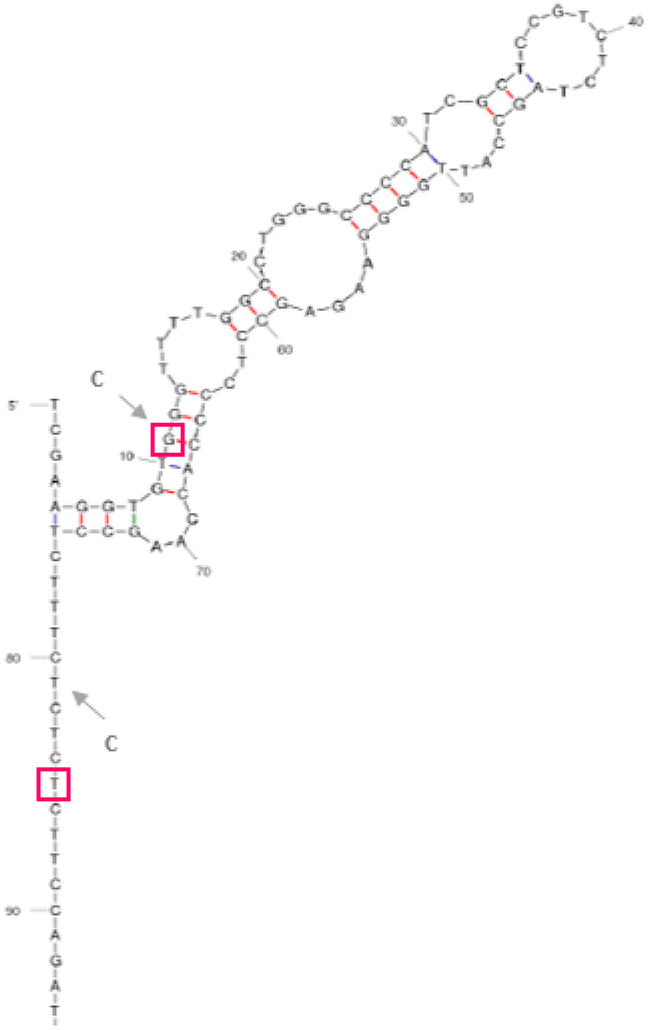
LDLR CL8



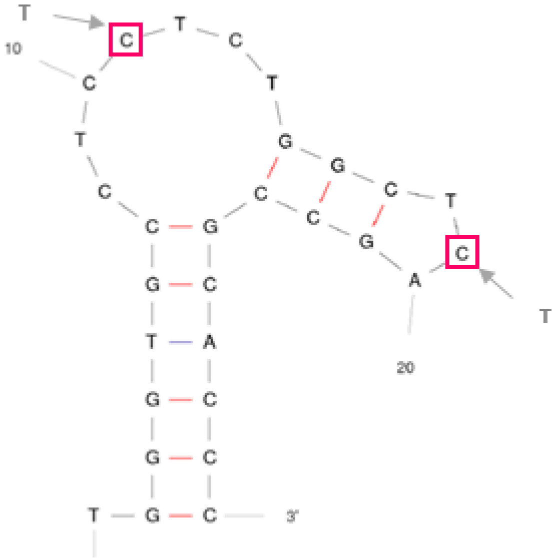
LDLR CL9



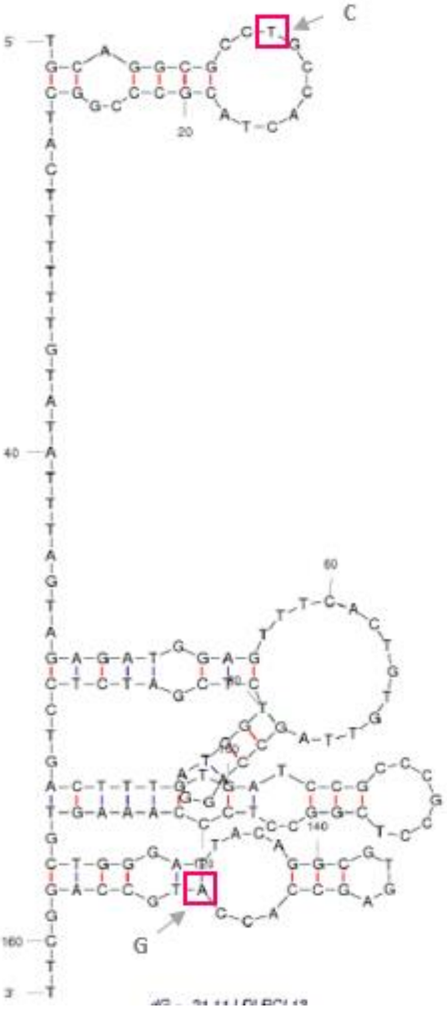
LDLR CL10



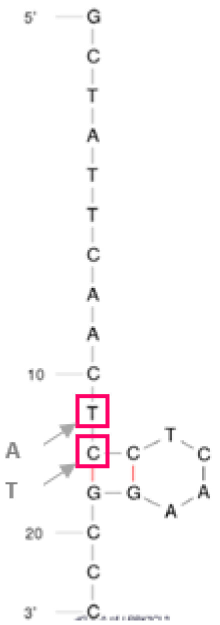
LDLR CL12



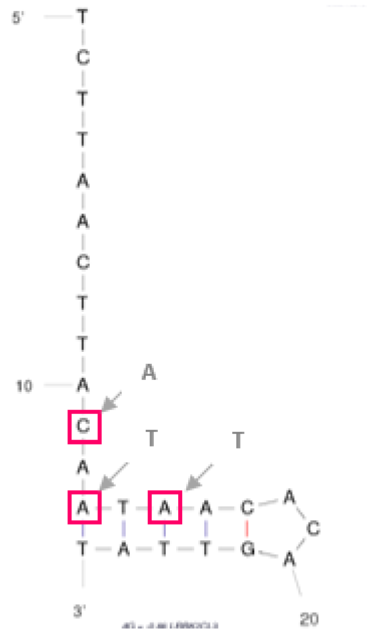
LDLR CL13



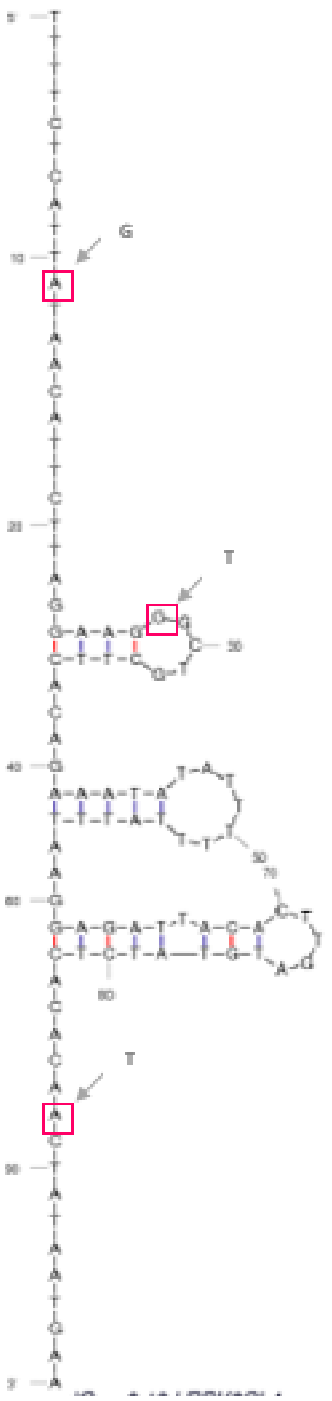
LRRK2 CL2



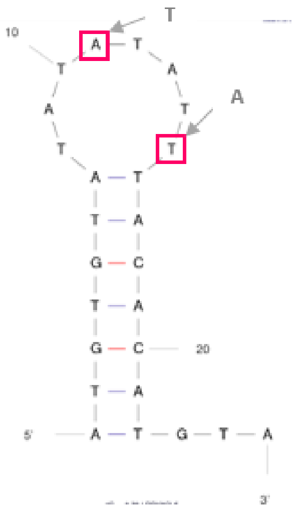
LRRK2 CL3



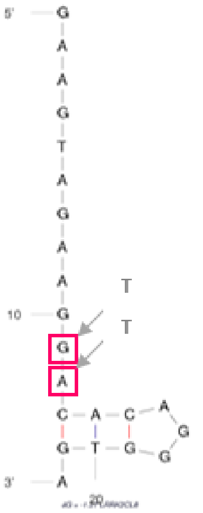
LRRK2 CL4



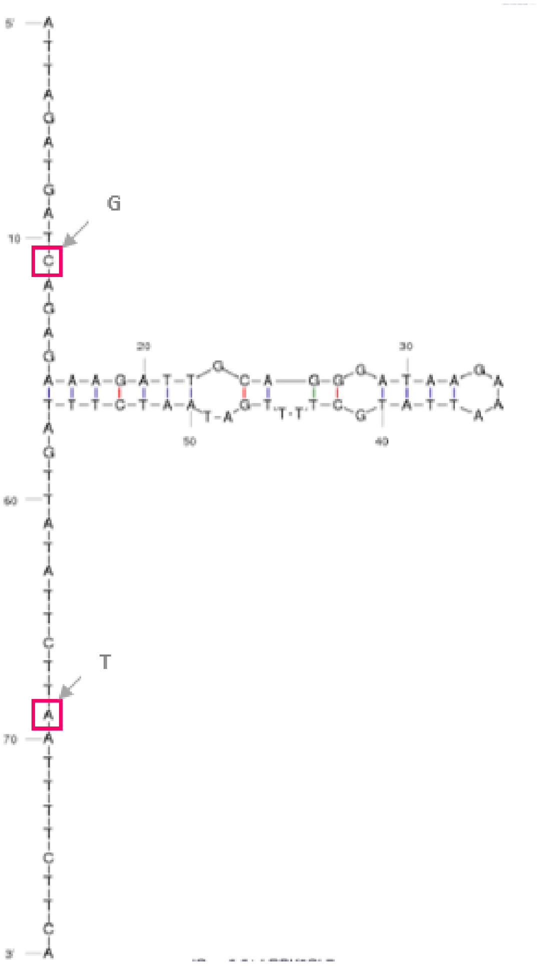
LRRK2 CL5



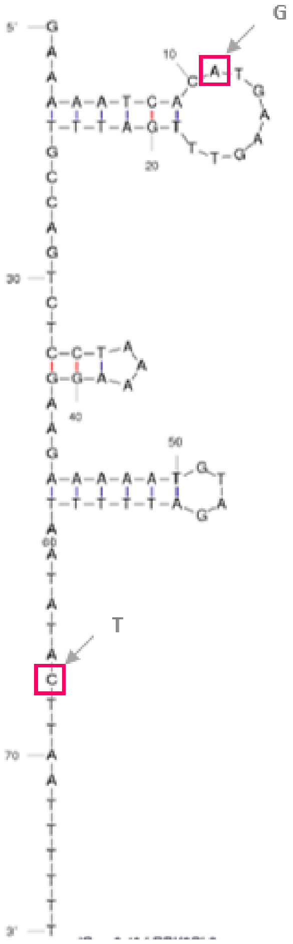
LRRK2 CL6



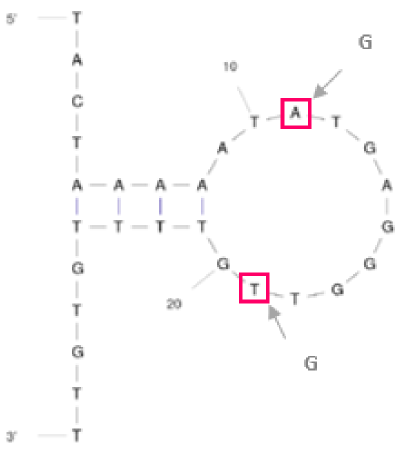
LRRK2 CL7



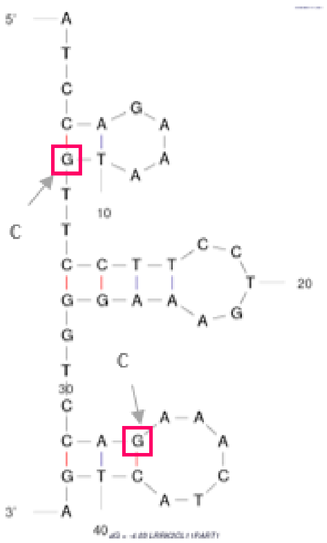
LRRK2 CL8



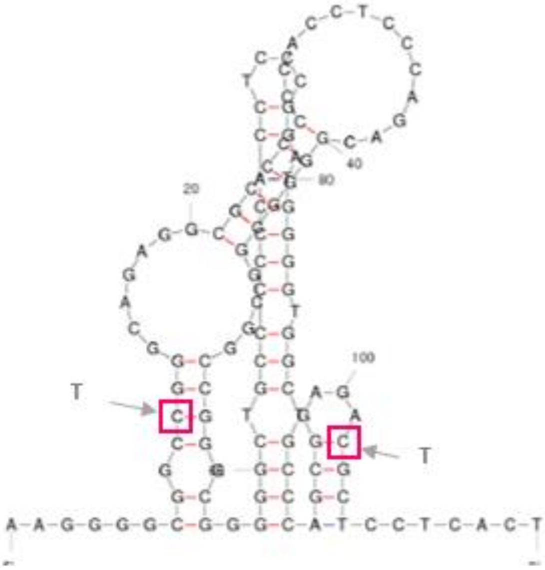
LRRK2 CL9



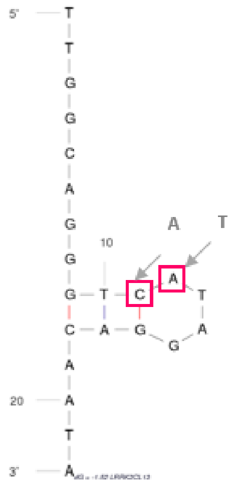
LRRK2 CL11 PART1



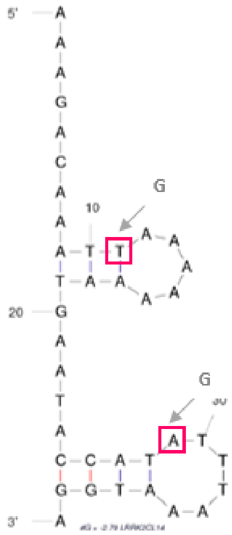
LRRK2 CL11 PART2



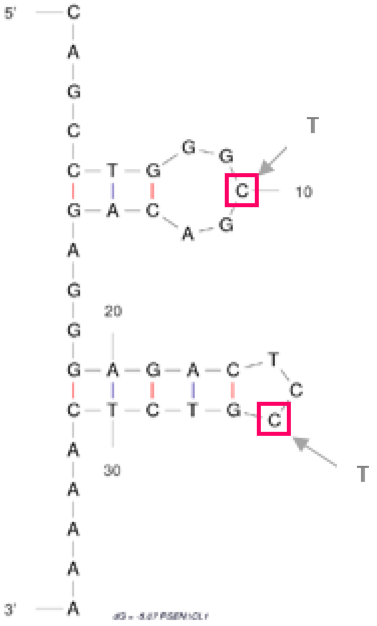
LRRK2 CL13



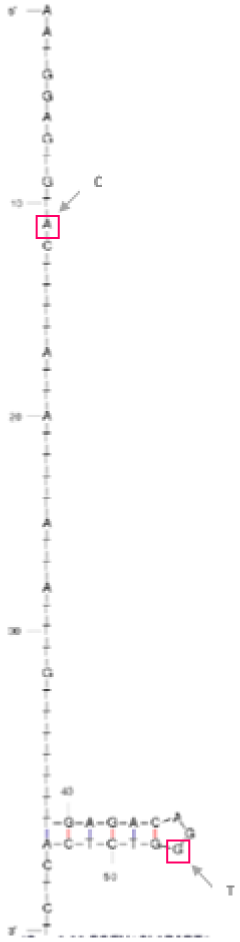
LRRK2 CL14



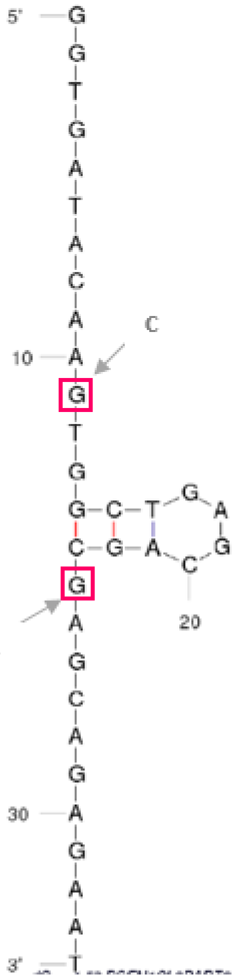
PSEN1 CL1



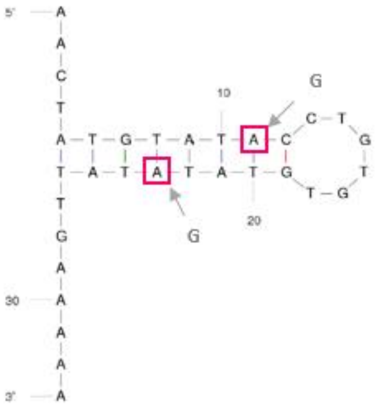
PSEN1 CL2 PART1



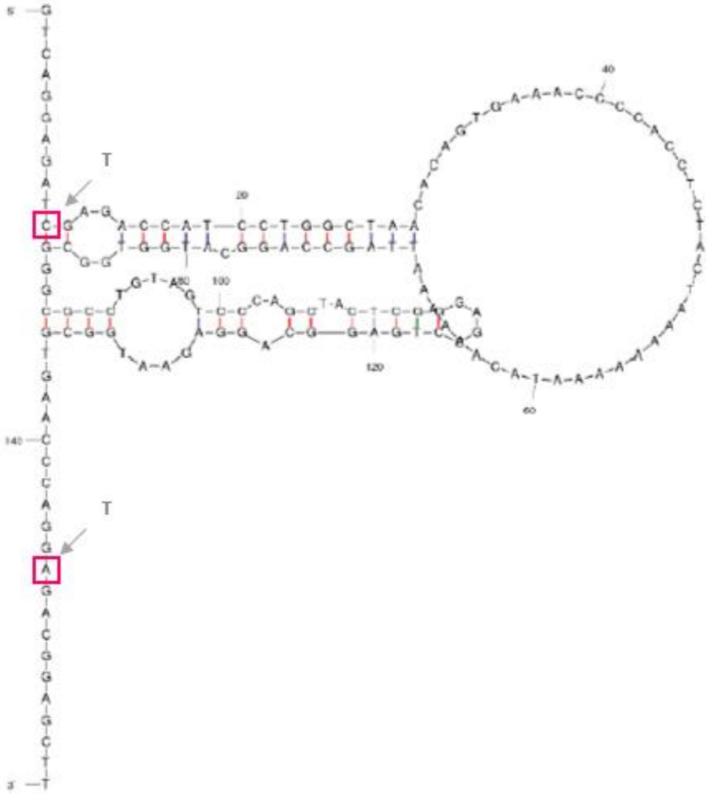
PSEN1 CL2 PART2



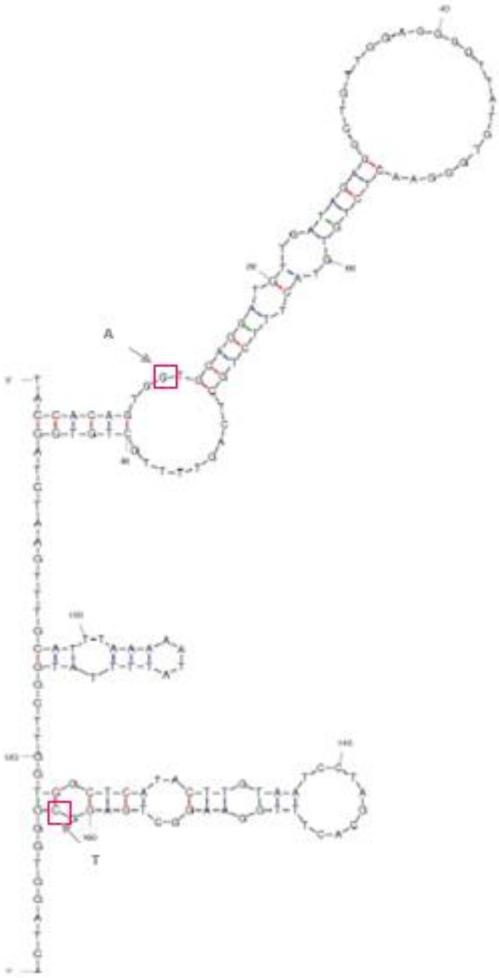
PSEN1 CL3 PART1



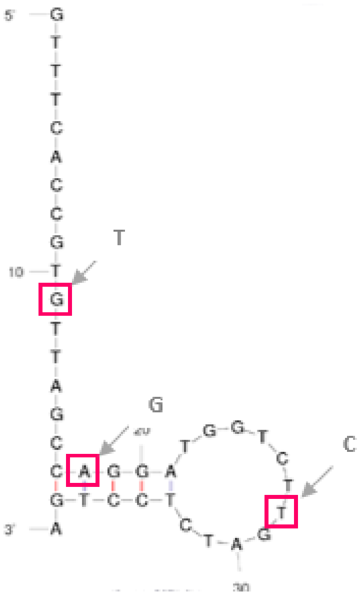
PSEN1 CL3 PART2



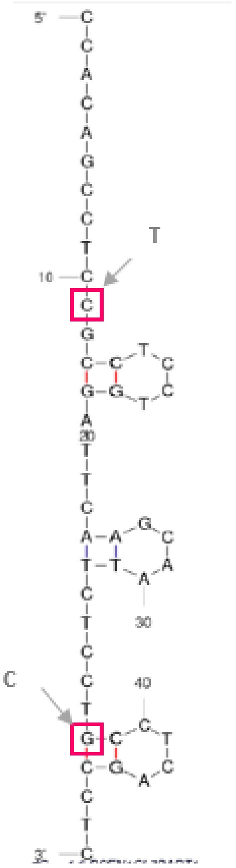
PSEN1 CL4



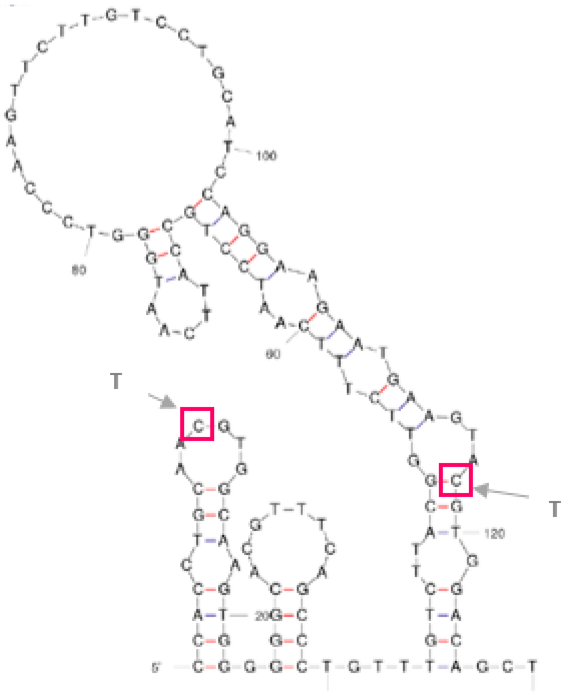
PSEN1 CL6



PSEN1 CL7 PART1

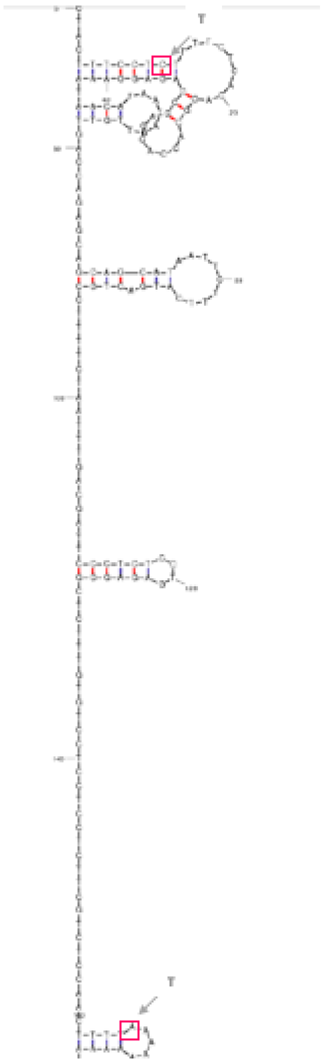


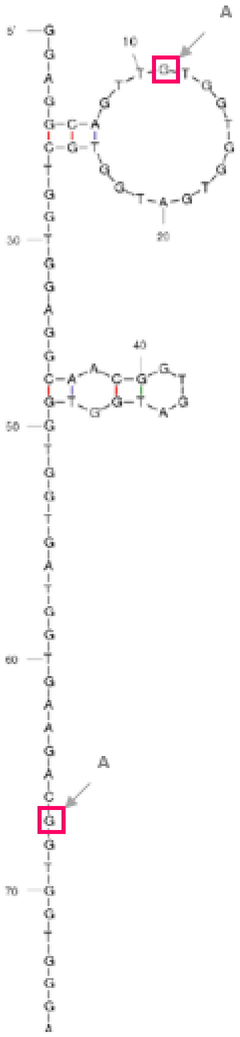
PSEN1 CL7 PART2



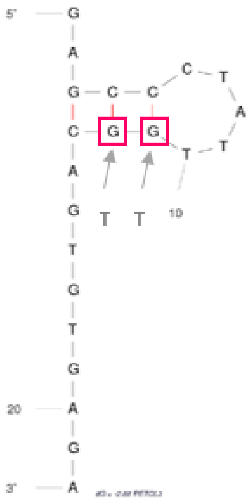
RET CL1

PSEN1 CL7 PART3

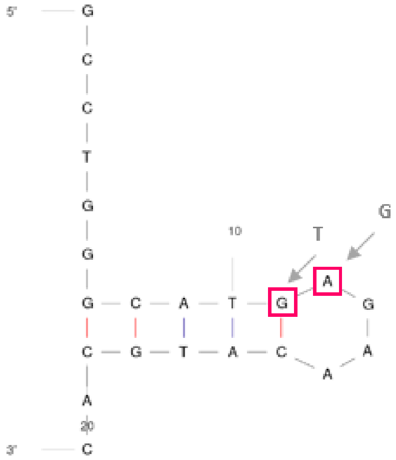




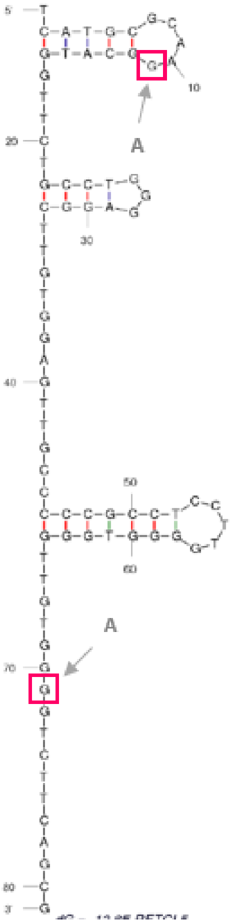
RET CL4



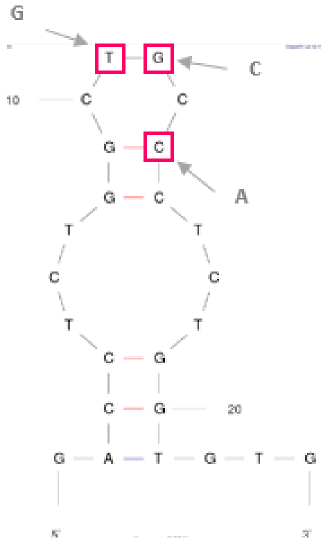
RET CL5



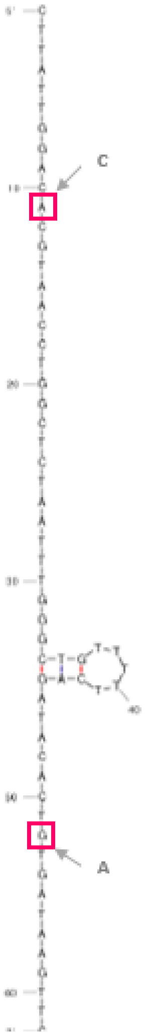
RET CL6



RET CL7



RET CL8



References

1. Johnston, M.O., *Mutations and New Variation: Overview*, in *eLS*. 2001, John Wiley & Sons, Ltd.
2. MacDouga, D.T., *Species and varieties, their origin by mutation; lectures delivered at the University of California by Hugo De Vries ed. by Daniel Trembly Mac Dougal*. 1905, Chicago: The Open court publishing company;.
3. De Vries, H., *The Principles of the theory of mutation.* Science, 1914. **40**(1020): p. 77-84.
4. Antonarakis, S.E., *Molecular genetics of coagulation factor VIII gene and haemophilia A.* Haemophilia, 1998. **4**: p. 1-11.
5. Mullaney, J.M., et al., *Small insertions and deletions (INDELs) in human genomes.* Hum Mol Genet, 2010. **19**(R2): p. R131-6.
6. Vogelstein, B. and K.W. Kinzler, *The Genetic Basis of Human Cancer*. 2002: McGraw-Hill, Medical Pub. Division.
7. Bettencourt, C. and M. Lima, *Machado-Joseph Disease: from first descriptions to new perspectives.* Orphanet Journal of Rare Diseases, 2011. **6**(1): p. 35.
8. Kondrashov, A.S., *Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases.* Human Mutation, 2003. **21**(1): p. 12-27.
9. Nachman, M.W. and S.L. Crowell, *Estimate of the Mutation Rate per Nucleotide in Humans.* Genetics, 2000. **156**(1): p. 297-304.
10. Haldane, J.B.S., *The rate of spontaneous mutation of a human gene.* Journal of Genetics, 1935. **31**(3): p. 317-326.
11. Cooper, D.N., Krawczak, M., *Human gene mutation, by, Bios Scientific Publishers, Oxford, U.K., 1993,402 pp, $99.* Molecular Reproduction and Development, 1994. **38**(3): p. 356-356.
12. Kimura, M., *Evolutionary Rate at the Molecular Level.* Nature, 1968. **217**(5129): p. 624-626.
13. Haldane, J.B.S., *The mutation rate of the gene for haemophilia, and its segregation ratios in males and females.* Annals of Eugenics, 1946. **13**(1): p. 262-271.
14. F., V.A.G., Motulsky, *Human Genetics: problems and approaches*, Springer-Verlag, Editor. 1997: Berlin.
15. Drake, J.W., et al., *Rates of Spontaneous Mutation.* Genetics, 1998. **148**(4): p. 1667-1686.
16. Ségurel, L., M.J. Wyman, and M. Przeworski, *Determinants of Mutation Rate Variation in the Human Germline.* Annual Review of Genomics and Human Genetics, 2014. **15**(1): p. 47-70.
17. Kong, A., et al., *Rate of de novo mutations and the importance of father's age to disease risk.* Nature, 2012. **488**(7412): p. 471-5.
18. Conrad, D.F., et al., *Variation in genome-wide mutation rates within and between human families.* Nature genetics, 2011. **43**(7): p. 712-714.
19. Loeb, L.A. and R.J. Monnat, *DNA polymerases and human disease.* Nat Rev Genet, 2008. **9**(8): p. 594-604.
20. Conaway, R.C. and I.R. Lehman, *Synthesis by the DNA primase of Drosophila melanogaster of a primer with a unique chain length.* Proceedings of the National Academy of Sciences of the United States of America, 1982. **79**(15): p. 4585-4588.
21. Sobol, R.W., et al., *Requirement of mammalian DNA polymerase-[beta] in base-excision repair.* Nature, 1996. **379**(6561): p. 183-186.

22. Hudson, G. and P.F. Chinnery, *Mitochondrial DNA polymerase-γ and human disease.* Human Molecular Genetics, 2006. **15**(suppl 2): p. R244-R252.

23. Pursell, Z.F., et al., *Yeast DNA Polymerase ε Participates in Leading-Strand DNA Replication.* Science (New York, N.Y.), 2007. **317**(5834): p. 127-130.

24. Byrnes, J.J., et al., *A new mammalian DNA polymerase with 3' to 5' exonuclease activity: DNA polymerase δ.* Biochemistry, 1976. **15**(13): p. 2817-2823.

25. Lehmann, A.R., *Translesion synthesis in mammalian cells.* Experimental Cell Research, 2006. **312**(14): p. 2673-2676.

26. Lehmann, A.R., et al., *Translesion synthesis: Y-family polymerases and the polymerase switch.* DNA Repair, 2007. **6**(7): p. 891-899.

27. Wit, N., et al., *Roles of PCNA ubiquitination and TLS polymerases κ and η in the bypass of methyl methanesulfonate-induced DNA damage.* Nucleic Acids Research, 2015. **43**(1): p. 282-294.

28. Hoege, C., et al., *RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO.* Nature, 2002. **419**(6903): p. 135-141.

29. Wang, G. and K.M. Vasquez, *Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability.* DNA Repair, 2014. **19**: p. 143-151.

30. Smith, G.R., *Meeting DNA palindromes head-to-head.* Genes & Development, 2008. **22**(19): p. 2612-2620.

31. Oussatcheva, E.A., et al., *Influence of Global DNA Topology on Cruciform Formation in Supercoiled DNA.* Journal of Molecular Biology, 2004. **338**(4): p. 735-743.

32. Frank-Kamenetskii, M.D. and S.M. Mirkin, *Triplex DNA structures.* Annu Rev Biochem, 1995. **64**: p. 65-95.

33. Sinden, R.R., M.J. Pytlos-Sinden, and V.N. Potaman, *Slipped strand DNA structures.* Front Biosci, 2007. **12**: p. 4788-99.

34. Majumdar, A. and D.J. Patel, *Identifying Hydrogen Bond Alignments in Multistranded DNA Architectures by NMR.* Accounts of Chemical Research, 2002. **35**(1): p. 1-11.

35. Herbert, A. and A. Rich, *Left-handed Z-DNA: structure and function.* Genetica, 1999. **106**(1-2): p. 37-47.

36. Zhao, J., et al., *Non-B DNA structure-induced genetic instability and evolution.* Cellular and Molecular Life Sciences, 2010. **67**(1): p. 43-62.

37. Wang, G. and K.M. Vasquez, *Models for chromosomal replication-independent non-B DNA structure-induced genetic instability.* Molecular Carcinogenesis, 2009. **48**(4): p. 286-298.

38. Cooper, D.N., et al., *On the Sequence-Directed Nature of Human Gene Mutation: The Role of Genomic Architecture and the Local DNA Sequence Environment in Mediating Gene Mutations Underlying Human Inherited Disease.* Human mutation, 2011. **32**(10): p. 1075-1099.

39. Wang, G., et al., *DNA Structure-induced Genomic Instability In Vivo.* JNCI Journal of the National Cancer Institute, 2008. **100**(24): p. 1815-1817.

40. Rooms, L., E. Reyniers, and R.F. Kooy, *Diverse chromosome breakage mechanisms underlie subtelomeric rearrangements, a common cause of mental retardation.* Human Mutation, 2007. **28**(2): p. 177-182.

41. Quental, R., et al., *Molecular mechanisms underlying large genomic deletions in ornithine transcarbamylase (OTC) gene.* Clinical Genetics, 2009. **75**(5): p. 457-464.

42. Masson, E., et al., *Co-inheritance of a novel deletion of the entire SPINK1 gene with a CFTR missense mutation (L997F) in a family with chronic pancreatitis.* Molecular Genetics and Metabolism, 2007. **92**(1–2): p. 168-175.

43. Hile, S.E. and K.A. Eckert, *Positive Correlation Between DNA Polymerase α-Primase Pausing and Mutagenesis within Polypyrimidine/Polypurine Microsatellite Sequences.* Journal of Molecular Biology, 2004. **335**(3): p. 745-759.

44. Voineagu, I., et al., *Replisome stalling and stabilization at CGG repeats that are responsible for chromosomal fragility.* Nature structural & molecular biology, 2009. **16**(2): p. 226-228.

45. Hartenstine, M.J., M.F. Goodman, and J. Petruska, *Base Stacking and Even/Odd Behavior of Hairpin Loops in DNA Triplet Repeat Slippage and Expansion with DNA Polymerase.* Journal of Biological Chemistry, 2000. **275**(24): p. 18382-18390.

46. Butler, D.K., L.E. Yasuda, and M.-C. Yao, *Induction of Large DNA Palindrome Formation in Yeast: Implications for Gene Amplification and Genome Stability in Eukaryotes.* Cell, 1996. **87**(6): p. 1115-1122.

47. Wittig, B., T. Dorbic, and A. Rich, *Transcription is associated with Z-DNA formation in metabolically active permeabilized mammalian cell nuclei.* Proceedings of the National Academy of Sciences of the United States of America, 1991. **88**(6): p. 2259-2263.

48. Youssoufian, H., et al., *Recurrent mutations in haemophilia A give evidence for CpG mutation hotspots.* Nature, 1986. **324**(6095): p. 380-2.

49. Cooper, D.N. and H. Youssoufian, *The CpG dinucleotide and human genetic disease.* Hum Genet, 1988. **78**(2): p. 151-5.

50. Pfeifer, G.P., *Mutagenesis at methylated CpG sequences.* Curr Top Microbiol Immunol, 2006. **301**: p. 259-81.

51. Halder, R., et al., *Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide.* Mol Biosyst, 2010. **6**(12): p. 2439-47.

52. Drake, J.W., *Mutations in clusters and showers.* Proceedings of the National Academy of Sciences, 2007. **104**(20): p. 8203-8204.

53. Buettner, V.L., et al., *Evidence that proximal multiple mutations in Big Blue® transgenic mice are dependent events.* Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 2000. **452**(2): p. 219-229.

54. Hill, K.A., et al., *Spontaneous multiple mutations show both proximal spacing consistent with chronocoordinate events and alterations with p53-deficiency.* Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 2004. **554**(1–2): p. 223-240.

55. Colgin, L.M., et al., *The unexpected landscape of in vivo somatic mutation in a human epithelial cell lineage.* Proceedings of the National Academy of Sciences, 2002. **99**(3): p. 1437-1442.

56. Chen, J.-M., C. Férec, and D.N. Cooper, *Closely spaced multiple mutations as potential signatures of transient hypermutability in human genes.* Human Mutation, 2009. **30**(10): p. 1435-1448.

57. Chen, J.-M., C. Férec, and D.N. Cooper, *Transient hypermutability, chromothripsis and replication-based mechanisms in the generation of concurrent clustered mutations.* Mutation Research/Reviews in Mutation Research, 2012. **750**(1): p. 52-59.

58. Wang, J., et al., *Evidence for mutation showers.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(20): p. 8403-8408.

59. Zhu, W., et al., *Concurrent Nucleotide Substitution Mutations in the Human Genome Are Characterized by a Significantly Decreased Transition/Transversion Ratio.* Human Mutation, 2015. **36**(3): p. 333-341.

60. Kondrashov, A.S., S. Sunyaev, and F.A. Kondrashov, *Dobzhansky–Muller incompatibilities in protein evolution.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(23): p. 14878-14883.

61. Chen, J.-M., et al., *Gene conversion: mechanisms, evolution and human disease.* Nat Rev Genet, 2007. **8**(10): p. 762-775.

62. Chen, J.-M., et al., *Genomic rearrangements in inherited disease and cancer.* Seminars in Cancer Biology, 2010. **20**(4): p. 222-233.

63. Chuzhanova, N., et al., *Gene conversion causing human inherited disease: Evidence for involvement of non-B-DNA-forming sequences and recombination-promoting motifs in DNA breakage and repair.* Human Mutation, 2009. **30**(8): p. 1189-1198.

64. Drake, J.W., *Too Many Mutants with Multiple Mutations.* Critical reviews in biochemistry and molecular biology, 2007. **42**(4): p. 247-258.

65. Drake, J.W., et al., *Clusters of mutations from transient hypermutability.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(36): p. 12849-12854.

66. Waters, L.S., et al., *Eukaryotic Translesion Polymerases and Their Roles and Regulation in DNA Damage Tolerance.* Microbiology and Molecular Biology Reviews : MMBR, 2009. **73**(1): p. 134-154.

67. Chen, J.-M., C. Férec, and D.N. Cooper, *Patterns and Mutational Signatures of Tandem Base Substitutions Causing Human Inherited Disease.* Human Mutation, 2013. **34**(8): p. 1119-1130.

68. Stone, J.E., S.A. Lujan, and T.A. Kunkel, *DNA Polymerase zeta Generates Clustered Mutations During Bypass of Endogenous DNA Lesions in Saccharomyces cerevisiae.* Environmental and molecular mutagenesis, 2012. **53**(9): p. 777-786.

69. Harris, K. and R. Nielsen, *Error-prone polymerase activity causes multinucleotide mutations in humans.* Genome Res, 2014. **24**(9): p. 1445-54.

70. Chen, J.M., C. Ferec, and D.N. Cooper, *Complex Multiple-Nucleotide Substitution Mutations Causing Human Inherited Disease Reveal Novel Insights into the Action of Translesion Synthesis DNA Polymerases.* Hum Mutat, 2015.

71. Chen, Z., et al., *Epidemiology of Doublet/Multiplet Mutations in Lung Cancers: Evidence that a Subset Arises by Chronocoordinate Events.* PLoS ONE, 2008. **3**(11): p. e3714.

72. Azevedo, L., et al., *Epistatic interactions modulate the evolution of mammalian mitochondrial respiratory complex components.* BMC Genomics, 2009. **10**(1): p. 266.

73. *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

74. Goldstein, J.L. and M.S. Brown, *History of Discovery: The LDL Receptor.* Arteriosclerosis, thrombosis, and vascular biology, 2009. **29**(4): p. 431-438.

75. Miki, Y., et al., *A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.* Science, 1994. **266**(5182): p. 66-71.

76. Hall, J., et al., *Linkage of early-onset familial breast cancer to chromosome 17q21.* Science, 1990. **250**(4988): p. 1684-1689.

77. Smith, S.A., et al., Allele losses in the region 17q12-21 in familial breast and ovarian-cancer involve the wild-type chromosome. Nature Genetics, 1992. **2**(2): p. 128-131.

78. Ford, D., et al., *RISKS OF CANCER IN BRCA1-MUTATION CARRIERS.* Lancet, 1994. **343**(8899): p. 692-695.

79. Clark, R.F., et al., *The structure of the presenilin 1 (S182) gene and identification of six novel mutations in early onset AD families.* Nat Genet, 1995. **11**(2): p. 219-222.

80. Sherrington, R., et al., *Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease.* Nature, 1995. **375**(6534): p. 754-760.

81. Humphries, C. and M.A. Kohli, *Rare Variants and Transcriptomics in Alzheimer disease.* Current genetic medicine reports, 2014. **2**(2): p. 75-84.

82. Campion, D., et al., *Mutations of the presenilin I gene in families with early-onset Alzheimer's disease.* Human Molecular Genetics, 1995. **4**(12): p. 2373-2377.

83. Cruts, M., et al., *Molecular genetic analysis of familial early-onset Alzheimer's disease linked to chromosome 14q24.3.* Human Molecular Genetics, 1995. **4**(12): p. 2363-2371.

84. Pützer, B.M. and M. Drosten, *The RET proto-oncogene: a potential target for molecular cancer therapy.* Trends in Molecular Medicine, 2004. **10**(7): p. 351-357.

85. Khan, N.L., et al., *Mutations in the gene LRRK2 encoding dardarin (PARK8) cause familial Parkinson's disease: clinical, pathological, olfactory and functional imaging and genetic data*. Vol. 128. 2005. 2786-2796.

86. Hernandez, D., et al., *The dardarin G2019S mutation is a common cause of Parkinson's disease but not other neurodegenerative diseases.* Neuroscience Letters, 2005. **389**(3): p. 137-139.

87. Cilia, R., et al., *LRRK2 mutations in Parkinson's disease: Confirmation of a gender effect in the Italian population.* Parkinsonism & Related Disorders, 2014. **20**(8): p. 911-914.

88. Singleton, A.B., *Altered α-synuclein homeostasis causing Parkinson's disease: the potential roles of dardarin.* Trends in Neurosciences. **28**(8): p. 416-421.

89. Long, H., et al., *Mutation Rate, Spectrum, Topology, and Context-Dependency in the DNA Mismatch Repair-Deficient Pseudomonas fluorescens ATCC948.* Genome Biology and Evolution, 2015. **7**(1): p. 262-271.

90. Markham, N. and M. Zuker, *UNAFold*, in *Bioinformatics*, J. Keith, Editor. 2008, Humana Press. p. 3-31.

91. Scully, R. and D.M. Livingston, *In search of the tumour-suppressor functions of BRCA1 and BRCA2.* Nature, 2000. **408**(6811): p. 429-432.

92. Wang, Q., et al., *BRCA1 and cell signaling.* Oncogene, 2000. **19**(53): p. 6152-6158.

93. Zheng, L., et al., *Lessons learned from BRCA1 and BRCA2.* Oncogene, 2000. **19**(53): p. 6159-6175.

94. Monteiro, A.N.A., *BRCA1: exploring the links to transcription.* Trends in Biochemical Sciences, 2000. **25**(10): p. 469-474.

95. Schuchardt, A., et al., *Defects in the kidney and enteric nervous system of mice lacking the tyrosine kinase receptor Ret.* Nature, 1994. **367**(6461): p. 380-383.

96. Sanchez, M.P., et al., *Renal agenesis and the absence of enteric neurons in mice lacking GDNF.* Nature, 1996. **382**(6586): p. 70-73.

97. Najam, O. and K.K. Ray, *Familial Hypercholesterolemia: a Review of the Natural History, Diagnosis, and Management.* Cardiology and Therapy, 2015. **4**(1): p. 25-38.

98. Sun, L.-Y., et al., *Identification of the gene defect responsible for severe hypercholesterolaemia using whole-exome sequencing.* Scientific Reports, 2015. **5**: p. 11380.

99. Braenne, I., et al., *Systematic analysis of variants related to familial hypercholesterolemia in families with premature myocardial infarction.* Eur J Hum Genet, 2015.

100. Figueiredo, M.S., et al., *High frequency of the Lebanese allele of the LDLr gene among Brazilian patients with familial hypercholesterolaemia.* Journal of Medical Genetics, 1992. **29**(11): p. 813-815.

101. Karran, E.H., et al., *Presenilins – in search of functionality*. Vol. 26. 1998. 491-496.

102. McGeer Patrick, L., T. Kawamata, and G. McGeer Edith, *Localization and Possible Functions of Presenilins in Brain*, in *Reviews in the Neurosciences*. 1998. p. 1.

103. Li, Y., et al., *Structural biology of presenilin 1 complexes.* Molecular Neurodegeneration, 2014. **9**: p. 59.

104. Pink, A.E., et al., *Mutations in the [gamma]-Secretase Genes NCSTN, PSENEN, and PSEN1 Underlie Rare Forms of Hidradenitis Suppurativa (Acne Inversa).* J Invest Dermatol, 2012. **132**(10): p. 2459-2461.

105. Takai, Y., T. Sasaki, and T. Matozaki, *Small GTP-binding proteins.* Physiological Reviews, 2001. **81**(1): p. 153-208.

106. Shen, J., *Protein Kinases Linked to the Pathogenesis of Parkinson's Disease.* Neuron, 2004. **44**(4): p. 575-577.