A comparison of the number of SNPs and mutations with synonymous (K_s) and nonsynonymous (K_a) substitution rates in human immunome

Master's thesis
Mesue Nicholas Kolle
Institute of Medical Technology
University of Tampere
September 2009

Acknowledgements

I am most grateful to the Almighty God for guidance and strength throughout the entire writing process especially in moments I took ill and felt like to continue no more.

I thank my supervisors Csaba Ortutay and Mauno Vihinen for accepting to be my mentors during the writing of my thesis. To Mauno Vihinen, I thank you so much for your critical comments. In a very special way, I am indebted to Csaba Ortutay for granting me audience whenever I needed help. It was nice working with you.

A big thank you to the administrators and my lecturers both from the University of Tampere and Turku especially Martti Tolvanen for the technical support and assistance to get me started on my first day of arrival in Tampere.

To all my colleagues, Ayodeji Olatubusun, Teku Gabriel, Ashok Aspartua and Emmanuel Ojefua, thank you for the assistance in both education and community life . It was really nice being with you on the programme.

To my family, thank you so much for the consistent words of encouragement and prayers that have strengthened me spiritually, mentally and otherwise for the time that I have lived abroad.

Tampere, September 2009,

Mesue Nicholas Kolle

MASTER'S THESIS

Place:	UNIVERSITY OF TAMPERE Faculty of Medicine Institute of Medical Technology
Author:	Mesue Nicholas Kolle
Title:	A comparison of the number of SNPs and mutations with synonymous (K_s) and nonsynonymous (K_a) substitution rates in human immunome
Pages:	57 pages
Supervisors:	Csaba Ortutay, PhD and Mauno Vihinen, Professor
Reviewers:	Professor Mauno Vihinen and Csaba Ortutay, PhD
Date:	September 2009

Abstract

Background

Changes that occur in a nucleotide sequence of a gene are known as mutations. Mutations in general and single nucleotide polymorphisms (SNPs) in particular are the major driving forces of both genetic variation/evolution and genetic diseases in humans and other organisms**.** An understanding of the evolutionary pattern of genes and proteins related to the human immune system (human immunome) is of prime importance due to the fundamental role they play in preventing pathogens from invading host organisms. The values of nonsynonymous substitution (K_a) and synonymous substitution rates (K_s) give us a clear picture into the evolution of the human immunome. However, since our knowledge of mutations is increasing day by day, estimating these rates in order to understand human immunome is very essential.

Methods

I collected four datasets, gene2RefSeq and HomoloGene from EntrezGene database, SNPs and mutations from Immunome Knowledgebase (IKB). In addition, I used a data file consisting of 874 human immunome genes collected from IKB. I used Perl/bioperl modules to download GenBank files for both human and mouse orthologs, picked up the coding sequences, compared them with the standard GenBank's, translated them, generated cDNA sequences using their protein sequences as a guide, aligned them globally and then estimated K_a and K_s rates for each ortholog pair.

Results

In a total of 755 human immunome genes, the mean nonsynonymous substitution rate (K_a) = 0.178 (0.158), mean synonymous substitution rate (K_s) = 0.685 (0.169), mean K_a/K_s = 0.394 (0.488) and mean Z-score = -13.15 (7.873). Most SNPs occurred in the intronic regions 123,265 (80.04%). Missense mutations had the highest frequency 1,878 (46.074%). The highest correlation was observed between Z-score and the number of coding synonymous SNPs ($r = -0.47$, $p < 2.2e-16$). Interestingly, the number of SNPs is associated with K_s and Z-score ($r = -0.116$, $p = 0.001$; $r = -0.37$, $p < 2.2e-16$) respectively.

Conclusion

Pooling ideas from the K_a, K_s and K_a/K_s estimates, human immunome genes are highly conserved at the protein level. Less than 3.3% of these genes evolving quickly, suggests a possibly adaptation of these genes. A strong evidence of a negative correlation between Z-score and number of coding synonymous SNPs despite a moderate correlation, suggests a biological relevance between these variables which is worth seeking, and interpreting.

CONTENTS

Abbreviations

Abbreviations

3'	3 prime; downstream of a DNA sequence
5'	5 prime; upstream of a DNA sequence
Bio::Align::DNAStatistics	Module for multiple alignments and calculation of K_a, K_s and Z-score statistics
Bio::Align::Utilities	Module for generating cDNA sequences
Bio::AlignIO	Module for reading and write multiple alignments files
Bio::Factory::EMBOSS	Module for multiple global alignments
Bio::SeqIO	Module for reading and write out files
bps	Base pairs
cDNA	Complementary DNA
CDS	Coding sequences
CNP	Copy number polymorphism
cSNP	SNP that appears in the coding regions
csv file	Comma separated value files
DNA	Deoxyribonucleic acid
gbk	GenBank
GI	GetInfo Identifier
gSNP	SNP that appears in the gapped (intergenic) regions
HGP	Human Genome Project
HIVP	Human Immunome Variome Project
hMLH1	Human mismatch repair gene
hMSH2	Human mismatch repair gene
IKB	Immunome Knowledge base
indel	Insertions and deletions
iSNP	SNP that appears in the intronic regions
K_a	Nonsynonymous substitution rate
K_s	Synonymous substitution rate
MEGA	Molecular Evolutionary Genetics Analysis
MHC	Major histocompatibility complex

mRNA	Messenger RNA
mRNA-utr	Messenger RNA at the untranslated regions
NCBI	National Center for Biotechnology Information
ORF	Open reading frame
Perl	Practical Extraction and Reporting language
r	Pearson correlation coefficient
R	An open software for programming and data analysis
RFLPs	Restriction fragment length polymorphisms
RNA	Ribonucleic acid
RpII	RNA polymerase II
rSNP	SNP that appears in the regulatory region
RSV	Respiratory syncytial virus
sd	Standard deviation
SNP	Single nucleotide polymorphism
sqrt	Square root
sSNP	SNP that appears in the silent regions
SCFBIO	Supercomputing Facility for Bioinformatics & Computational Biology
STAT	Signal transducer and activator of transcription
Tpi	Triosephosphate isomerase
UTR	Untranslated region

One-letter code of amino acids

A	alanine
C	cysteine
D	aspartate or aspartric acid
E	glutamate or glutamic acid
F	phenylalanine
G	glycine
H	histidine
I	isoleucine
K	lysine
L	leucine
M	methionine
N	asparagines
P	proline
Q	glutamine
R	arginine
S	serine
T	threonine
V	valine
W	tryptophan
X	any amino acid
Y	tyrosine

Introduction

The ability of an organism to grow is one of the fundamental characteristics of a living thing. During this vital process, an organism undergoes both chemical and morphological changes that stem from changes that occur at the molecular level of that organism (Klug *et al.,* 2006). Usually, the growth process is initiated by chemical signals that the cell receives either from within or its environment. These signals trigger cell replication. A lot of chemical changes are observed at the S phase of cell cycle during DNA replication. Most of these changes are repaired by the DNA repair machinery and those that cannot be repaired by this mechanism, result to what is known as variations. Worthy of note is the fact that this project uses the term mutations when they cause diseases rather than mutations in the global sense. In a population, changes observed at the phenotypic level due to copying errors at the genotypic level result to polymorphisms. In particular, SNPs and mutations at the genotypic level are said to have a vital role in the evolution of genes and proteins. Proteins being conserved in nature due to their unique role they play in the functioning of an organism, and our knowledge of SNPs increasing day by day, the estimation of K_a and K_s to assess the evolution of genes and proteins have been over looked for the past two decades and a half, especially genes and proteins related to the human immune system (human immunome).

The immune system is a complex biological system whose functions depend on the action of many genes and proteins (Ortutay *et al*., 2006). The primary function of the immunune system is to avert pathogens from invading host organisms. Alterations of these genes and proteins escalate an organism's susceptibility to infections (Storey *et al*., 2008). To better understand the interplay between these molecules, a reference set of 874 essential human genes and proteins (human immunome) were identified, annotated and made available at http://bioinf.uta.fi/immunome (Ortutay *et al.,* 2006). Though most proteins are naturally conserved, very few may evolve slightly due to mutations or adaptations. To compare the number of SNPs and mutations with synonymous (K_s) and nonsynonymous (K_a) nucleotide substitution rates in human immunome, I collected 874 human immunome genes from IKB, HomoloGene (which consists of homologs) and gene2refseq (which consists of a reference set of genes and proteins) datasets from

EntrezGene database, and two other datasets (single nucleotide polymorphisms and disease causing mutations) from the Immunome Knowledge Base at http://bioinf.uta.fi/IKB/.

1.1 Brief introduction to DNA

DNA is a macromolecule that contains biological instructions. It is located in the nucleus of a cell. These instructions are unique and are vital to the development and functioning of a living organism. These biological instructions are divided into discrete units called genes. A complete set of these genes constitutes what is known as the genome (nuclear DNA). During sexual reproduction process, the genetic material in the DNA is passed on from parents to offsprings. The DNA also contains unique information for the processing of other macromolecules such as RNA and proteins (Klug *et al*., 2006).

A cell is the smallest, structural and functional unit of a living organism. It consists of many parts. Each part plays an important role for the well-being of the cell. The nucleus, located almost at the centre of a cell and mitochondria are some of them. Most of the information/instructions in the DNA are stored in the nucleus of a cell while some are located in the mitochondria. Instructions found in the nucleus are called nuclear DNA and those found in the mitochondria are called mitochondria DNA. Due to the lengthy nature of DNA molecule and the microscopic nature of a cell, the DNA is packed tightly into structures known as chromosomes (Klug *et al.,* 2006). The structural organization of DNA stemming from the nucleus in a cell, to the packing into chromosomes and then to the double helix structure is illustrated in figure 1.1.

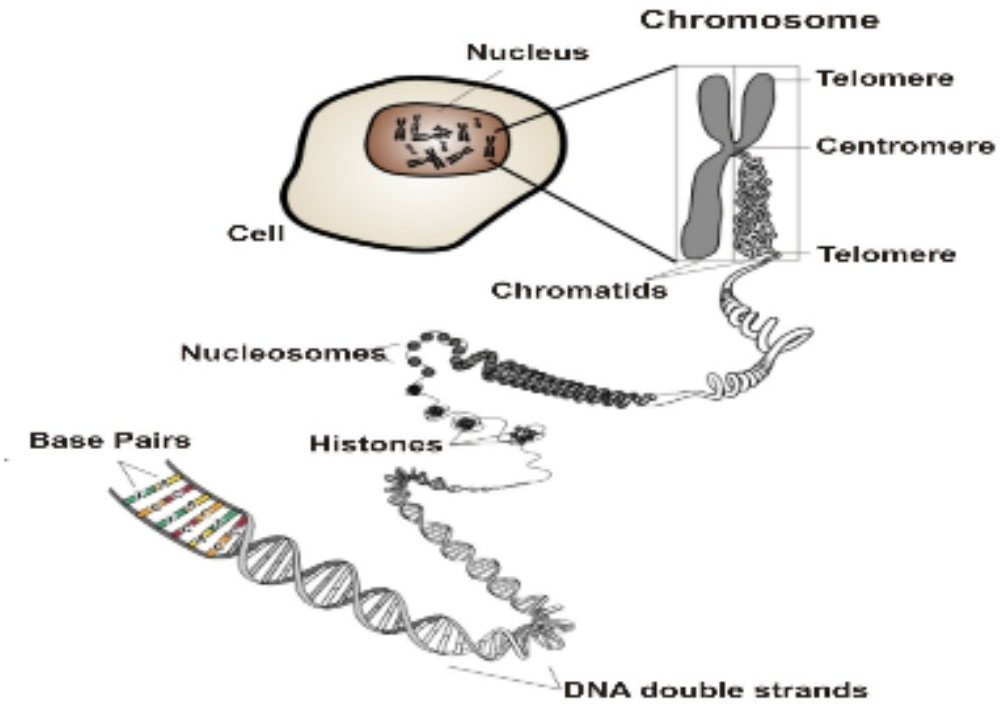

Figure 1.1 Structural organization of DNA in the nucleus of a cell.
Source: Wikipedia, September 16th, 2009. Adapted from (http://fr.wikipedia.org/wiki/Fichier:Chromosome_fr.svg)

DNA is made up of four Nitrogen bases namely: Adenine (A), Guanine (G), Cytosine(C) and Thymine (T). RNA is made up of four Nitrogen bases as well namely: Adenine (A), Guanine (G), Cytosine(C) and Uracil (U). Adenine and guanine are purines while cytosine, thymine and uracil are pyrimidines. Purines consist of a double benzene ring while pyrimidines consist of a single benzene ring. The structures of these bases, bonding types and their associated compounds are illustrated in Figure 1.2. Chemically, DNA is made up of a sequence of the four nitrogen bases called a strand. DNA exists as a double helical structure consisting of a leading and a complementary strand. For example, the two DNA strands ATTGAT and ATTGTA contain separate instructions. Their corresponding complementary strands are TAACTA and TAACAT respectively, consisting of separate instructions as well.

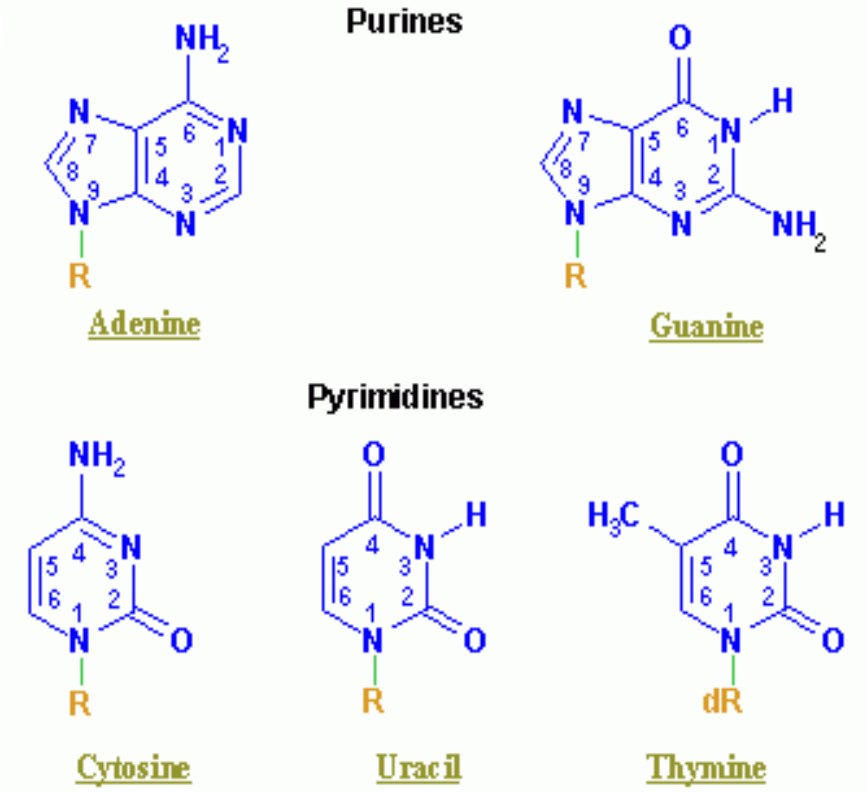

Figure 1.2 The four bases of a DNA molecule.
Source: Encyclopedia of creative science, September 16th, 2009.

DNA has a double helix chain structure. The two chains are held together by hydrogen bonds. Adenine pairs with thymine with two hydrogen bonds (weak bonds) while guanine pairs with cytosine with three hydrogen bonds (strong bonds). The backbone of a DNA molecule consists of a sugar and a phosphate molecule that are covalently bonded together by phosphodiester bonds. The 5' terminal of the molecule is linked to the phosphate (P) group while the 3' terminal is linked to the hydroxyl (OH) group. The base paring, bond type and the direction of a double helix structure of a DNA are illustrated in Figure 1.3.

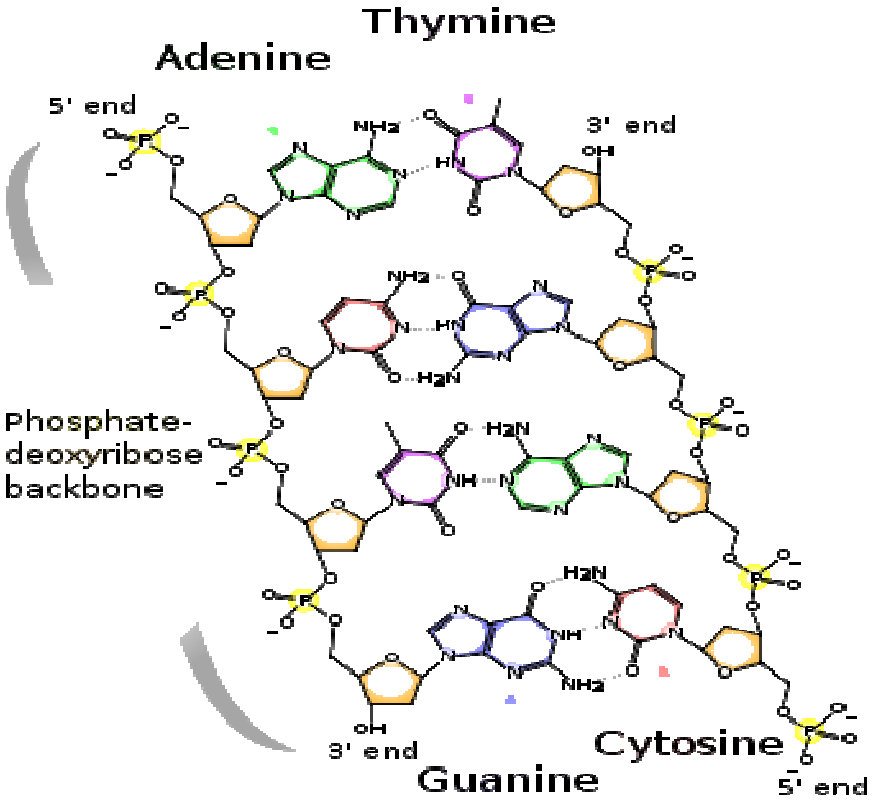

Figure 1.3 Chemical structure of DNA structure showing complementary base pairing, bond type and direction. Hydrogen bonds are shown as dotted lines.
Source: Wikipedia, September 16th, 2009.

1.2 Mutations

A lot of chemical changes are observed at the S phase of cell cycle during DNA replication. Most of these changes are repaired by the DNA repair machinery and those that cannot be repaired by DNA repair machinery result to what is known as variations. Variations occurring in a nucleotide sequence could either lead to mutations (abnormal variations) or polymorphisms (different forms of the same organism). Variations that result to permanent changes of a nucleotide sequence at the DNA level are called mutations and those that result to normal variations amongst organisms in a randomly mating population are known as polymorphisms

(Condit *et al.,* 2002). These variations account for slight differences that are observed amongst individuals at the phenotypic level due to allelic differences that had occurred at the genotypic level of the individuals. Examples of phenotypic differences that could occur in a population are the human blood group system, physical make up of organisms, eye and hair color in humans. Changes in a nucleotide sequence that do not lead to a normal variation between living organisms in a population are referred to as mutations. Given other potential factors of genetic variation such as sexual reproduction, outbreeding and diploidy, mutation is the ultimate source of genetic variation at the DNA level (Klug *et al*. 2006). It is both the substrate of genetic variation and the root cause of genetic diseases that can lead to cancer and cell death (Nachman *et al.*, 2000; Huppke *et al*., 2002).

Generally, mutations can be divided into two main categories namely, somatic (acquired) mutations and germ line mutations. Mutations that occur in the cells of the body are called somatic cell mutations. Such mutations are not inherited by the descendants and thus, do not affect evolution. Somatic mutations arise as a result of either copying errors during DNA replication or environmental factors. Mutations that can be passed on to the descendants during reproduction are called germ line or hereditary mutations. Such mutations are present in the egg and sperm cells of organisms that reproduce sexually. Mutations that occur in the egg and sperm cells after fertilization that do not have a family history are called new (de vivo) mutations. Mutations range in sizes from single nucleotide base (point mutations) to chromosomal mutations (Sun *et al*., 2009).

Mutation rates reflect the recent evolutionary divergence and human nucleotide diversity (Stamatoyannopoulos *et al*., 2009). The rate is thought to vary across the human genome on several different scales. At the chromosomal level, the Y chromosome evolves faster than the X chromosome (Hodgkinson *et al*., 2009; Miyata *et al*., 2009). This is because the CpG sites are highly methylated, thus resulting in a high transition rate (Hodgkinson *et al*., 2009). Taking into account the different types of mutations that could be present at the nucleotide level and an average generation time of about 25 years, it was estimated that the average mutation rate per nucleotide in human was between 1.3×10^{-8} and 3.5×10^{-8} per sites (Nachman *et al.*, 2000).

1.2.1 Factors that could predispose mutations

Mutations can be triggered by two major factors namely: internal or environmental. Internal factors such as copying or repair errors arising during DNA replication and repair processes respectively, could lead to mutations (Klug *et al.* 2006). In addition, environmental factors such as

i. the exposure of a cell to ultra violet rays and other radiations
ii. the exposure of a cell to viruses and chemicals

could trigger or catalyze mutations in the human genome and other genomes (Nelson *et al.,* 2005).

1.2.2 Types of base changes

There are essentially two types of base changes that can occur in human or other genomes. These two categories are

i. transitions
ii. transversions

A transition is a swop either between the purines or between the pyrimidines whereas; a transversion is a swop either between the purines and the pyrimidines or vice-versa. Figure 1.4 illustrates how these base changes occur.

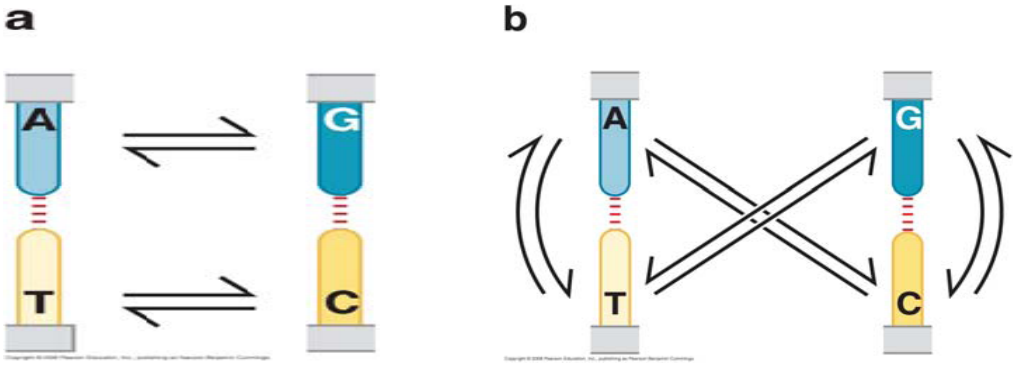

Figure 1.4 a) Transition mutations: Purine to purine or Pyrimidine to pyrimidine b) Transversion mutations: pyrimidine to purine or purine to pyrimidine

Source: Lehninger, Principles of Biochemistry, 2005: 4th edition, by Nelson and Cox.

1.2.3 Mutation nomenclature

There is ample need to document a standard and unequivocal way of reporting mutations due to the fact that our knowledge of mutations is increasing day by day. Consistent gene mutation nomenclature is essential for efficient, accurate reporting and testing of the disease causing mutations and polymorphisms occurring in an organism's genome (den *et al.,* 2000). Therefore, standards for communicating variants and mutations that occur in an organism's genome in an unequivocal and easy fashion are worth putting in place (Ogino *et al*., 2007). To this end, though a committee was formed to suggest standards for the description of sequence variants in DNA, RNA and protein sequences (Antonarakis *et al.*, 1998; den *et al.*, 2003), additional suggestions needed to be made on the nomenclature of complex mutations (den *et al.,* 2000). When talking about a mutation, it is of prime importance to mention the level at which such a mutation is observed in a genome (den *et al*., 2000). Various levels at which mutations could be detected are:

i. the DNA level
ii. the RNA level and
iii. the protein level.

Nonsense mutations at the protein levels could be as a consequence of the insertions and deletions (indels) originating from point mutations at the DNA level (Richard *et al.*, 1995).

1.2.4 Types of mutations

An organism's genome consists of several types of mutations. The mutation data collected from the Immunome KnowledgeBase consists of 12 types of mutations. These mutations are:

1) 3' UTR
2) 5' flanking
3) 5' UTR
4) Complex
5) Frame shift
6) Indels

7) Intron
8) Missense
9) Nonsense
10) Promoter
11) Silent mutations
12) Chromosomal mutations

Each of the above mutation types will be defined and explained explicitly below. The entire structure of a typical eukaryotic gene stemming from DNA through transcription, pre-messenger RNA, mRNA to protein product is illustrated in Figure 1.5. This figure aids in the understanding of some mutation types especially their location in human or other genomes.

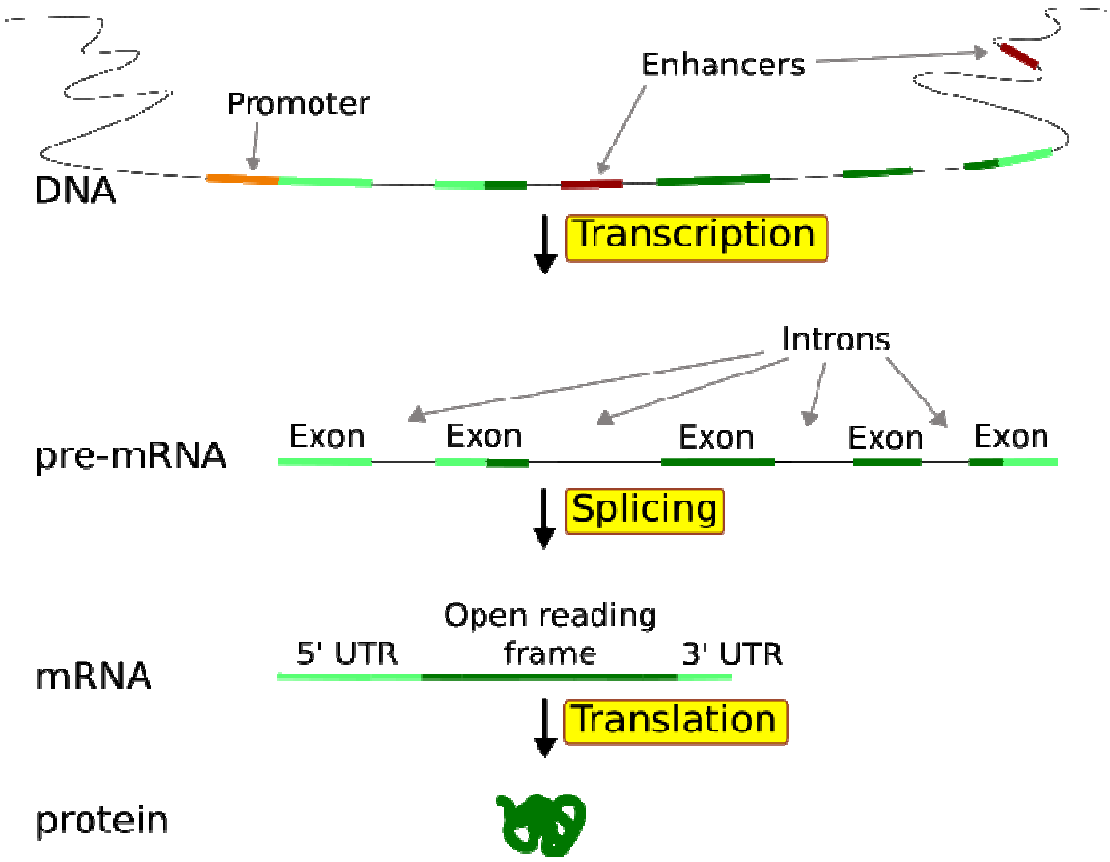

Figure 1.5 A typical structure of a typical eukaryotic protein-coding gene.
Source: Wikipedia, September 16th, 2009.

3' UTR mutations

These are mutations that appear at the downstream region of a protein coding gene close to its PolyA tail. These mutations appear outside the coding sequences (open reading frame) of the protein coding gene and thus, the protein coding sequences are not altered. Figure 1.5 clearly illustrates the location of such mutations on a typical eukaryotic gene.

5' UTR mutations

These are mutations that appear at the upstream region of a protein coding gene between the cap region and the start codon of a gene. These mutations occur outside the coding sequences (open reading frame) of a protein coding gene and thus, the coding sequences are not altered.

5' flanking mutations

These are mutations that occur just before the initiation codon at the upstream of a gene. This region is also called the promoter region as it contains important signals that initiate the transcription process.

Complex mutations

These types of mutations can arise as a combination of indels, changes within codons or frame shift. Mutations of these kinds are associated with changes in the Mitochondria (den *et al*., 2000).

Frame shift mutations

These are mutations caused by insertions/deletions of nucleotides. Due to the triplet nature of the genetic code, indels may result in a sequence of nucleotides that is not divisible by three, thus leading to an abnormal (too short or too long) protein.

Insertion and deletion mutations

These types of mutations arise as a result of either an insertion or a deletion of one or more nucleotides in a genome. Situations may also arise where a nucleotide is inserted or deleted more than once or a combination in a genome sequence.

Intron mutations

These are mutations that occur/appear in the intronic regions of a protein coding gene. Mutations occurring in these regions can affect the protein coding gene and consequently, the protein product. Such mutations can lead to eliminated splice sites resulting to transcribed introns that would lead to longer proteins and consequently, frame shifts if the length of the intron is not 3N (where N is the number of nucleotides) or a premature stop codon.

Missense mutations

They arise when a wrong amino acid is synthesized due to a substitution of a nucleotide in a codon. One or more substitutions can occur in a codon or different codons in a genome resulting in the non synthesis of a target protein. As such mutations occur in the coding regions of a protein coding gene, such synthesized proteins could lead to either a disease/defect (Choi *et al*., 2009; Byrne *et al*., 2009) or a new function in that organism.

Nonsense mutations

Such mutations arise when a substitution of a nucleotide in a codon leads to a premature stop codon within an open reading frame (Rowe *et al.,* 2009). Nonsense mutations account for about 20% of diseases associated with single basepair substitutions in the coding region (Mort *et al*., 2008). There are three stop codons occurring in the nuclear DNA namely; TAA, TGA, and TAG with a percentage frequency of approximately 21.1, 38.5 and 40.4 respectively (Mort *et al*., 2008). Such proteins are rather too short and are generally quite different from the target protein and thus are more likely to cause diseases in humans and other organisms. The earlier a stop codon, the more the protein is truncated and the more unlikely is the protein to function.

Silent mutations

These are mutations that still give rise to the synthesis of a target protein despite a substitution of a nucleotide in a codon (Britten, 1993). This is due to the degeneracy nature of the genetic code. Apart from a nucleotide substitution in a codon, there could be more than one nucleotide substitution in a codon or different codons of a gene. Silent mutations occur in the coding regions of a protein coding gene. Such mutations can affect methylation signals or alter the codon usage in the protein synthesis resulting in a slow or fast protein production (Chamary *et al*., 2009). In addition, silent mutations can also affect the secondary structure of a protein coding gene which can alter the speed and efficiency of the translation process.

Chromosomal rearrangements

Modifications of the number of chromosomes such as a change in the total number of chromosomes, rearrangement of chromosomes, the arrangements of genetic materials in chromosomes and deletions or duplications of genes or the chromosome segments are called chromosomal rearrangements. Such rearrangements contribute significantly to speciation (Raskina *et al*., 2008). Two main types of chromosomal mutations exist namely:

i) changes in the number of chromosomes and
ii) alterations in chromosome structure

Changes in the number of chromosomes

During meiosis, nature's intention is for a progeny to inherit 46 chromosomes, 23 each from both parents. This is not usually the case with some offsprings due to non-disjunction of chromosomes during sexual reproduction. For example, a non-disjunction occurring on chromosome 21 would result to an offspring with Down's syndrome, 47 chromosomes (Klug *et a*l., 2006). Changes in the number of chromosomes could further divided into two main sub categories

i) aneuploidy and
ii) euploidy

Aneuploidy is a gain or loss of one or more chromosomes but not a complete set. A gain of a single chromosome is called trisomy (2n+1) whereas a loss of a single chromosome is called monosomy (2n-1). On the other hand, euploidy is the presence of a complete set of chromosomes such as diploidy (2n) and polyploidy (triploidy (3n), tetraploidy (4n)).

Alterations in chromosome structure

This category is further divided into five sub categories namely:

i. Inversions
ii. Deletions
iii. Transpositions
iv. Duplications
v. Translocations

Inversions

Inversions arise as a result of an insertion of a chromosome fragment in a reverse manner. Figure 1.6 illustrates the mechanisms involved during gene (chromosomal segment) inversion.

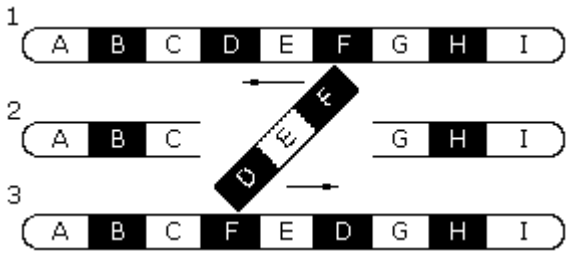

Figure 1.6 Inversion of a gene.
Source: biology-online (http://www.biology-online.org/2/8_mutations.htm).

Deletions

Deletions arise when a chromosome segment is lost. A chromosomal segment simply breaks off resulting to an entirely new chromosome segment. Figure 1.7 shows a loss of gene DEFG in gene1 to give rise to a completely truncated new gene 3.

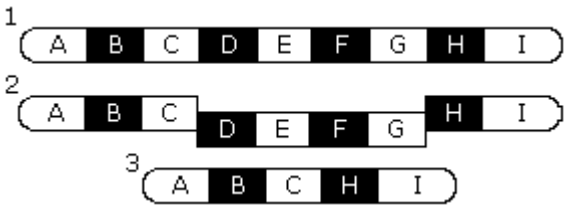

Figure 1.7 Deletion of a gene.
Source: biology-online (http://www.biology-online.org/2/7_mutations.htm)

Transpositions

Transpositions or transposons are DNA sequences that can move from one part of a genome to another. During this process, they can change the DNA content of a genome and thus result to mutations. Figure 1.8 shows the movement of gene JK from one genome to the other. Such movements may give rise to mutations in both genomes.

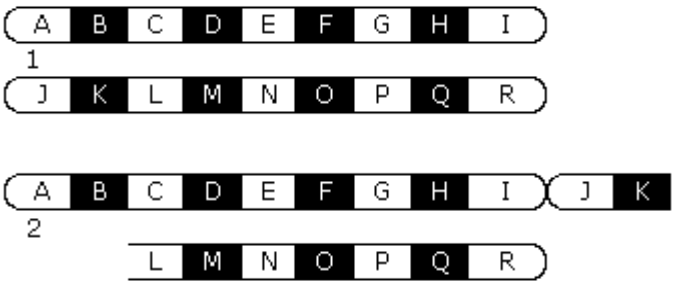

Figure 1.8 Transposition of a gene.
Source: biology-online (http://www.biology-online.org/2/8_mutations.htm).

Duplications

Duplications are repeats of a DNA or a protein sequence. They involve either a duplication of a certain DNA sequence or the whole chromosome. Most of the duplications that occur in a genome are tandem repeats with a definite pattern. Figure 1.9 illustrates the repetition of CD in gene 3. Repeats of such kinds may also lead to mutations.

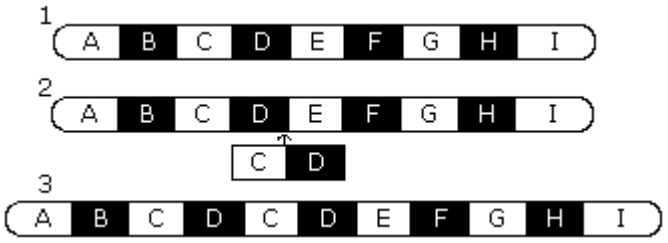

Figure 1.9 Duplication of a gene.
Source: biology-online (www.biology-online.org/2/7_**mutations**.htm*)*.

Translocations

Translocations are attachments of chromosome fragments to non homogenous chromosomes in a genome. They arise when a chromosome breaks at one point and then attaches itself to another chromosome.

Mutations could also be categorized based on their functional properties. Some of these categories are

i) loss of function mutations
ii) gain of function mutations
iii) lethal mutations

Loss of function mutations are mutations that arise when either the function of a gene product has reduced or it is lost completely. On the other hand, gain of function mutations arise when a gene product has gained a new function or an existing function is enhanced (Carla *et al*., 2000). Mutations that lead to the death of the affected cell in an organism are called lethal mutations.

1.3 Polymorphisms

Polymorphism could be defined as the co-existence of multiple variants in at least 1% of the same population. Organisms within this population are assumed to be mating randomly. Common examples of polymorphisms include the ABO blood groups in humans and major histocompatibility complex (MHC). Many genetic diseases arise as a result of polymorphisms at a single locus. When these multiple variants are rare in an interbreeding population (i.e. less than

1%), they may be referred to as mutations. Public discourse about genetics and hereditary indicates that mutations have negative connotations when compared to normal sequence variants or polymorphisms (Condit *et al.,* 2002). There are over 14 million polymorphisms spanning the whole human genome (Pang *et al*., 2009).

1.3.1 Types of polymorphisms

There are four basic types of polymorphisms namely:

i) protein polymorphisms
ii) restriction fragment length polymorphisms (RFLPs)
iii) copy number polymorphisms (CNPs)
iv) single nucleotide polymorphisms (SNPs)

1.3.1.1 Protein polymorphisms

Existence of multiple variants of a protein arising from amino acid polymorphisms, splicing variants or amino acid substitutions such as SNPs is referred to as protein polymorphism. Protein polymorphism is said to be associated with the development of respiratory diseases in neonates such as respiratory syncytial virus (RSV) bronchiolitis (Hallman *et al*., 2006).

1.3.1.2 Restriction fragment length polymorphisms

DNA sequences between various individuals have different digestion patterns by restriction enzymes. These variations (polymorphisms) in DNA sequence lengths due to restriction enzymes are called RFLPs. These fragments could be analyzed by gel electrophoresis for usage in genetic fingerprinting and markers to either identify culprits during a criminal investigation or particular groups of people at risk for a certain genetic disorder (Osakabe *et al*., 2008; Sertoz *et al*., 2008).

1.3.1.3 Copy number polymorphisms

These are polymorphisms that arise due to a variation in the number of copies of a sequence within the DNA molecule. CNPs are widely distributed in human and other genomes but are under estimated despite their great contribution to genetic diversity (Buckley *et al*., 2005).

1.3.1.4 Single nucleotide polymorphisms

Single nucleotide polymorphism (SNP) is a nucleotide point substitution in a genome**.** For instance, a SNP might change a DNA sequence A**A**GGTAATC to A**T**GGTAATC. Here, an A is substituted by a T. Two in every three nucleotide substitutions involve a change from a cytosine (C) to a thymine (T). Substitutions and deletions are random in nature. SNPs are the most abundant genetic variation in the human and other genomes. They account for more than 90% of all differences between individuals and they occur in every 100-300 bases along the 3-billion-bases in the human genome (Twyman *et al*., 2003). Genetic variations in the human and other genomes occur predominantly as single nucleotide polymorphisms (Twyman, 2004). SNPs are said to be the most dominating type of variations to have been explored in the human and other genomes due to their great contribution to genetic diversity (Buckley *et al.,* 2005; Marth *et al*., 1999). SNPs are said to also have a significant contribution to variations in drug response (Pang *et al*., 2009).

The human DNA consists of about 10 million SNPs of which three million or more are likely to differ between any two unrelated individuals (Twyman, 2004). Most human sequence variations are attributed to SNPs, while the rest are attributed to insertions and deletions of one or more nucleotide bases, repeat length polymorphisms and rearrangements. On average, SNPs occur on every 300 bases in a human genome (Sachidanandam *et al*., 2001). SNPs can occur in coding and non coding regions in a genome.

Of particular interest are those SNPs that appear within the protein coding genes. These SNPs are most likely to affect or alter the biological function of a protein and thus a start point of molecular evolution or disease. Our knowledge on SNPs has been on an increase in the recent past. This has led to an in-depth knowledge in molecular evolution/genetic diseases in humans and other organisms, paving the way to a wider study and understanding of the genes and proteins related to the human immune system. To this end, a reference set of 874 essential genes and proteins related to the human immune system was identified, annotated and characterized (Ortutay *et al.*, 2007).

1.3.1.4.1 Classification of SNPs

The human genome as well as other genomes contain various types of SNPs. To understand the role of SNPs in greater detail, it is crucial to classify them based on the region in which they appear in the human or other genomes. Below are types of SNPs and their locations in a genome.

i. cSNP, SNP that appears in the coding region of a genome
ii. iSNP, SNP that appears in the intronic region of a genome
iii. rSNP, SNP that appears in the regulatory region of a genome
iv. gSNP, SNP that appears in the gapped (intergenic) region of a genome
v. sSNP, SNP that appears in the silent region of a genome

1.3.1.4.2 Identification of SNPs

SNPs are identified principally by two methods namely:

i) the sequencing method and
ii) the databases method

While sequencing chips are used by the sequencing method for SNPs identification, a host of databases such as

a) dbSNP,
b) the SNP consortium (TSC)
c) human gene variation database (HGVbase)
d) environmental genome project (EGP)
e) Janpanese SNPS (JSNP)

can be used as well.

1.3.1.4.3 Applications of SNPs

SNPs are widely applied in biomedical research. Below are some of the areas in which their applications are essential.

i. They help in disease diagnosis
ii. They help pharmaceutical companies in drug development (Twyman *et al*., 2003).

iii. They serve as biological markers for pinpointing diseases. SNPs are the most abundant molecular genetic markers (Duran *et al*., 2009).

1.3.1.4.4 Types of SNPs

Six major types were identified and are categorized into two major groups.

i) SNPs that appear in the non coding regions of a genome and
ii) SNPs that appear in the coding regions of a genome

During the translation process, only coding sequences of the messenger RNA are translated into the target protein. Thus, cSNPs could lead to amino acid alterations. These alterations in amino acids could lead to truncated proteins which are usually non functional or disease causing in living organisms.

1.3.1.4.5 SNPs that appear in the non coding regions

Examples of SNPs that appear in the non coding region of a gene include locus-region, mrna-utr, splice-site, and introns. Such SNPs may occur in the messenger RNA and so, there may be alterations in the coding sequences (open reading frame). The non alteration of coding sequences in a genome would lead to the synthesis of a target protein. Figure 1.5 illustrates the structure of an mRNA from where the coding sequences are translated.

1.3.1.4.5.1 Locus-region

These are SNPs that appear in a gene region but not in the transcribed region. Such SNPs are difficult to be located with precision in a genome. They may be found in the regulatory region and they constitute about 2000 bases.

1.3.1.4.5.2 mrna-utr

These are SNPs that appear between the Cap and the Start codon at the upstream (5' UTR), and between the Stop codon and PolyA tail downstream (3' UTR) of a messenger RNA. These SNPs

appear within an exon but are never translated. The structure of an mRNA gene depicting the positions of 3' UTR and 5' UTR is illustrated in Figure 1.5.

1.3.1.4.5.3 Splice-site

These are SNPs that appear on either the first two or the last two bases of introns at splice sites. Since introns are located between exons, such SNPs are not translated but can contribute significantly to human genetic diseases (Hyo *et al.,* 2005).

1.3.1.4.5.4 Intron

These are SNPs that appear in intron region where the first and the last two bases remain unaltered. As introns are the non coding sequences, SNPs that appear in this region are never translated and thus have an effect on the target protein. The location of introns in a typical eukaryotic gene is illustrated in Figure 1.5.

1.3.1.4.6 SNPs that appear in the coding regions

Two major types of SNPs are located in the coding region of a gene. They are, synonymous and the nonsynonymous SNPs. These SNPs are located on the coding sequences of the matured messenger RNA and thus play a vital role in the evolution and biological function of a protein. SNPs that appear in these regions could lead to the synthesis of a target protein, a missense or a truncated protein depending on the resulting amino acid.

1.3.1.4.6.1 Synonymous SNPs in the coding regions

Despite a substitution of a nucleotide within a codon in a genome, the resulting amino acid is not altered and thus the target protein is still synthesized. The genetic code table indicates that all substitutions at the second nucleotide positions of codons result in amino acid replacement whereas a fraction of the nucleotide changes at the first and third positions are synonymous (Nei

et al., 2000). For example the codon CTT that codes for Leucine, a change at the third position by any of the remaining three bases (A, G and C) would still give rise to the same amino acid. This is due to the fact that the codon table is degenerate as more than one codon could code for the same amino acid.

1.3.1.4.6.2 Nonsynonymous SNPs in the coding regions

In nonsynonymous substitution, a substitution of a nucleotide base within a codon, results in a change in an amino acid sequence and hence a change in the target protein to be synthesized. The synthesis of a non target protein could result to a protein that would function quite differently or not at all. This type of SNP is further divided into two sub categories namely: missesnse and nonsense SNPs.

Missense SNPs

In this category, a substitution of a nucleotide base within a codon in a gene will lead to an alteration of an amino acid and thus the target protein. The genetic code table indicates that all substitutions at the second nucleotide positions of codons result in amino acid replacement whereas a fraction of the nucleotide changes at the first and third positions are synonymous (Nei *et al*., 2000). For example, a replacement of an A (adenine) by a T (thymine) at the second nucleotide position of the sixth codon position of hemoglobin chain (GAG to GTG) leads to the synthesis of a Valine instead of Glutamine. A person with such a variation is said to suffer from sickle-cell disease. Approximately 80% of missense SNPs in coding region are neutral (neither helpful nor harmful) while the rest are deleterious to protein function and hence disease causing (Wang *et al*., 2003).

Nonsense SNPs

In this category, a substitution of a nucleotide base within a codon in a gene leads to a premature stop codon which results to a truncated protein. The earlier a stop codon appears in a gene, the more truncated the protein becomes, and the more unlikely is the protein to function.

1.4 K_a and K_s substitutions rates

K_a is the rate of nonsynonymous nucleotide substitutions per nonsynonymous site while K_s is the rate of synonymous nucleotide substitutions per synonymous site (Jukes *et al*., 1999). The central dogma in molecular Biology states that DNA is transcribed to mRNA and the mRNA is in turn translated into a protein. The flow of this genetic material from DNA to protein through RNA is illustrated by Figure 1.10**.**

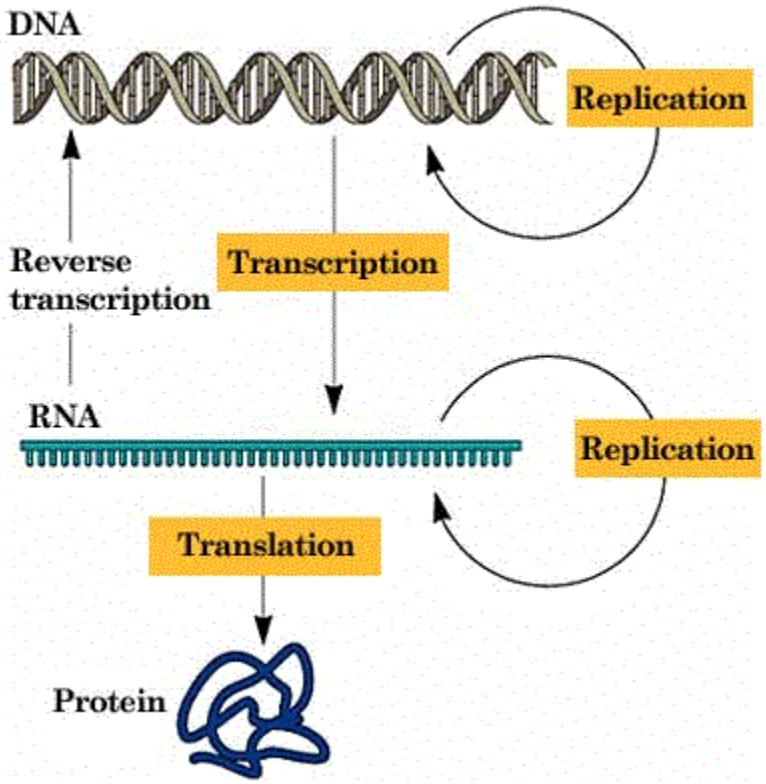

Figure 1.10 The flow of genetic material from DNA to protein through RNA.
Source: SCFBIO (www.scfbio-iitd.res.in/tutorial/orf.html)

There are twenty standard amino acids that make up a protein. DNA is made up of four nucleotides namely adenine (A), guanine (G), cytosine (C) and thymine (T). During DNA replication for example, errors may arise where a nucleotide or two are substituted within a codon. These substitutions occurring at the DNA level may or may not alter the final amino acid

and consequently, the target protein. Substitutions that will result to amino acid alterations are called nonsynonymous whereas those that will leave an amino acid unchanged due to degeneracy nature of the translation table are called synonymous. Most evolutionary models currently in use work on the assumption that synonymous substitution rates remain constant while nonsynonymous substitution rates vary over sites (Mayrose *et al.,* 2007).

1.5 Estimation of substitution rates

Several algorithms exist for calculating both K_a and K_s rates though they all have a similar approach originally proposed by Nei and Gojobori in 2000. Other authors such as Miyata and Yasunaga implemented a similar approach. The basic idea about this algorithm is to align two homologous sequences and compare the number of synonymous and nonsynonymous nucleotide differences codon by codon for a whole gene. In this project, I used human immunome sequences and their corresponding mouse orthologs. When there is just a single nucleotide difference between a codon pair, estimating the number of synonymous and nonsynonymous differences is easy. When two or more substitutions occur in a codon, computer or simulation methods are needed due to the increasing complexity of the algorithm (Nei *et al.,* 2000). In an event to study the evolution of a protein at the molecular level, substitution rates could as well be estimated by either codon or amino acid substitution models (Seo *et a*l., 2008).

1.6 Z-score

Z-score is the difference between K_a and K_s in each codon pair when two genes are aligned globally. **Z-score = K_a-K_s;** where K_a and K_s are as defined above. In a given gene, the number of synonymous substitutions is always greater than the number of nonsynonymous substitutions (LIopart *et al.,* 1999). A negative Z-score indicates that the number of synonymous substitutions per synonymous site is greater than the number of nonsynonymous substitutions per nonsynonymous site. A positive Z-score means the opposite. A gene with a positive Z-score is said to be undergoing positive selection and a gene with a negative Z-score is said to be conserved (Mayrose *et al*., 2007). The concept of Z-score is relatively new when compared to

that of the ratio of K_a to K_s for inferring whether a gene is conserved or is it undergoing positive selection.

1.6.1 Relevance of the Z-score

The relevance of K_a and K_s could be approached from two angles.

The Z-score approach

A negative Z-score means that the gene is conserved. This may be due to its vital role in a given organism. On the other hand, a positive Z-score means that the gene is evolving quickly and thus gaining new functions.

The ratio of K_a to K_s (K_a/K_s) approach

The ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site is a powerful indicator whether a protein is conserved or evolving/gaining new functions (Mayrose *et al*., 2007; Seo *et al*., 2008). When the ratio is approaching zero from the right, (or is between 0 and 0.05), the gene is conserved especially at the protein level (Ortutay *et al.* 2007). Genes that are conserved are said to play a vital role in living organisms. The slightest evolution of those genes may be harmful to such organisms. On the other hand, when the ratio K_a/K_s is far greater than 1, the gene is said to be evolving quickly. When a gene evolves in a manner that a new function or new functions are gained, we say the gene has undergone positive section. It is not obvious that when $K_a/K_s = 1$, then the gene is undergoing neutral section. In this case, further analysis such as multiple sequence alignment may be done to arrive at a valid conclusion.

1.7 Rationale for the study

Mutations in general and SNPs in particular are the root cause of genetic variation and thus, a base for evolution (Nachman *et al*., 2000). The preponderance of the available data for analyzing DNA sequence evolution is from the coding regions of a gene (Kreitman *et al.,* 1999). Mutation

is a major driving force for evolution (Nei *et al.,* 2000). The human immunome consists of genes and proteins related to the human immune system (Ortutay *et al*., 2006). These genes and proteins function primarily in preventing pathogens from invading the humans. The functions of these genes and proteins are conserved or modified depending on the role they play. Synonymous and nonsynonymous substitution rates within these genes and proteins are used to explain whether or not these genes and proteins are conserved or undergoing positive selection. These rates are used to assess the evolutionary speed of human immunome genes. The fact that our knowledge of mutations and SNPs is increasing day by day (den *et al.,* 2000; Ogino *et al*., 2007), and the crucial role played by the human immunome genes in preventing pathogens from invading humans, it is worth while estimating K_a and K_s rates to assess the evolutionary speed of the human immunome genes. To this effect, a set of 874 human immunome genes were analyzed. This set is available at http://bioinf.uta.fi/immunome (Ortutay *et al.,* 2006).

2. Objectives of the study

A comparison of the number of SNPs and mutations with synonymous (K_s) and nonsynonymous (K_a) substitution rates in human immunome.

Specific objectives:

1. To estimate the number of synonymous substitutions per synonymous sites (K_s) and the number of nonsynonymous substitutions per nonsynonymous sites (K_a).

To achieve this,

a) Mouse orthologs were collected for the human immunome genes

b) cDNA of both mouse and human orthologs were downloaded

c) Substitution rates were then calculated using bioperl modules.

2. To perform a comparative analysis of substitution and mutation rates.

To achieve this,

a) Mutation/SNP rates were calculated per gene per codon pair

b) Correlation analyses between the different rates were performed

c) Assessment of the relevance of these rates to evolution.

3. Materials and methods

3.1 Databases and datasets

In this project, four datasets and a file that consists of a reference set of 874 human immunome genes were used. The gene2refseq and HomoloGene datasets were downloaded from the EntrezGene database in NCBI while SNPs and mutation datasets were downloaded from the Immunome KnowledgeBase (IKB).

Gene2refseq

It is a large dataset that contains 13 variables. The first six variables of this dataset are tax_id, GeneID, status, RNA_nucleotide_accession.version, RNA_nucleotide_gi, and protein gi. This dataset contains the corresponding RefSeq accession numbers and GetInfo Identifier (GI) numbers for each gene pair in the dataset. Thus from this dataset, we can identify the protein and mRNA sequence entries from the RefSeq database which represents the genes in our analysis.

HomoloGene

This dataset contains 6 variables namely, HomoloGene group id (HID), Taxonomy ID, Gene ID, Gene symbol, protein gi, and protein accession respectively. This dataset contains information for orthologous genes (i.e. close descendants of the same gene but in different genomes having the same function). HomoloGene applies strict method of defining orthologs, therefore we can safely use this dataset to find the mouse orthologs for the human immune genes.

SNPs and mutation datasets

Mutation dataset contains the number of mutations occurring in each gene of the human immunome genes and the types of mutations whereas the SNPs dataset contains the number of SNPs occurring in each gene of the human immunome genes and the types of SNPs. These datasets were obtained from Immunome KnowledgeBase (IKB) which is available at (http://bioinf.uta.fi/IKB/).

Why mouse orthologs?

The mouse genome was chosen to study human orthologs due to the following reasons: The mouse genome is

i. a traditional animal model in immunology
ii. well annotated
iii. available and is used most often in other studies.

3.2 Computational environments used in the analysis

Perl

Perl is an acronym for Practical Extraction and Reporting language. It is a high level programming language that has gained grounds in bioinformatics due to its diversity and flexibility. It has an extension to the computing aspects in molecular biology called Bioperl. Bioperl has powerful methods embedded in its modules that are meant to solve specific tasks especially in molecular biology and bioinformatics. Though both Perl and Bioperl could be used in multiple platforms, the Linux platform is popular. Various modules available in Bioperl can be accessed from its website (http://www.bioperl.org/wiki/Main Page).

R

R is an open source implementation of the well-known S language. It is a programming or computational software tool which provides an environment in which one can perform statistical analysis. It is free software from Bristol University's website http://www.stats.bris.ac.uk/R/ . R is becoming more popular in statistical analyses because it is connected to the internet thus providing powerful help services. Most of its modules are written as packages which must be installed before being utilized. The software was first created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. It can be used on multiple platforms such as Windows and Linux. The software is maintained by R core team developers.

3.3 Algorithm of the substitution rate calculations

A careful interplay between the reference set (874 human immunome genes) and the four datasets (gene2refseq, HomoloGene, SNPs and mutations) made it possible to estimate the substitution rates with the aid of perl/bioperl modules. Below is the algorithm to estimate the K_a and K_s values. The algorithm was divided into two phases.

I. Derivation of human mouse orthologs
II. Calculation of the substitution rates (i.e. K_a and K_s)

Phase I: Derivation of human mouse orthologs

Phase I is subdivided into two arms namely: the derivation of human GenBank files and mouse GenBank files.

First arm: Derivation of human GenBank files.

- A combination of the reference set (874 human immunome genes) and gene2refseq dataset gave the corresponding accession numbers for the human immunome genes from where human GenBank files were downloaded using Bioperl modules/methods.

Second arm: Derivation of mouse GenBank files.

- A combination of the reference set (874 human immunome genes) and HomoloGene dataset enabled the acquisition of the corresponding mouse orthologs. A further combination of mouse ids and gen2refseq dataset gave mouse accession numbers, from where mouse GenBank files were downloaded as well using Bioperl modules/methods.

A concatenation of data files from both arms gave the human_mouse_orthlogs.csv file which consists of five columns namely: Homology group id, human gene id, human gene accession numbers, mouse gene id and mouse gene accession numbers.

Phase II: Calculation of the substitution rates

A step-by-step algorithm

i. The modules Bio::SeqIO, Bio::AlignIO, Bio::Align::Utilities, Bio::Align::DNAStatistics, and Bio::Factory::EMBOSS were employed in the calculation process.
ii. The human_mouse_orthologs.csv file is read in line by line into a Perl program.
iii. Human and mouse ids were captured in a hash and converted into GenBank files using the extension ".gbk"
iv. Checks are done to ensure that only one (coding sequence) CDS exists for both human and mouse GenBank files on their feature tables.
v. Coding sequences for both human and mouse genes were translated into their corresponding protein sequences and checked to ensure that the script's translation equals the standard GenBank's.
vi. Needleman-Wunsch algorithm which is embedded in the module Bio::Factory, was used to perform global alignment for each pair of the human and mouse protein sequence.
vii. A method, aa_to_dna_aln embedded in the module Bio::Factory::EMBOSS was used to generate cDNA using both human and mouse protein sequences as a guide.
viii. K_a_K_s pair statistics were calculated using Bio::Align::DNAStatistics module. This module generated an additional statistic including Z-score. The calculated K_a and K_s values were stored in the **human_kaks_statistics.csv** file that had four columns namely: human gene id, K_a value, K_s value and Z-score value respectively.

To achieve the second objective, more information (such as number of mutations, number of SNPs, number of deletion mutations, number of nonsense mutations, number of insertion mutations, number of intronic mutations, number of coding nonsynonymous SNPs, number of coding synonymous SNPs, number of intronic SNPs, number of locus regions SNPs, number of mrna-utr SNPs, number of missense mutations, and K_a/K_s) from mutation and SNP datasets were added to the human_kaks_statistics.csv file. This file was then read into R in Linux platform to perform correlation analyses. Figure 3.1 illustrates a diagrammatic view of the whole algorithm.

3.4 Flow chart.

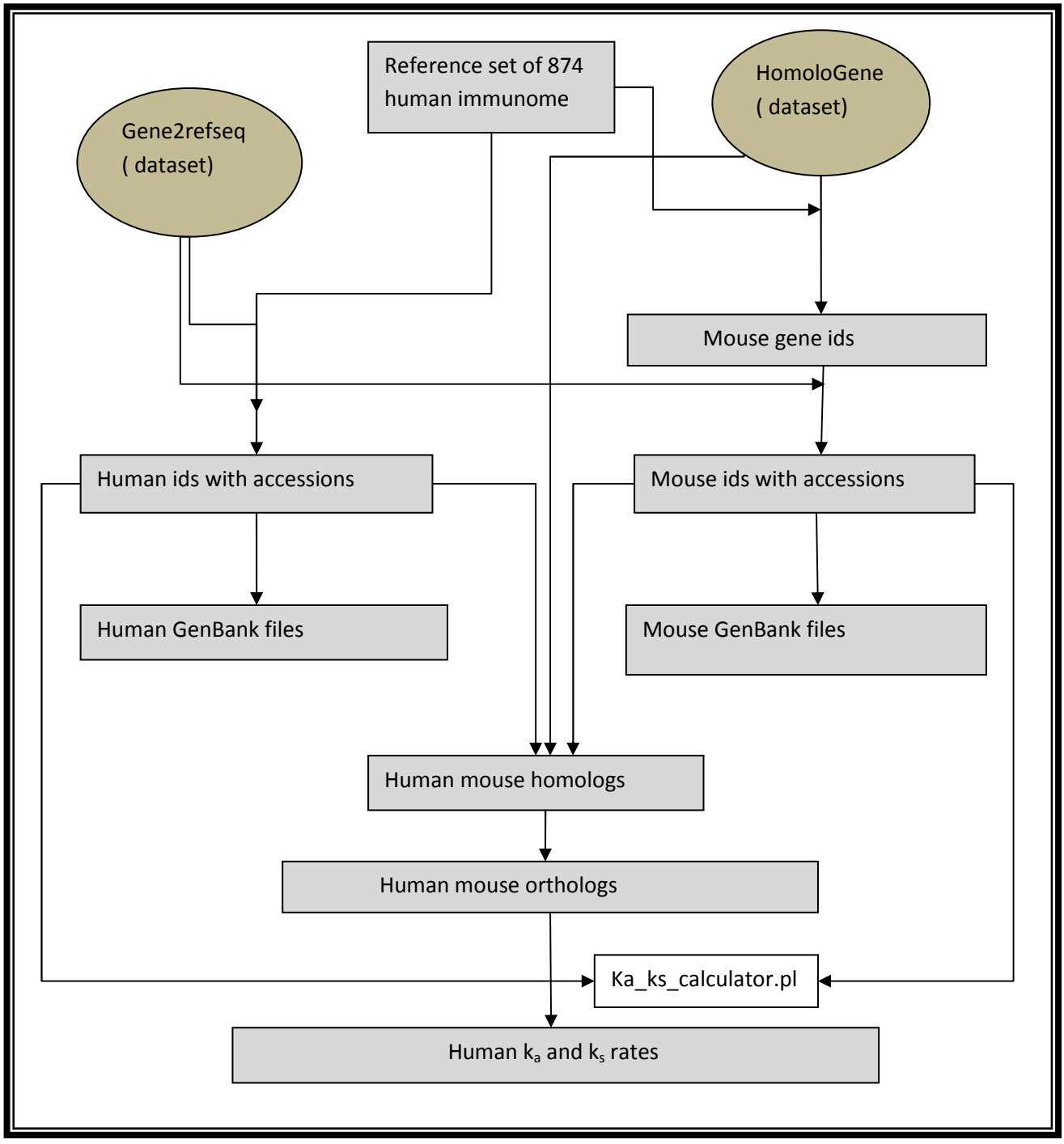

Figure 3.1 Flow chart showing the interplay between databases, datasets and a reference set that leads to the derivation of human_$k_a k_s$_statistics.csv that contains the calculated K_a and K_s values.

Brown (i.e. spheres): Datasets
Grey (i.e. rectangles): Files

3.5 Statistical analysis

Correlation

The analysis was done in R statistical environment using the Pearson Product Moment Correlation. Correlational techniques are used to study relationships between two or more variables. Pearson product moment correlation coefficient is a parametric approach whose assumptions rely on that the variables in question are random and distributed normally or approximately. Correlation coefficient is denoted by r, where r is a real number between -1 and 1 inclusive (i.e. r Є [-1, 1]).

Types of correlations

Four basic types of correlations exist, namely: Bivariate, multiple, partial and semi-partial.

Bivariate: Correlation between two independent variables.

Multiple: Correlation between three or more independent variables.

Partial: Correlation between two independent variables removing the effect of a third independent variable from both variables.

Semi-partial: Correlation between two independent variables removing the effect of a third variable from just one of the independent variables.

When the assumptions of independence and normality are not met, non parametric equivalence of bivariate correlation such as Kendall's Tau or Spearman rank correlation coefficient could be used to estimate r. They do not make use of the data values themselves rather, their ranks. Kendall's Tau and Spearman rank methods both make use of different computational formulae based on ranks but similar results/conclusions are reached.

Estimating the bivariate correlation coefficient (r)

Different formulae exist by which r can be estimated. They virtually arrive at the same value of r though computed slightly differently.

Given two independent variables X and Y, r can be computed as:

$$r = \frac{\sum XY - \frac{\langle \sum X \rangle \langle \sum Y \rangle}{n}}{\sqrt{\langle \sum X^2 - \frac{\langle \sum X \rangle^2}{n} \rangle \langle \sum Y^2 - \frac{\langle \sum Y \rangle^2}{n} \rangle}}$$

Where $\sum$ means summation sign and n, the total number of random variables.

The strength of a relationship

A value of r = +1 implies a perfect positive correlation between the two variables meaning, an increase or decrease in one variable results to an increase or decrease in the other respectively. A value of r = -1 implies a perfect negative correction meaning, an increase or decrease in one variable results to a decrease or increase in the other and an r = 0, implies the two variables are not related at all. The value of r indicates the strength of which two variables are related. The closer r is to either 1 or -1, the stronger the relationship between the two variables and the closer it is to naught, the weaker the relationship. An r = 1 or -1 does not mean causality between the two random variables. Though such an r is one the requirements for causality, more rigorous steps need be under taken to ascertain causality between the two random variables.

3.6 Comparing correlation coefficients when zeros are controlled and when they are not

Two correctional analyses were performed, one when the zeros were controlled in the estimation of r and the other, when they were not. The idea behind was that variables such as, number of deletion mutations, number of nonsense mutations, number of insertion mutations, number of intronic mutations, number of coding nonsynonymous SNPs, number of coding synonymous SNPs, number of intronic SNPs, number of locus region SNPs, number of mrnautr SNPs, and number of missense mutations have a high frequency of zeros which were thought to confound correlation coefficients and thus their associated p-values. These analyses are displayed in Tables 4.3 and 4.4.

4. Results

4.1 Effective sample size used for analysis

To compare the number of SNPs and mutations with synonymous (K_s) and nonsynonymous (K_a) substitution rates in human immunome, a reference set of 874 human immunome genes were collected. The reference set is available on http://bioinf.uta.fi/IKB/. A final sample size of 755 human immunome genes was obtained because some genes did not have accession numbers, some had a prefix different from the usual NM and some had more than one mouse gene id in the HomoloGene dataset. Thus, 755 human immunome GenBank files and their corresponding mouse ortholog GenBank files were downloaded and used in the analysis and calculations of K_a and K_s rates in human immunome.

4.2 Exploratory analysis

Applying Bio::Align::DNAStatistics module and Needleman-Wunsch global alignment on each of the 755 gene pairs (human and mouse protein sequences), and generating their corresponding cDNA sequences, estimates of the variables K_a, K_s, K_a/K_s, and Z-score were obtained. In addition to those variables, SNP and mutation data were described and analyzed comprehensively as well.

Nonsynonymous substitution rate (K_a)

The minimum and maximum K_a rates are 0.000 and 1.794, respectively. Figure 4.1 shows the distribution of K_a rates in human immunome genes. These values are clustered between 0 and 0.32. The mean K_a rate is 0.178 (0.158) and the median K_a rate is 0.153. The mean is over estimated here because K_a rates are positively skewed and thus, the median is a better statistic for the measure of central tendency. However, its square root transformation is approximately normal as shown in Figure 4.2.

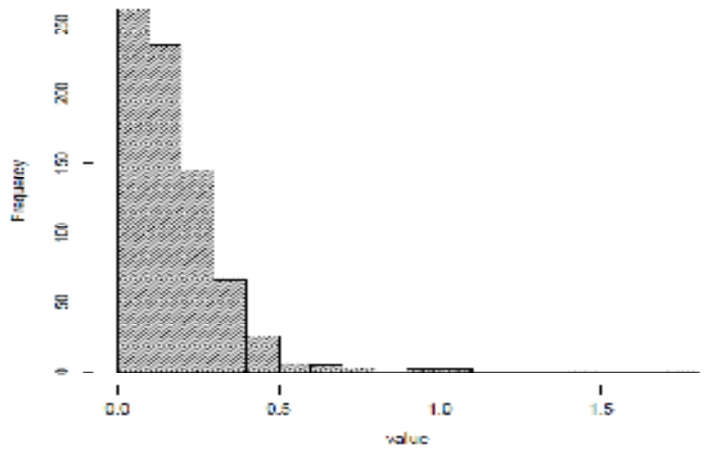

Figure 4.1 The distribution of nonsynonymous substitution rates for human immunome genes.

The distribution of the square root values of nonsynonymous substitution rates as shown in Figure 4.2 is approximately normal.

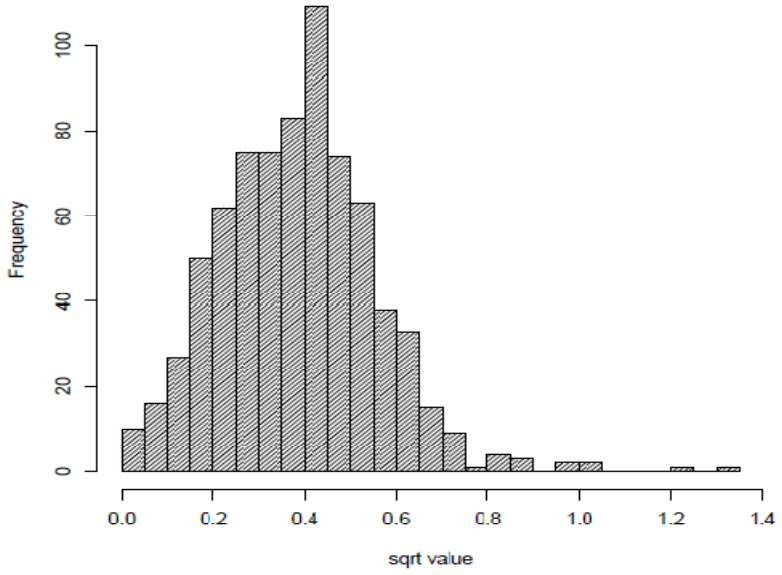

Figure 4.2 Square root distribution of nonsynonymous substitution rates for human immunome genes.

Synonymous substitution rate (K_s)

The minimum and the maximum K_s rates are 0.169 and 1.490, respectively. The mean K_s rate is 0.685 (0.169). Figure 4.3 shows the distribution of K_s rates for human immunome genes being approximately normal.

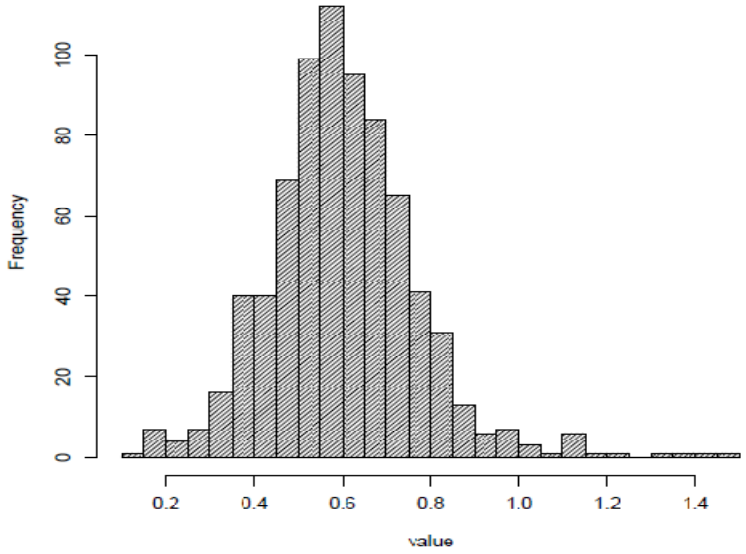

Figure 4.3 The distribution of synonymous substitution rates for human immunome genes.

The ratio of nonsynonymous to synonymous substitution rates (K_a/K_s)

The minimum and the maximum K_a/K_s rates are 0.000 and 7.852, respectively. The mean K_a/K_s rate is 0.394 (0.488) and a median K_a/K_s rate of 0.246. Figure 4.4 shows a positively skewed distribution of K_a/K_s rates for the human immunome genes. The mean is over estimated here as well because K_a/K_s rates are positively skewed and thus, the median is a better statistic for the measure of central tendency.

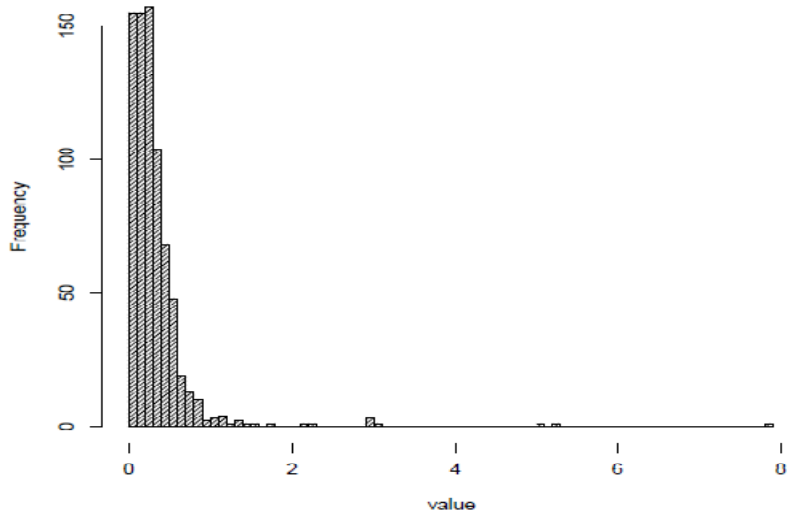

Figure 4.4 The distribution of K_a/K_s rates for human immunome genes.

Figure 4.5 shows the distribution of the log values of K_a/K_s rates for human immunome genes. This distribution is almost negatively skewed.

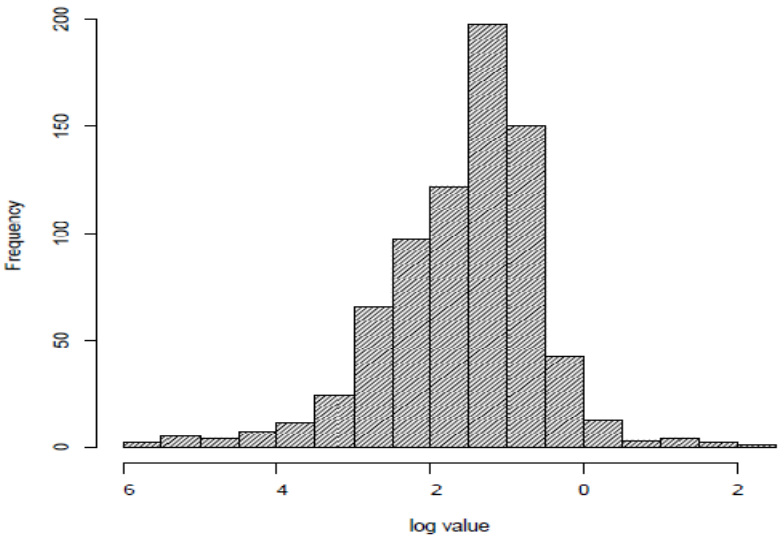

Figure 4.5 Log distribution of the ratio of nonsynonymous to synonymous substitution rates for human immunome genes.

Z-score

The minimum and the maximum Z-score values are -62.730 and 11.150, respectively. The median Z-score value is -11.980, while the mean Z-score is -13.150. Figure 4.6 shows the distribution of Z-score values for human immunome genes which is approximately normal.

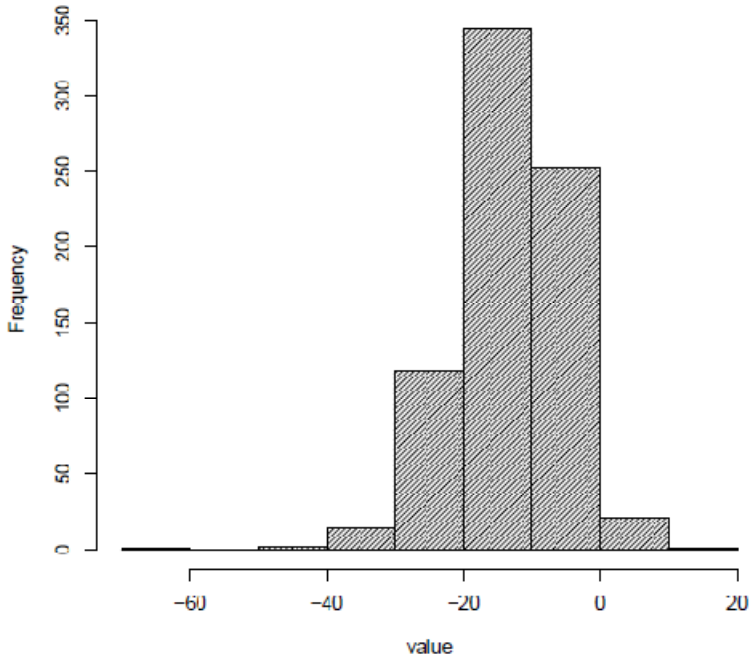

Figure 4.6 The distribution of Z-score values for human immunome genes

SNP data

Table 1 illustrates the frequencies of various types of SNPs. A total of one hundred and fifty four thousand and thirteen (154,013) SNPs were available in human immunome genes. The preponderance of SNPs were the introns 123,265 (80%) while splice-site accounted for the least frequency 22 (0.01%). Table 4.1 does not include non-applicable (NA) type of SNPs with a percentage occurrence of 4.17. This justifies the sum of the total percentage frequency being less than one hundred.

Table 4.1 Frequency distribution of various types of SNPs.

SNP type	Frequency	% Frequency
Coding non-synonymous	4,124	2.67
Coding synonymous	2,696	1.75
Intron	123,265	80.04
Locus- regions	11,472	7.44
Mrna-utr	6,005	3.89
Splice-site	22	0.01
Total	**154,013**	**100**

Figure 4.7 illustrates the distribution of the number of SNPs in human immunome genes. This distribution is positively skewed. The lowest number of SNPs is 0 while the highest number of SNPs is 3207. The median number of SNPs is 99.

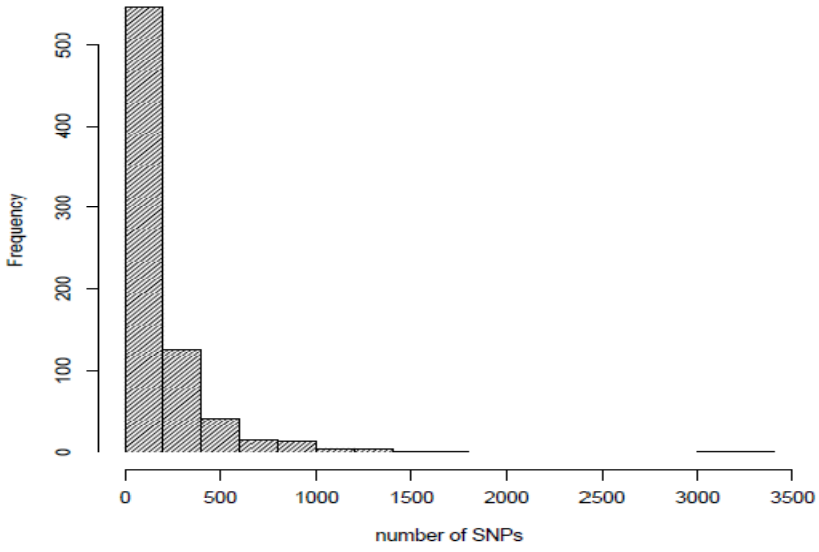

Figure 4.7 The distribution of the number of SNPs for human immunome genes.

Mutation data

Table 2 illustrates the frequencies of the various types of mutations in human immuome genes. A total of four thousand and seventy-six (4,076) mutations were available. Missense mutations accounted for the highest frequency 1,878 (46.07%). Mutation types such as 3' UTR, 5' flanking, 5'UTR, complex, frame shift, indel and promoter were discarded from Table 4.2 due to their low percentage frequency of less than 0.3. This omission accounts for a total percentage frequency of less than one hundred.

Table 4.2 Frequency distribution of the various types of mutations.

Mutation types	**Frequency**	**% Frequency**
Deletion	716	17.56
Frame shift	16	0.39
Intron	371	9.10
Missense	1,878	46.07
Nonsense	639	15.67
Sense	26	0.63
Uncertain	34	0.83
Total	**4,076**	**100**

4.3 Correlation analysis

Tables 4.3 and .44 illustrate the results of correlation analyses in two parts namely:

i) When the effect of the zeros is not controlled and

ii) When the effect of the zeros is controlled.

Table 4.3 Correlation analysis when the effect of zeros is not controlled

Red: Very strong evidence of association between the two random variables despite a moderate correlation.

Blue: Strong evidence of association between the two random variables despite a weak correlation.

Variables	**K_a**	**K_s**	**K_a/K_s**	**Z-score**
number_mutations	r = 0.02 p = 0.4888	r = -0.07 p = 0.0368	**r = 0.10** **p = 0.0042**	r = -0.05 p = 0.1020
number_deletion_ mutations	r = 0.0373852 p = 0.3056	r = -0.08 p = 0.0243	r = 0.12 p = 0.1269	r = -0.05 p = 0.1310
number_nonsense_ mutations	r = 0.01 p = 0.8408	r = -0.08 p = 0.0192	r = 0.08 p = 0.0151	r = -0.07 p = 0.0476
number_insertion_ mutations	r = 0.01 p = 0.7929	r = -0.07 p = 0.0523	r = 0.03 p = 0.2945	r = -0.07 p = 0.0321
number_intron_ mutations	**r = 0.13** **p = 0.0002**	r = -0.07 p= 0.0378	**r = 0.25** **p = 2.017e-12**	r = -0.00 p = 0.9874
number_missense- mutations	r = 0.01 p = 0.6457	r = -0.05 p = 0.1237	r = 0.07 p = 0.0285	r = -0.04 p = 0.1913
Snp	**r = -0.13** **p = 0.0002**	**r = -0.11** **p = 0.0014**	r =-0.02 p = 0.5127	**r = -0.37** **p < 2.2e-16**
number_coding_ nonsynonymous_snps	r = -0.00 p = 0.9972	r = 0.001 p= 0.8258	r = 0.06 p = 0.0591	**r = -0.39** **p < 2.2e-16**
number_coding synonymous_snps	r = -0.06 p = 0.0655	r = 0.01 p= 0.7994	r = 0.02 p = 0.5643	**r = -0.47** **p < 2.2e-16**
number_intronic_snps	**r = -0.14** **p = 6.031e-05**	r = 0.00 p= -0.1071	r = -0.03 P = 0.2982	**r = -0.38** **p < 2.2e-16**
number_locusregions_s nps	r = NA , std= 0 p = NA	r = NA , std=0 p = NA	r = NA, std = 0 p = NA	r = NA, std = 0 p-value=NA
number_mrnaute _snps	r = -0.01 p = 0.8204	r = -0.04 p = 0.2309	r = 0.048 p = 0.1869	r = -0.04 p-value= 0.2064

Table 4.4 Correlation analysis when the effect of zeros is controlled.

Red: Very strong evidence of association between the two random variables despite a moderate correlation.

Blue: Strong evidence of association between the two random variables despite a weak correlation.

Variables	**K_a**	**K_s**	**K_a/K_s**	**Z-score**
number_mutations	r = 0.05 p = 0.5431	r = -0.21 p = 0.0184	r = 0.13 p = 0.1426	r = -0.08 p = 0.3664
number_deletion_ mutations	r = 0.05 p = 0.6904	r = -0.28 p = 0.02901	r = 0.14 p = 0.2732	r = -0.12 p= 0.3532
number_nonsense_ mutations	r = 0.01 p = 0.9497	r = -0.29 p = 0.0069	r = 0.09 p = 0.4166	r = -0.1081125 p = 0.3336
number_insertion_ mutations	r = -0.06 p = 0.6392	r = -0.24 p = 0.0897	r = -0.02 p= 0.8903	r = -0.19 p = 0.1796
number_intron_ mutations	**r = 0.26** **p = 0.03901**	r = -0.27 p = 0.0288	**r = 0.33** **p = 0.0078**	r = 0.07 p= 0.5416
number_missense- mutations	r = 0.04 p = 0.6454	r = -0.21 p = 0.0420	r = 0.09 p = 0.3896	r = 0.00 p = 0.9837
Snp	**r = -0.13** **p = 0.0002**	**r = -0.12** **p = 0.0010**	r = -0.02 p = 0.5446	**r = -0.38** **p < 2.2e-16**
number_coding_ nonsynonymous_snps	r =-0.01 p = 0.8522	r = 0.00 p = 0.9658	r = 0.06 p = 0.0863	**r = -0.39** **p < 2.2e-16**
number_coding synonymous_snps	**r = -0.03** **p = 0.4547**	r = -0.00 p = 0.9294	r = 0.05 p = 0.2007	**r = - 0.45** **p < 2.2e-16**
number_intronic_snps	**r = -0.15** **p = 3.862e-05**	**r = -0.11** **p = 0.0032**	r = -0.04 p = 0.2868	**r = -0.39** **p < 2.2e-16**
number_locusregions_s nps	r = NA , std= 0 p = NA	r = NA , std=0 p = NA	r = NA, std = 0 p = NA	r = NA, std = 0 p = NA
number_mrmautr _snps	r = 0.02 p = 0.5457	r = -0.03 p = 0.3373	r = 0.07 p = 0.0718	r = -0.01 p = 0.744

5. Discussion

The study began with a reference set of 874 human immunome genes. Details about these genes were obtained from the IKB at http://bioinf.uta.fi/IKB/. Given a total of 874 human immunome genes, 755 of them were analyzed. The reason being, some human immunome genes did not have accession numbers, some had accession numbers but their prefixes were different from the usual NM, some had more than one mouse gene id in the HomoloGene dataset, and some were not refseq entries. Mouse orthologs were chosen for this study because the mouse genome is a traditional animal model in immunology, well annotated, available, and used most often in other studies.

The mean K_a and K_s for the human-mouse pair were 0.178 (0.158) and 0.685 (0.169) respectively, while the median K_a was 0.153. In the study conducted by Nei *et al.,* 2000, K_a and K_s rates were estimated at 0.056 and 0.354 respectively. These estimates are consistent with the fact that nonsynonymous substitution rates are always smaller than synonymous substitution rates in a given gene LIopart *et al*., 1999; Mayrose *et al*., 2007. The slight discrepancy between my results and that of Nei *et al*., 2000, could be due to the fact that Nei and Kumar used just a pair of human β globin and rabbit β globin ortholog genes whereas, I used a mean substitution rate obtained from 755 human-mouse pairs. Another reason could be that the lineage between the human and mouse is a bit different from the lineage between the human and rabbit.

In a total of 755 human-mouse pairs, 744 (98.5%) of the human immunome genes had higher K_s rates when compared to K_a rates. This finding is consistent with that of Nei *et al*., 1994; Nei *et al*, 2000. In this study, K_s rates being greater than K_a rates in the human immunome genes go to support the fact that there are more synonymous substitutions in human immunome genes than nonsynonymous substitutions. This statement is consistent with the findings of Kreitman *et al.,* 1999; LIopart *et al*., 1999; Mayrose *et al.,* 2007 where they said "K_s rates exceed K_a rates in a given protein except the protein is undergoing positive selection". The K_a and K_s rates depend on the human ortholog used in estimating them. Nei *et al*., 2000, illustrated that K_s rates between two human paralogs (human β and human α globins) are higher than K_s rates between orthologs

of different organisms (e.g. human and chicken genes; human and mouse genes). Conversely, K_a rates between human orthologs are smaller than K_a rates between human paralogs. This confirms the fact that evolution is very slow within the same organisms and conversely within different organisms dan *et al.,* 1999.

The distribution of K_a rates is positively skewed while its square root transformation is approximately normal. This, together with the fact that nonsynonymous substitution rates are random, justifies the application of a Pearson correlation in carrying out inferences. The distribution of K_a/K_s rates is positively skewed with a mean value of 0.394 and a median K_a/K_s value as 0.246. The highest frequency of K_a/K_s occurred in a range of about 0.04 to 0.16, highlighting the slow nature of evolution of the human immunome genes. This result is in line with the study carried out by Ortutay *et al*., 2007. The log distribution of K_a/K_s rates is approximately normal. These rates were a bit lower (0.02 to 0.13) in the study by Hurst, 2002 where he used mouse and rat orthologs. The reason could be that the evolutionary distance between the mouse and rat is smaller than between the human and mouse orthologs.

Z-score values are approximately normally distributed (Figure 4.6), and coupled with the fact that these values are random, Pearson correlation could be used for association search in order to carry out inference. The Z-score value for any ortholog gene pair is an important indicator on the evolutionary speed of that gene. It is assessed based on whether it is positive or negative. In a total of 755 human-mouse pairs analyzed, 744 (98.54%) of the human-mouse pairs had negative Z-score values. A negative Z-score value is interpreted as the rate of synonymous substitutions per synonymous sites exceeding the rate of nonsynonymous substitutions per nonsynonymous sites. Thus, in evolutionary perspective, one would say the speed of evolution is relatively slow in human immunome genes since the separation of the human and mouse common ancestor approximately 70 million years ago and so, evolution among these genes is just by chance or adaptive. A slow speed of evolution among the reference set of human immunome genes would be due to the fact that proteins are highly conserved. The highly conservative nature of proteins could be due to the fact that they adhere to their respective functions in humans and other organisms. A minimum Z-score value of -62.730 and a mean Z-score value of -13.150 go a long way to buttress the timid evolution in the reference set of human immunome genes.

Looking at SNP data in Table 4.1, the majority of the number of SNP types are the introns with a frequency of 123,265 (about 80%) that occurred in the human genome. This finding goes to support the fact that the largest part of the human genome is made up of non coding DNA which does not affect the evolution of the gene. This is in conformity with the fact that the greatest part of DNA sequence evolution is from the coding region of a gene Kreitman *et al.,* 1999. However, 4,124 (2.67%) of SNPs occurring in the human genome being coding nonsynonymous and 2,696 (1.75%) being coding synonymous, would mean that 10 (1.45%) of the human immunome genes underwent positive selection. This is because the number of SNPs occurring in the coding nonsynonymous region (2.67%) of the human immunome genes is far greater than the number of SNPs occurring in the coding synonymous region (1.75%). This signifies a positive Z-score value or a higher K_a rate when compared to the K_s rate.

In our mutation dataset (Table 4.2), missense mutations accounted for the highest frequency 1,878 (46.07%), nonsense mutations accounted for 639 (15.67%) and silent mutations accounted for 26 (0.63%). Silent mutations accounting for just 0.63% would mean that synonymous substitution rates are so low and thus the protein(s) in question is/are evolving and consequently undergoing positive selection. This result supports the fact that very few 10 (1.45%) human immunome genes maybe undergoing positive selection. In addition, the fact that the number of missense mutations (46.07%) is far bigger than the number of nonsense mutations (15.68%) elucidates the fact that most of the substitutions will result to amino acid altering rather than truncated proteins that may attract diseases. Thus the odds of a nucleotide substitution resulting to a missense mutation (amino acid altering for new protein function) is about three times higher than a nucleotide substitution resulting to a nonsense mutation.

Correlation analyses between various variables were performed under two main sections i) when the zeros were not controlled and ii) when the zeros were controlled in the correlation analysis. The rationale for this is that the occurrence of zeros was anticipated to bias or confound the real correlation coefficients among the variables concerned. The act of controlling zeros with the intention of having unbiased correlation coefficients was not worth the trouble as there were no significant differences in both the correlation coefficients and their associated p-values between

the variables. This means that statistical conclusions were the same in both Tables 4.3 and 4.4 despite slight alterations in their correlations coefficients.

In order to avoid misclassification bias, Table 5.1 (Munro, 2005) illustrates the standard measure to classify the strength of correlation between any two random variables

Table 5.1 Classification of Pearson correlation coefficient.

Correlation coefficient (|r|)	**Correlation**
0.00 – 0.25	Very low
0.26 – 0.49	Low
0.50 – 0.69	Moderate
0.70 – 0.89	High
0.90 – 1.00	Very high

Taking a look at Table 4.3, K_a, K_s and K_a/K_s show very low correlation coefficients between variable pairs (records/rows) with corresponding $p > 5\%$ except for the blue and red colored p-values that declare significance despite very low correlation coefficients. Despite a very low correlation coefficient between K_a/K_s and the number of intronic mutations, the association between the variables is highly significant ($r = 0.25$, $p = 2.017e\text{-}12$). The highest correlation is shown in column five between Z-score and the number of coding synonymous SNPs ($r = -0.47$, $p < 2.2e\text{-}16$). This signifies a strong biological linear relationship among those variables. Interestingly, the number of SNP is associated with K_a, K_s and Z-score with $r = -0.13$, $p = 0.0002$; $r = -0.11$, $p = 0.0014$; and $r = -0.37$, $p < 2.2e\text{-}16$, respectively, but failed to have a significant correlation with K_a/K_s (i.e. $r = -0.02$, $p = 0.5127$).

The entries in Table 4.4 are quite similar to those in Table 4.3. This justifies the fact that zeros have no significant effect on the statistical decision between the variables in question despite a slight perturbation of their correlation coefficients.

6. Conclusion

From the reference set of 874 human immunome genes, 755 (86.38%) were analyzed. In addition to this, four datasets, SNPs and mutation downloaded from IKB, gene2refseq and HomoloGene downloaded from EntrzGene were also used in the analyses. The mean K_s rate was estimated at 0.685, the median K_a rate was estimated at 0.153, and the median K_a/K_s value was estimated at 0.246. These values are consistent with the previous ones by Nei and Gojobori since a decade. The slight differences in the values above are dependent on the ortholog pairs employed in the calculations of both the K_a and K_s values. However, immunome genes are highly conserved since the separation of human and mouse lineages undergo positive selection with $K_a/K_s > 1$.

Despite a moderate correlation coefficient between the number of intron mutation and K_a/K_s ($r = 0.25$; $p = 2.017e-12$, SNPs and Z-score ($r = -0.37$; $p < 2.2e-16$), the number of coding nonsynonymous SNPs and Z-score ($r = -0.39$; $p < 2.2e-16$), the number of coding synonymous SNPs and Z-score ($r = -0.47$; $p < 2.2e-16$) and the number of intronic SNPs and Z-score ($r = -0.38$; $p < 2.2e-16$), their respective p-values are highly significant. These strong evidences suggest plausible biological relevance among these variables. Seeking, ascertaining and interpreting these relationships can provide more insights on the evolution of human immunome genes.

7. References

Ainsworth C. (2005): Nonsense mutations: running the red light. *Nature.* 438: 726-8.

Akashi H. (1994): Synonymous codon usage in *Drosophila malanogaster*. Natural selection and translational accuracy. *Genetics*. 136: 927-935.

Akashi H. (1999): Inferring the fitness effects of DNA mutations from polymorphisms and divergence data: statistical power to detect directional selection under stationarity and free recombination. *Genetics.* 151: 221-238.

Anagostopoulos T., P. M. Green, G. Rowley, C. M. Lewis and F. Giannelli. (1999): DNA variation in 5-Mb region of X chromosome and estimates of sex- specific type specific mutation rates. *Am. J. Hum Genetics*. 64: 508-517.

Antonarakis SE (1998): Recommendations for the nomenclature system in human gene mutations. Nomenclature Working Group *Hum. Mutat*. 11: 1-3.

Antonarakis SE., and McKusick VA. (1993): Discussion on mutation nomenclature. *Hum. Mutat*. 2: 248-8.

Beaudet AL. (1996): Update on nomenclature on human gene mutations. Ad Hoc committee on mutation nomenclature. *Hum. Mutat.* 8: 197-202.

Beaudet AL., and Tsui LC. (1994): A suggested nomenclature for designating mutations. *Hum. Mutat*. 4: 166.

Beutler E. (1993): The designation of mutations. *Am. J. Hum. Genet*. 53: 783-5.

Beutler E., McKusick VA., Motulsky AG., Scriver CR., and Hutchinson F. (1996): Mutation nomenclature: nicknames, systematic names and unique identifiers. *Hum. Mutat.* 8: 203-6.

Brian d foy, Pheoenix T., and Schwartz RL. (2005): Learning Perl 4th Edition. O'Reilly.

Britten RJ. (1993): Forbidden synonymous substitutions in coding regions. *Mol. Biol. Evol.* 10: 205-20.

Buckley PG., Mantripragada KK., Piotrowski A., Diaz de Ståhl T., and Dumanski JP. (2005): Copy-number polymorphism: mining the tip of an iceberg. *Trends Genet*. 21: 315-7.

Buetow KH., Edmonson MN., and Cassidy AB. (1999): Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet*. 21: 223-5.

Byrne JA., Strautnieks SS., Ihrke G., Pagani F., Knisely AS., Linton KJ., Mieli-Vergani G., and Thompson RJ. (2009): Missense mutations and single nucleotide polymorphisms in ABAB11 impair bile salt export pump processing and function or disrupt pre-messenger RNA splicing. *Hepatology*. 49: 553-67.

Cargill M., Altshuler D., Ireland J., Sklar P., Ardile K., Patil N., Lane CR., Lim EP., Kalayanraman N., Nemesh J., Ziaugra L., Friedland L., Rolfe A., Warrington J., Lipshutz R., Daley GO., and Lander ES. (1999): Characterization of single nucleotide polymorphisms in coding regions human genes. *Nat. Genet*. 22: 231-8.

Chamary JV., and Hurst LD. (2009): The price of silent mutations. *Sci. Am.* 300: 46-53.

Chamary JV., Parmley JL., and Hurst LD. (2006): Hearing silence: non neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* 7: 98-108.

Choi SY., Park HJ., Lee KY., Dinh EH., Chang Q., Ahmad S., Lee SH., Bok J., Lin X., and Kim UK. (2009): Different functional consequences of two missense mutations in GJB2 gene associated with non-syndromic hearing loss. *Hum. Mutat*. 30: E716-27.

Collins A., Lonjou C., and Morton NE. (1999): Genetic Epidemiology of single nucleotide polymorphisms. *Proc. Natl. Acad. Sci. U.S.A*. 96: 15173-7.

Comeron JM., and Aguade' M. (1996): Synonymous substitution at the *Xdh* gene of drosophila: heterogenous distribution along the coding region. *Genetics.* 144: 1056-62.

Comeron JM., Kreitman M., and Aguade' M. (1999): Natural selection on synonymous sites is correlated with gene length and recombination rate in Drosophila. *Genetics.* 151: 239-49.

Condit CM., Achter PJ., Lauer J., Sefcovic E. (2002): The changing meaning of "mutations:" A contextualized study of public discourse. *Hum. Mutat*. 19: 69-75.

Cotton RG. (2002): Communicating "mutation" modern meaning of connotations. *Hum. Mutat*. 19: 2-3.

Curtis JD. (2003): Perl programming for biologists. Wiley-Liss.

Daniel C, Salvekar A., and Schindler U. (2000): A gain of function in *STAT6*. *J. Biol. Chem.* 275: 14255-9.

den Dunnen JT., and Antonarakis SE. (2001): Nomenclature for the description of human sequence variations. *Hum. Genet*. 109: 121-4.

den Dunnen JT., and Antonarakis SE. (2003): Mutation nomenclature. *Curr. Protoc. Hum. Genet.* Chapter 7: unit 7.13.

den Dunnen JT., and Antonarakis SE. (2000): Mutation nomenclature extensions and suggestions to describe complex Mutations. *Hum Mutat.* 15: 7-12.

den Dunnen JT., and Paalman MH. (2003): Standardizing mutation nomenclature: why bother? *Hum. Mutat*. 22: 181-2.

Duran C., Appleby N., Vardy M., Imelfort M., Edwards D., and Batley J. (2009): Single nucleotide discovery in barley using autoSNP db. *Plant Biotechnol. J*. 7: 326-33.

Eanes WF., Kichner M., and Yoon J. (1993): Evidence for adaptive of the G6pd gene in the *Drosophila melanogaster* and *Drosophila simulans* lineages. *Proc. Natl. Acad. Sci. U.S.A*. 90: 7475-9.

Eyre-Walker A., and Bulmer M. (1993): Reduced substitution rate at the start of enterobacterial genes. *Nucleic Acids Res*. 21: 4599-603.

Glazko GV., and Nei M. (2003): Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol*. 20: 424-34.

Gojobori T. (1983): Codon substitution in evolution and "saturation" of synonymous changes. *Genetics*. 105: 10011-27.

Gojobori T., Li WH., and Graur D (1982): Pattern of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* 18: 360-9.

Golding GB., and Dean MA., (1999): The structural biases of molecular adaptation. *Mol. Biol. Evol.* 15: 355-69.

Gorlov IP, Gorlova OY., Franzier ML., and Amos CI. (2003): Missense mutations in hMLH1and hMSH2 are associated with exonic splicing enhancers. *Am. J. Hum. Genet*. 1157-61.

Graur D. (1984): Pattern of nucleotide substitution and the extent of purifying selection in retroviruses. *J. Mol. Evol*. 21: 221-31.

Graur D., and Wen-Hsiung L. (1999): Fundamentals of Molecular evolution. 2nd Edition.

Gu Z., Hillier L., and Kwok PY. (1998): Single nucleotide polymorphisms and hunting in cyber space. *Hum Mutat.* 12: 221-5.

Guda K., Moiova H., He J., Jamison O., Ravi L., Lutterbuagh J., Lawrence E., Lewis S., Willson JK., Lowe JB., Wiesner GL., Parmigiani G., Barnholtz-Slon J., Dawson DW., Veculescu VE., Kinzler KW., Papadopoulos N., Vogelstein B., Willis J., Gerken TA., and Markowtz D.(2007): Inactivating germ-line and somatic mutations in polypeptide N-accetylgalactosaminyltransferase 12 in human colon. *Proc. Natl. Acad. Sci. U.S.A.* 17.

Hallmam M,. and Haataja R. (2006): Surfactant protein polymorphisms and neonatal lung disease. *Semin. Perinatol*. 30: 350-61.

Hasson E., Wang IN., Zeng LW., Kreitman M., and Eanes WF. (1998): Nucleotide variation in the triosephosphate isomerase (Tpi) locus of *Drosophila melanogaster* and Drosophila. *Mol. Biol. Evol*. 15: 756-69.

Hodgkinson A, Ladoukakis E., and Eyre-Walker A. (2009): Cryptic variation in the human mutation rate. *PLoS Biol.* 7: e1000027.

Hughes AL., Friedman R., Rivailler P., and French JO. (2008): Synonymous and nonsynonymous polymorphisms versus divergence in bacterial genomes. *Mol. Biol. Evol*. 25: 2199-209.

Huppke P., Held M., Hanefeld F., Engel W., and Laccone F. (2002): Influence of mutation type and location of phenotype in 123 patients with Rett syndrome. *Neuropediatrics.* 33: 63-8.

Hurst LD. (2002): The K_a/K_s ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18: 486.

Hurst LD. (2009): Evolutionary genomics: A positive becomes a negative. *Nature*. 457: 543-4.

Hurst LD., Feil EJ., and Rocha EP. (2006): Protein evolution: Causes of trends in amino-acid gain and loss. *Nature.* 442: E11-2; discussion E12.

Irizarry K., Kustanovich V, Li C., Brown N., Nelson S., Wong W., and Lee CJ. (2000): Genome-wide analysis of single nucleotide polymorphisms in human expressed sequences. *Nat. Genet*. 26: 233-6.

Jarcho J. (2001): Restriction length fragment polymorphism analysis. *Curr. Protoc. Hum. Genet*. Chapter 2: Unit 2.7.

Jin L., and Nei M. (1990): Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.* 7: 82-102.

Kafatos F.C. Efstratiadis, B. G. Forget, and S.M. Weissman (1977): Molecular evolution of human and rabbit β-globin mRNAs. *Proc. Nat. Acad. Sci. U.S.A*. 74:5618-22.

Kang HJ., Choi KO., Kim BD., Kim S., and Kim YJ (2005): FESD: a fundamental element SNPs database in human. *Nucleic Acids Res*. 33 (Database issue): D518-22.

Kimura M. (1977): Preponderance of synonymous changes as evidence of the neutral theory of molecular evolution. *Nature:* 267: 275-6.

Kimura M. (1980): A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol*. 16: 111-20.

Kimura M. (1983): The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Klug WS., Cummings MR., and Spencer CA. (2006): Concepts of Genetics. 8th edition.

Krawczak M., Reiss J., and Cooper DN. (1992): The mutational spectrums of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Gene*t. 90: 41-54.

Kreitman M., and Akashi H. (1995): Molecular evidence of neutral selection. *Annu. Rev. Ecol. Syst*. 26: 403-22.

Kreitman M., and Cameron JM. (1999): Coding sequence evolution. *Curr. Opin. Genet. Dev*. 9:637-41.

Kumar S., Nei M., Dudley J., and Tamura K. (2008): MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief. Bioinformatics*. 9: 299-306.

Lawn RM., Efstratiadis A., O'Connell C., and Maniatis T. (1980): The nucleotide sequence of human β-globin gene. *Cell*: 21: 647-51.

LeBlanc DC. (2003): Statistics: Concepts and applications for science. Jones and Bartlett Pubishers.

Lercher MJ., and Hurst LD. (2003): Gene expression, gene clusters, and genomic regionality in rates of evolution. German conference on bioinformatics: 83-7.

Lewin B. (2004): Genes VIII. Pearson Education.

Li WH., Wu CI., and Luo CC. (1885): A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likihood of nucleotide and codon changes. *Mol. Biol. Evol*. 2: 150-74.

LIopart A., and Aguade M. (1999): Synonymous rates at the RpII215 gene of drosophila: variation among species and across the coding regions. *Genetics*. 152: 269-80.

Mao X., Young BD., and Lu YJ. (2007): Application of single nucleotide polymorphism microarrays in cancer research. *Curr. Genomics*. 8: 219-28.

Marth GT., KOrf I., Yandell MD., Yeh RT., Gu Z., Zakeri H., Stitzier NO., Hillier L., Kwok PY., and Gish WR. (1999): A general approach to single nucleotide polymorphism discovery. *Nat. Genet*. 23: 452-6.

Mayrose I., Doron. Faigenboim A., Bacharach E., and Pupko T. (2007): Towards realistic codon model: among site variability and dependency of synonymous and nonsynonymous rates. *Bioinformatics*. 23: 319-27.

Miyata T., and Yasunaga T. (1980): Molecular evolution of mRNA: a method of estimating evolutionary rates synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. *J. Mol. Evol*. 16: 23-6.

Miyata T., Hayashida H., Kuma K., Mitsuyasu K., and Yasunaga T. (1987): Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb. Symp. Quant. Biol* 52: 863-7.

Miyata T., Miyazawa S., and Yasunaga T. (1979): Two types of amino acid substitutions in protein evolution. *J. Mol .Evol*. 12: 219-36.

Miyata T., Nishida T., and T. Nishida (1980): Nucleotide sequence divergence and functional constraints in mRNA evolution. *Proc. Natl. Acad. Sci*. 77: 7328-32.

Mort M., Ivanov D., Cooper DN., and Chuzhanova NA (2008): A meta-analysis of nonsense mutations causing human genetic disease. *Hum. Mutat*. 29: 1037-47.

Mouchiroud D., Gautier C., and Bernaedi G. (1995): Frequencies of synonymous substitutions in mammals are gene specific and correlated with frequencies of nonsynonymous substitutions. *J. Mol. Evol.* 40:107-13.

Munro BH. (2005): Statistical methods for health care research. Lippincott Williams & Wilkins, 5th edition.

Nachman MW. (1998): Deleterious mutations in animal mitochondrial DNA. *Genetica.* 102-103: 61-9.

Nachman MW., and Crowell SL. (2000): Estimate of the mutation rates per nucleotides in human. *Genetics.* 156: 298-304.

Nei M. (2007): The new mutation theory of phenotypic evolution. *Proc. Natl. Acad. Sci. U.S.A.* 104: 12235-42.

Nei M., and Gojobori T. (1986): Simple methods for estimating numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3: 418-26.

Nei M., and Graur D. (1984): Extention of protein polymorphism and the neutral mutation theory. *Evol. Biol.* 17: 73-118.

Nei M., and Kumar S. (2000): Molecular evolution and phylogenetics. Oxford University Press.

Nei M., and Miller JC. (1990): A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. *Genetics*. 125: 873 79.

Nei M., and Takezaki N. (1994): Estimation of genetic distances and phylogenetic trees from DNA analysis. *5th world congress Genet. Appl. Lifestock production*. 21: 405-12.

Nelson DL, and Cox MM (2005): Lehninger, Principles of Biochemistry, 4th edition. W.H. Freeman & Co.

Ogino S., Gulley ML., den Dunnen JT., Wilson RB.; Association of molecular patholpogy training and education committee. (2007): Standard nomenclature in molecular diagnostics: practical and educational challenges. *J. Mol. Diagn.* 9: 1-6.

Ohta T. (1995): Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Mol Evol*. 40: 56-63.

Ortutay C., and Vihinen M. (2006): Immunome: a reference set of genes and proteins for systems biology of the human immune system. *Cell Immunol.* 244: 87-9.

Ortutay C., and Vihinen M. (2009): Immunome knowledge base (IKB): an integrated service for immunome research. *BMC Immunol*. 10:3.

Ortutay C., Siermala M., and Vihinen M. (2007): Molecular characterization of immune system: emergence of proteins, processes and domains. *Immunogenetics.* 59: 333-48.

Osaka is M., Kotsubo Y., Tajima R., and Hinomoto N. (2008): Restriction fragment length polymorphism catalog for molecular identification of Tapanese Tetranychus spider mites (Acari: Tetranychidea). *J. Econ. Entomol*. 101: 1167-75.

Pang GS., Wang J., Wang Z., and Lee CJ. (2009): Predicting potentially functional SNPs in drug response genes. *Pharmacogenomics.* 10: 639-53.

Peter Dalgaard (2003): Introductory statistics with R; Statistics and computing. Springer.

Piirilä H., Väliaho J., and Vihinen M. (2006): Immunodeficiency mutation databases (IDbases). *Hum. Mutat.* 27: 1200-8.

R Development Core Team (2007): R: A Language and Environment for Statistical computing R Foundation for Statistical Computing: Vienna, Austria: url = {http://www.R-project.org}

Raskina O., barber JC., Nevo E. Belyayev A. (2008): Repetitive DNA and chromosomal rearrangements: speciation-related events in plant genomes. *Cytogenet. Genome Res*. 120: 351-7.

Richard I., and Beckmann JS (1995): How neutral are synonymous codon mutations? *Nat. Genet.*10: 259.

Rowe SM., and Clancy JP. (2009): Pharmaceuticals targeting nonsense mutations in genetic diseases: progress in development. *BioDrugs*. 23: 165-74.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D; International SNP Map Working Group. (2001): A map of human genome sequence variation containing 1.42 million single nucleotide p polymorphisms. *Nature*. 409**:** 298-933.

Salcedo T., Geraldes A., and Nachman MW. (2007): Nucleotide variation in wild and inbred mice. *Genetics.* 177: 2277-91.

Seo TK., and Kishino H. (2008): Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. Syst. Biol. 57: 367-77.

Sertoz RY., Erensoy S., Pas S., Ozacar T., and Niesters H. (2008): Restriction fragment length polymorphism analysis and direct sequencing for determination of HBV genotype in a Turkish population. *New Microbiol*. 31: 189-94.

Shen LX., Basilion JP., and Stanton VP. Jr. (1999): Single nucleotide polymorphism can cause different structural folds of mRNA. *Proc. Natl. Acad. Sci. U.S.A.* 96: 7871-6.

Sherry ST., Ward MH., Kholodov M., Baker J., Phan L., Smigielski EM., and Sirotkin K. (2001): dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 29: 308-11.

Singh KK., Ayyasamy V., Owens KM., Koul MS., and Vujcic M. (2009): Mutations in mitochondrial DNA polymerase-gamma promote breast tumorigenesis. *J. Hum. Genet*. 24.

Stamatoyannopoulos JA., Adzhubei I., Thurman RE., Kryukov GV., Mirkin SM., and Sunyaev SR., (2009): Human mutation rate associated with replication timing. *Nat Genet*. 41: 393-5.

Storey M., and Jordan S. (2008): An overview of the immune system. *Nurs. Stand*. 23: 15-17.

Sun S., and Xu J. (2009): Chromosomal rearrangements between serotype A and D strains in Cryptococcus neoformans. *PLoS One.* 4: e5524.

Tabatha N., and Kimura M. (1981): A model of evolutionary base substitution and its applications with special reference of rapid change of pseudogenes. *Genetic*s. 98: 641-57.

Tajima F., and Nei M. (1984): Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* 1: 269-85.

Tang H. Wu L., (2006): A new method for estimating synonymous substitutions and its applications to detecting positive selection. *Mol. Biol. Evol*. 23: 372-9.

Tisdal J. (2001): Beginning Perl for Bioinformatics. An introduction to perl for Bioinformatics. O'Reilly.

Tsaur SC. and Wu CI. (1997): Positive selection and the molecular evolution of a gene of male evolution, Acp26Aa of Drosophila. *Mol. Biol. Evol*. 14: 544-9.

Tsaur SC., Ting CT, and WC. CI. (1998): Positive selection driving the evolution of a gene of male reproduction, Acp26Aa of Drosophila: II divergence vs. polymorphism. *Mol. Biol. Evol*. 15:1040-6.

Twyman RM. (2004): SNP discovering and typing technologies for pharmacogenomics. *Curr. Top. Med. Chem*. 4: 1423-31.

Twyman RM., and Primrose SB. (2003): Techniques patents for SNP genotyping. *Pharmacogenomics.* 4: 67-79.

Wang DG. (1999): Large scale identification, mapping, and genotyping of single nucleotide polymorphisms in the human genome. *Science.* 280: 1077-82.

Wang Z., and Moult J. (2003): Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain. *Proteins.* 53: 748-57.

Warnecke T., and Hurst LD. (2007): Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. *Mol. Biol. Evol.* 24: 2755-62.

Xue D., Yin J., Tan M., Yue J., Wang Y., and Liang L. (2008): Prediction of functional synonyonymous single nucleotide polymorphisms in human G-protein-coupled receptors. *J. Hum Genet.* 53: 379-89.

Yang Z. (1994a): Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39: 105: 11.

Yang Z., and Bielaswski JP. (2000): Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15: 496-503.

Yang Z., and Kumar S. (1996): Approximate methods for estimating the pattern of nucleotide substitution rates among sites. *Mol. Biol. Evol*. 13: 650-59.

Yang Z., Kumar S., and Nei M. (1995b): A new method of inference ancestral nucleotides and amino acid sequences. *Genetics*. 141: 1641-50.

Yang Z., Wong GK., Eberle MA., Kibukawa M., Passey DA., Hughes WR., Kruglyak L., and Yu J. (2000): Sampling SNPs. *Nat. Genet*. 26: 13-4.

Zeng LW., Cameron JM., Chen B., and Kreitman M. (1998): The molecular clock revisited: the rates of synonymous vs. replacement change in Drosophila. *Genetica*. 102-3 (1-6): 369-82.