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<sub>Ts/Tv</sub>) 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 .................................................................................................................. 8
Tables Index .................................................................................................................. 9
Abbreviations list ......................................................................................................... 10
Introduction .................................................................................................................. 11
  The concept of mutation: a very brief history ......................................................... 11
  Types of mutations on the DNA sequence .............................................................. 12
  The human mutation rate ....................................................................................... 14
  Source of Mutation ................................................................................................. 16
  Mutational Clusters ............................................................................................... 21
Aims ............................................................................................................................... 26
Material and Methods .................................................................................................. 27
  Data mining ............................................................................................................. 27
  Data analyses ......................................................................................................... 27
Results and discussion ................................................................................................. 29
Conclusions .................................................................................................................. 39
Future perspectives ....................................................................................................... 40
Appendix ....................................................................................................................... 41
  Supplementary Files .............................................................................................. 41
  Appendix I: Detailed description of the studied genes ........................................ 42
  Appendix II: Sequence of the genes BRCA1, LDLR, LRRK2, PSEN1 and RET with
  the annotation of the concurrent MNS and the respective population distribution. ... 44
  Appendix III: Predicted structures ..................................................................... 67
References ...................................................................................................................... 81
Figures Index

Birth of multiple mutations ................................................................. 21
Mutation shower .............................................................................. 22
MNS classification ............................................................................ 22
Cluster distribution according to distance between the component mutations .......... 34
DNA sequence from LDLR-CL9, LRRK2-CL2, RET-CL7 and the respective population
distribution .............................................................................................. 36
Analyzes of the type of base change .................................................... 37
Non-B structure for BRCA-CL1 ............................................................. 40
Tables Index

Main function of the classical DNA polymerases .......................................................... 16
Summary of the sequence requirement by the common non-B structures .................. 18
Mutation clusters detected ............................................................................................ 30
Distributions of the concurrent TBS ............................................................................ 31
$R_{Ts/Tv}$ for different considered groups ........................................................................ 32
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 \textit{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.

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

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.

<table>
<thead>
<tr>
<th>Name</th>
<th>Conformation</th>
<th>Sequence Requirements</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cruciforme/Hairpin</td>
<td>Inverted repeats</td>
<td></td>
<td>TCGGTAACCGA, AGCCATGGCCT</td>
</tr>
<tr>
<td>Triplex</td>
<td>Mirror repeat symmetries in polypurine/polypyrimidine regions</td>
<td></td>
<td>AAGAGGGGAGAA, TTCTCC, CCTCTT</td>
</tr>
<tr>
<td>Slipped strand DNA DNA</td>
<td>Misalignment of simple repeats</td>
<td></td>
<td>TCGGTTCGGT, AGCCAAGCCA</td>
</tr>
<tr>
<td>Tetraplex</td>
<td>Sequence rich in Guanine</td>
<td></td>
<td>AG₃[T₂AG₃]₃, single strand</td>
</tr>
<tr>
<td>Z-DNA</td>
<td>Sequence with alternating pyrimidine and purine</td>
<td></td>
<td>CGCGGGCTGGTG, GCGCAGCGACAC</td>
</tr>
</tbody>
</table>

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.](image-url)
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).

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].
The origin of concurrent MNS results mainly from two mechanisms: gene conversion and transient hypermutability [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 in 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 (\( LDLR \ p = 2.21171 \times 10^{-26} \); \( PSEN1 \ p = 1.86026 \times 10^{-12} \); \( BRCA1 \ p = 3.13311 \times 10^{-07} \); \( RET \ p = 0.000337 \); \( LRRK2 \ p = 4.2233 \times 10^{-08} \)). 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 \( PSEN1 \) 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.

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

<table>
<thead>
<tr>
<th>TBS</th>
<th>Gene Cluster</th>
<th>Replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LDLR-CL1</td>
<td>TC&gt;CA</td>
</tr>
<tr>
<td>2</td>
<td>LDLR-CL4</td>
<td>TGCC&gt;ACTG</td>
</tr>
<tr>
<td>3</td>
<td>LDLR-CL11</td>
<td>TG&gt;CA</td>
</tr>
<tr>
<td>4</td>
<td>LRRK2-CL1</td>
<td>GC&gt;AT</td>
</tr>
<tr>
<td>5</td>
<td>LRRK2-CL2</td>
<td>TC&gt;AT</td>
</tr>
<tr>
<td>6</td>
<td>LRRK2-CL6</td>
<td>GA&gt;TT</td>
</tr>
<tr>
<td>7</td>
<td>LRRK2-CL12</td>
<td>CC&gt;AT</td>
</tr>
<tr>
<td>8</td>
<td>LRRK2-CL13</td>
<td>CA&gt;AT</td>
</tr>
<tr>
<td>9</td>
<td>PSEN-CL5</td>
<td>GC&gt;TT</td>
</tr>
<tr>
<td>10</td>
<td>RET-CL2</td>
<td>GA&gt;TG</td>
</tr>
<tr>
<td>11</td>
<td>RET-CL3</td>
<td>TG&gt;GA</td>
</tr>
<tr>
<td>12</td>
<td>RET-CL4</td>
<td>GG&gt;TT</td>
</tr>
<tr>
<td>13</td>
<td>RET-CL5</td>
<td>GA&gt;TG</td>
</tr>
</tbody>
</table>
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).
Identification and analyzes of concurrent multiple nucleotide substitutions

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 interesting 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 opposite is verified, where the concurrent MNS only appears 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 is responsible for most of the TBS [67]. It is important to clarify that TBS distribution ahead from the pathologic mutational spectrum was accessed. In this study, we used the polymorphic variation present in the entire sequence of five genes to demonstrate the occurrence and frequency of TBS across human populations (Figure 5).

The complex MNS (mutations that have noncontiguous and contiguous 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).
Identification and analyzes of concurrent multiple nucleotide substitutions

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].

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 Noncontiguous 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 Noncontiguous 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 ($\leq$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 is 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 $R_{Tv/Tv}$ values for 133 Concurrent MNS and 4933 Genome-wide de novo SNS mutations, 28 TBS and 69 Noncontiguous MNS and 79 concurrent MNS with a separation $\leq$100 nucleotides and 24 concurrent MNS separated by more than 100 nucleotides. The $R_{Tv/Tv}$ 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.

<table>
<thead>
<tr>
<th>Category</th>
<th>$N$</th>
<th>$R_{Tv/Tv}$</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome-wide de novo SNS mutations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concurrent MNS (All)</td>
<td>4933</td>
<td>2.1 (3344/1589)</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>103</td>
<td>1.1 (54/49)</td>
<td></td>
</tr>
<tr>
<td>TBS</td>
<td>28</td>
<td>0.75 (12/16)</td>
<td>0.0486</td>
</tr>
<tr>
<td>Noncontiguous MNS</td>
<td>69</td>
<td>1.76 (44/25)</td>
<td></td>
</tr>
<tr>
<td>$\leq$100 nucleotides</td>
<td>79</td>
<td>1.03 (40/39)</td>
<td>0.3352</td>
</tr>
<tr>
<td>$&gt;$100 nucleotides</td>
<td>24</td>
<td>1.4 (14/10)</td>
<td></td>
</tr>
</tbody>
</table>
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.

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 role 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**

*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]. Innumerable 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 TAGACTACAGTGAGCTATGATCACTCCACTGCAGTCAGCGTGGGCGGCAAAGC 41254701
**LDLR**

**CL1**

11200818  CGCCCGCCGGGACCCTCCTTCGAACTCGGAGCTGGGTGTGGGGACCTCCAGTCCTAAA  11200877

**CL2**

11204838  TGAGATGAGTTTTGCTTTGCTGCCAGGCTGGAGTGAATGGGCAAGCTCTCTCAAGCTCGCTCAC  11204897

11204898  CGCAACCTCCACCTCGTTCAAGCTCTCTCTGCCTCACGCTCCTGAGTAGCTGGGA  11204957
CL3
11204898 CGCAACCTCCACCTCTGTTCAAGCGATCTCTGCTCCTAGCCTCTGAGTAGCTGGGA 11204957

CL4
11212218 CACTGCAAGCTGTGCCCTGGTTCAAGCGATCTCTGCTCCTAGCCTCTGAGTAGCTGGGA 11212277
Identification and analyzes of concurrent multiple nucleotide substitutions

CL5

11212818  GCGTGAACCAGGAGGCAGGCTTGAGCTGAGCTGATCGGACTTGCAGTGAGCCAAGATCGTGCCACTGTACTCCAGCCT  11212877

CL6

11214558  CGATAGAAAAATGCAAGTTAGCTGACAGCAACCTCGGCATTATAAGTACTGACAAAGCCTCTTCAGCGCACAGCACAAGCACCAT  11214617
11214618  CCTCTTCAGCGCACAGCACAAGCACCATTCTATAAAAATCTCCAGCAAGCGCAGGCGAGGTCA  11214677
CL7

11216958 ACGATTATGGCTCAGCCTTGGGCTCAAGATCTCTCAGCCTCAGCC 11217017

11217018 TCTCAAGCAGTGGAAACCACAGCTGTACCATCAGCTGCTTATTTTTGTACAGAC 11217077

CL8

11217918 ACAGGCAACACACCCGTGCCAAGCGTTTTCTTAATGAATCCATTTGCATGCGTTCTT 11217977

11217978 ATGTGAATAACTATATTATATGATGAGTGCCAAAACCTGTAGGCTCAGACACACGTCAGACACACTGAC 11218037
Identification and analyzes of concurrent multiple nucleotide substitutions

CL9

11218218 AGTTTGTGGGACGACAGACATGAGCTGTAGGGTTTTGGGAACTCC
11218277

11218278 ACTCTGCCAACCTGTGCAAGGGGCTCTCCTTCACTTTGAGACAGTCGCGACGGTGC
11218337

11218338 CCCAGCTGGGCAATAGCCGAGTTCCTCGGCTCAGTGCAACCTCTGCTCCCAAGGTCA
11218397

CL10

11222058 ATTACATCTCCGAGAGGCTGGGCTGTCTTGGCTGCCTTCGAAGGTGGGTTTTGGC
11222117

11222118 TAGGGCCCATCAGCCTCGTCTCTAGCCATTGGGGAGACCCCCAAGGCTCCCGAC
11222177

11222178 TCTCTTCCAGATATCGATGAGTCCAGGATCCCGACACCTTCAGCCAGCTGCTGTA
1122237
Identification and analyzes of concurrent multiple nucleotide substitutions

**CL11**

11222958  TGTATTTTAGAGACGGGTTTCAAGGTTAGCCAGGATGCTCGATCTCCTGAC  11223017

**CL12**

11223798  CCTGGCGGTGAGGTCTTTTCCACCTCTTTTTCTGGTGCTGCTT  11223857
CL13
11232918  CGCCCTACCTACGGCCGCTACTTTTTTGTATATTTAGTAGAGATGGAGTTTCACTGTG  11232977
11232978  TTAGCCAGGATGGGTCTGATCTCGATCTCCCTGACCTTTGTGATCCGCCGCTCGCCCTCCCAAAGT  11233037
11233038  GCTGGGATTACAGGCCTGAGCACAAGCCTCGGCTTTTTTTTTTTTTTTTTTTTTTGAGA  11233097
**LRRK2**

**CL1**

40604826 ACCACCTGTGACCTGGAAGCCCCCTCCCGAGTCCAAGTTGCTTAGGGAAAAATAAA 40604885

**CL2**

40605126 TCAAGGCCAGGACTATCATGGAAGGTGGGTGTGAGAATGTAAGTCAAT 40605185
Identification and analyzes of concurrent multiple nucleotide substitutions

CL3
40606626  TGTAGTGTTCTTAACTTTAGCAGTTATTTGTATAAGTGCAGTGAAAATCTGTT  40606685

CL4
40634106  TAGCTGTGTCTCTTTTAGGAACATCTTCTAGGAACCTGCTTCACAGAAATATTTTTTT  40634165
40634166  ATTTAAGGAGATTACACTTGATGTATCTCACACAACTATAATGAATATTGTAATTTTGA  40634225
CL5

40641246 TATATAAAATATGTATATAATATACCGTTATATAAATATGTATATAAATATATAC 40641305

CL6

40667166 AGATGATAATGAAATGATGTCCAAGCTGAGCAATTAAAGTGTGAAGTAGAAGGACACAG 40667225
CL7
40675266 GATGAGGAGAAAAGATTGCAGGGATAAGAAATTATGCTTTTGATAATCTTTAGTTATAT 40675325
40675326 TCTTGATTTTCTTACATTATTATTAAATGTAAAAAATATCTCTGAGCGATAGTTTTT 40675385

CL8
40677546 TTGCCTTATTTTATTTGTTTCTATCCTTAAATTGAGATGTAGAGAAAAATCACATGAAGTGATTGAAGT 40677605
40677606 TTGATTTGAGCTCTCCTAAAAGGAAGAAAAATGTAGATTTTATAATATTTTTTTTTTTT 40677665
Identification and analyzes of concurrent multiple nucleotide substitutions

CL9
40688166 TGAAGACCTGTTCTAACCTATTTCCATTTGAATTATAGCTGACTACTAAAATTGA 40688225
40688226 GGGTTTTGTGTTAAATTCTAGATCTTAAAGATGGGTGAATGACGACTTGAGTTGAAT 40688285

CL10
40723146 TGCTATGTGAGATGAGGAAAATTAACGCTATTCTTTCTCCTTTTCCCA 40723205
40723206 GTCCTTTAATTGTTCTATTATTTATGTAGGAAATTTATATTTATTTCTGTT 40723265

LRRK2-CL9

LRRK2-CL10

Population
Identification and analyzes of concurrent multiple nucleotide substitutions

CL11
40730406 CTTCTTTCTTTCTGAAATGCTATGAAATGCTTTTAGTAGATCCAGAAACTCTCT 40730465
40730466 TCCTGAAAGGGTGGCTCAACTACTGAAAACGTACAGAATGAAATGAAACAGGGGCA 40730525
40730526 .......................................................... 40746785
40746786 CAGACGGGTGCCAGCCGGCAGAGGGGCTCCTCACTTCCCAGAAGGGGCGGCCGG 40746845
40746846 AGGCACCCCCACCTCCACAGACGGGCACCCGCCGCCGGCGCCGGGGCTGCCCCACCTCCCG 40746905
40746906 GATGGGATTGGGCTCCAAGGCAGCTCTCTCACTTCCAGACCGGCGCTGCTGGCC 4074696

CL12
40738746 GTGTAGTAGGGTTATCATCATACTCAATGATCTCCATCCTCCT 40738805

LRRK2-CL11

LRRK2-CL12
Identification and analyzes of concurrent multiple nucleotide substitutions

**CL13**

40746306  CGCAGAGGGGATTTGCAAGGGTCATAGGACAATAATGGAGGGAAGGTCAGCAGATAAAC 40746365

**CL14**

40762986  TTCATTGTTACTTTGTATTTGCAATTTTTTTACAAAGACAAATTAAAAAAATGAATAC 40763045

40763046  CAGTTTTAATGGAATAATAAGGTTTTTAAAAACTTTAA 40763087

**PSEN1**
Identification and analyzes of concurrent multiple nucleotide substitutions

73620886 TAAACACAACATGTATACCTGTATATATTTGAAAAACGTAGGAGTGGTGTCG 73620945
73620944 ………………………………………………………………………………………………………….. 73621945
73671946 AGGCTGAGGCAGGGATACGCAGGAGAGATAGACCATCCTGGCTAACAGTGA 73672005
73672006 AACCCACACCTACTAAATACAAAAATTAGCCAGGCATGTTGCCGCGGCCTGT 73672065
73672066 AGTCCAGCTACTCGGGAGGCTAGGAGAAATTGGCGTGAAACCAGGACACGGAGCTT 73672125

CL4

73622626 ATGTAGGTTCTCATCTGTGTAACGGTTTACCACAGTG3GCAGGATTTAGAGCTGT 73622685
73622686 TGGAGGGGTATGTGGGATACGGTTTACCACAGTGGTGCAGGATGTTGATAGAGGCTGT 73622745
73622746 TGCATTTAAAATATTTTATGCTGGTGCTGCTCAGATATGCTACTTTTGGCACTTTGGAAAG 73622805
73622806 GCTGAGGCTAGGTTTACCACAGTG3GCAGGATTTAGAGCTGT 73622865

PSEN-CL4

Number of alleles

| Population | AFR | AMR | EUR | SAS | AFR | AMR | EUR | SAS |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|            |     |     |     |     | 8   | 1   | 1   | 1   |
|            | 4   | 3   | 3   | 3   | 3   | 3   | 3   | 3   |
|            | 2   | 2   | 2   | 2   | 2   | 2   | 2   | 2   |
|            | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1   |

ACB ASW ESN GWLW MSL YRI CLM
CL5
73657966  CTACTCTCTCTCACACTCTTTGGATAATGGGATCTAAATGGGATCTATATTTTTT  73658025

CL6
73663786  ACGCCACCACACACCAGCTATTTTTGTATTTTTTAGTAGACAGGGTTTCACCGTG  73663845
73663846  TTAGCG5AGATGCTCTGATCTCTGACCTTGATCAGCCACCCACCTCGCCCTCCAAAGT  73663905

CL7
Identification and analyzes of concurrent multiple nucleotide substitutions

**PSEN-CL7**

![Chart](chart1.png)

**RET-CL1**

![Chart](chart2.png)

**RET**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>73670746</td>
<td>GCCCAGGCTGGAGTGCAATGGGACAGCTCACCAGCCCTCCGGCTCTCTGGATTC</td>
</tr>
<tr>
<td>73670806</td>
<td>AAGCAATTCTCTCTCCTCCTCAGCCCTCCGGAGTGCCGAGGATATAGACATGGACACCACAC</td>
</tr>
<tr>
<td>73670866</td>
<td>TCTTCCTCCTCCTCGACCTGAGAATGGCAGTGAGGGGCACGGTCTGACAGCTCCCAATGGGATCTTCGAGCCTGAGGATG</td>
</tr>
<tr>
<td>73670926</td>
<td>AGAATGAGTTGTGGACAGCTGAGGAGGAGGAGAAGGCTTTTATCTGAGGCA</td>
</tr>
<tr>
<td>73670986</td>
<td>TTCTCTCTCTTCACCTGCAACGTGGCAAGTGGGGGGCACGTTCAGCCTGTTTGT</td>
</tr>
<tr>
<td>73671046</td>
<td>CTACGTTACACGTAATTAGTGGCTCCTCTACTTTTGCTTCTCAGGACCAGCAGCAGCATATTCTCTCTCTGAG</td>
</tr>
<tr>
<td>73671106</td>
<td>AAAAAAAAAAGTGAAACTATCAAGTATTGCTCCTGAACTTCG</td>
</tr>
</tbody>
</table>

**RET**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>73670746</td>
<td>GCCCAGGCTGGAGTGCAATGGGACAGCTCACCAGCCCTCCGGCTCTCTGGATTC</td>
</tr>
<tr>
<td>73670806</td>
<td>AAGCAATTCTCTCTCCTCCTCAGCCCTCCGGAGTGCCGAGGATATAGACATGGACACCACAC</td>
</tr>
<tr>
<td>73670866</td>
<td>TCTTCCTCCTCCTCGACCTGAGAATGGCAGTGAGGGGCACGGTCTGACAGCTCCCAATGGGATCTTCGAGCCTGAGGATG</td>
</tr>
<tr>
<td>73670926</td>
<td>AGAATGAGTTGTGGACAGCTGAGGAGGAGGAGAAGGCTTTTATCTGAGGCA</td>
</tr>
<tr>
<td>73670986</td>
<td>TTCTCTCTCTTCACCTGCAACGTGGCAAGTGGGGGGCACGTTCAGCCTGTTTGT</td>
</tr>
<tr>
<td>73671046</td>
<td>CTACGTTACACGTAATTAGTGGCTCCTCTACTTTTGCTTCTCAGGACCAGCAGCAGCATATTCTCTCTCTGAG</td>
</tr>
<tr>
<td>73671106</td>
<td>AAAAAAAAAAGTGAAACTATCAAGTATTGCTCCTGAACTTCG</td>
</tr>
</tbody>
</table>
CL1

43583575 TGATGGTGCTGGAAGCCAGTCCGGTGCTGGTGGCTGGGGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGTGCTGGTGGAGGCAACGGTGATGGT

43583635 TGGTGATGGTGAGGAATGGTGGTGGGAATGGTGGTGGTGGAGGCAATGGTGGTGGTGATGG

CL2

43583635 TGGTGATGGTGAGGAATGGTGGTGGGAATGGTGGTGGTGGAGGCAATGGTGGTGGTGATGG

43583694
Identification and analyzes of concurrent multiple nucleotide substitutions

CL3
43583635  TGGTGATGGTGAGACGGTGGTGGCAATGGTGGTGGTGGAGGCAATGGTGGTG
43583694

CL4
43590235  GCTTGAGCGGCTATGCTTGTAGTGACATGGACATGGGAGAAAATGTATGTAT
43590294

RET-CL4

<table>
<thead>
<tr>
<th>Population</th>
<th>MSL</th>
<th>AFR</th>
<th>AMR</th>
<th>EAS</th>
<th>EUR</th>
<th>SAS</th>
</tr>
</thead>
</table>
CL5
43590895 AGGCCGCTGGGCATGAGAA
43590954 CATGCACACCTGGACGGAGTGTGTGGATTGACAAGGTGT

CL6
43599295 GACATGTTCTGCTTGGGAGGCTTGTGGAGTTGCCCCCGCCTCCTTGGGGTGGGTTGTG
43599354 GGGTCTTCAGCGGTGACCTTCCGCCTTCAGCATGCCCCCTTTTCTGGATTTTCTGGAG
43599414

CL7
43599295 TCTTCAGCGGTACCTCGGCTTCAGCACGCTGCCCCCTTTTCTGGATTATTTTCTGGAG
43599414
Identification and analyzes of concurrent multiple nucleotide substitutions

CL7
43618495 CGCCCAGTGACCTGTGCTCCGTGTGACATGTATGCTTTTT
43618554

CL8
43625275 ACGTAACCTGGCTCTAATTTGGGCTGTTTTTCAGATACACTGATAAGTTCTTTTACAA
43625334

Subtitle:
- Intron
- Exon
- SNP
Appendix III: Predicted structures

BRCA1 CL1

LDLR CL2

LDLR CL3
Identification and analyzes of concurrent multiple nucleotide substitutions
Identification and analyzes of concurrent multiple nucleotide substitutions

LDLR CL8

LDLR CL9
Identification and analyzes of concurrent multiple nucleotide substitutions

LDLR CL10

LDLR CL12

LDLR CL13
Identification and analyzes of concurrent multiple nucleotide substitutions
Identification and analyzes of concurrent multiple nucleotide substitutions

LRRK2 CL5

LRRK2 CL6

LRRK2 CL7

LRRK2 CL8
Identification and analyzes of concurrent multiple nucleotide substitutions

LRRK2 CL9

LRRK2 CL11 PART1

LRRK2 CL11 PART2
Identification and analyzes of concurrent multiple nucleotide substitutions
Identification and analysis of concurrent multiple nucleotide substitutions
Identification and analyzes of concurrent multiple nucleotide substitutions

PSEN1 CL4
Identification and analyzes of concurrent multiple nucleotide substitutions
Identification and analyzes of concurrent multiple nucleotide substitutions

RET C11
References

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.


