The effect of pathogenic and neutral variants to the formation of nucleosome

Master's thesis
Zhang Qin
Institute of Biomedical Technology (IBT)
University of Tampere, Finland
August 2013

DEDICATION

I dedicate the Master's degree and work to my father **ZHANG Jinnan** and my mother **HE Huizhen**.

ACKNOWLEDGEMENT

The thesis work was carried out in the Bioinformatics research group, Institute of Biomedical Technology (IBT), University of Tampere, Finland.

I respect and owe my profound gratitude to Professor Mauno Vihinen, leader of Bioinformatics groups of University of Tampere and University of Lund (Sweden), for giving me the opportunity to work in his honored group and assigning me this interesting topic. Your enthusiasm, keen insight and wisdom in scientific research have inspired and motivated me to achieve success in the thesis work.

I would like to express my greatest gratitude to Adjunct Professor Csaba Peter Ortutay for his patient guidance and tremendous expertise throughout the entire project. Your dedication and generosity with your time in supervision was indispensable in completing the project.

I am grateful to Martti Tolvanen, a university lecturer and PhLic, for his guidance, motivation, inspiration and continuous support throughout the entire period of the Master of Science, Bioinformatics program. It would be difficult for me to complete this degree without your superior teaching techniques, suggestions and encouragement.

I am thankful to Professor Bairong Shen, Professor and Director of Center for System Biology, Soochow University, Suzhou, China for availing me the opportunity to pursue my Master's degree in Finland, a beautiful land with ingenuous people. Your academic rigor and diligence has been a source of inspiration and motivation for my work.

I am much indebted to Jani Kristian Haukka for his selfless assistance and arduous support so far in my life. His qualities of patience, honesty and innocence have irrefutably influenced my attitude towards life and enhanced spiritual growth. It is beyond reality that without his abled guidance that I could have conceived my master work away from home. I would like to express my special thanks to Shanmugapriya Kalaichelvan for her amicability and assistance with language and study.

Last but not least, I owe my deepest gratitude to my parents for their endless care and support in life since my birth not only financially but also spiritually. Their support, encouragement and understanding for my eager to pursue a Master's degree abroad are the source of my success and are tremendously appreciated.

Zhang Qin

April 10, 2013

Tampere, Finland

MASTER'S THESIS IN BIOINFORMATICS

Place	UNIVERSITY OF TAMPERE, FINLAND Institute of Biomedical Technology (IBT)
Author	Zhang Qin
Title	The effect of pathogenic and neutral variants to the formation of nucleosome
Pages	62 pages + Appendix 7 pages
Supervisors	Adjunct Professor Csaba Ortutay and Professor Mauno Vihinen
Reviewers	Adjunct Professor Csaba Ortutay and Professor Mauno Vihinen
Time	August 2013

Abstract

Background and Aims: Chromatin contains enormous genetic information, with further folding and compaction processes owing to the special structure of nucleosomes, which ensure the genetic continuity of organisms. The organization of chromatin structures like the nucleosome positioning pattern is found to have critical importance in contributing faithful gene regulations. Nucleosome shows special preference to specific DNA sequences, thus DNA sequence may participate in establishing their positioning patterns. To this concern, analysis of DNA sequence variations associated with nucleosome stability may provide important information underlying the associations. This study was aimed to find out whether pathogenic and neutral variations have different impact on the stability of nucleosome in terms of nucleosome binding affinities and occupancy levels. In addition to this regard, analysis of the variant variability and degree of pathogenicity and thereby possibly providing information for the prediction of pathogenicity of novel variants was one of the concerns in this research.

Methods: Two datasets including neutral and pathogenic variations were obtained from VariBench database. DNA sequences of specified length were downloaded with respective identifiers and were submitted to NuPoP to predict nucleosome positions and attributes. Variations were studied based on the location (within nucleosome core regions and at linker regions). Finally, statistical analysis was performed in R to examine the stability of nucleosome for neutral and pathogenic variations. In addition, variability and degree of pathogenicity of each nucleotide, substitution pattern and dinucleotide were calculated.

Results: There was no obvious difference in variant positions between neutral and pathogenic types along the DNA. Variations occurring inside nucleosomes displayed higher binding affinity and occupancy than variations at linker regions irrespective of the type of variations. Pathogenic variations showed higher nucleosome binding affinity and occupancy than neutral ones. However, significant occupancy difference between pathogenic and neutral variations were not observed for several substitution patterns, e.g. C→T, G→A and T→G. Transition variations showed higher frequency than transversion variations. Three out of four types of transition variations displayed higher probability to cause pathogenic variations; C→T and G→A were found to have highest substitution frequencies. CpG was observed with a high variation frequency and is more likely to become pathogenic type. TpG within nucleosome core regions and CpA at linker regions showed most and least pathogenicity, respectively. ApA and GpA might cause different variation types based on the location of variations.

Conclusion: Pathogenic and neutral variants distributed a similar positioning pattern along DNA. Variations occurring inside nucleosomes favor nucleosome stability more than variations at linker regions irrespective of the type of variations. Pathogenic variations are more likely to contribute to better nucleosome stability than neutral ones. Transition variations are more common. Dinucleotide CpGs are common variation sites and show high degree of pathogenicity. The location of variations has impact on nucleosome stability, dinucleotide variability and degree of pathogenicity.

CONTENTS

ABBREVIATIONS

AA	Amino Acid
bp	Base Pair
cDNA	complementary Deoxyribonucleic Acid
dATP	deoxyadenosine Triphosphate
dCTP	deoxycytidine Triphosphate
dGTP	deoxyguanosine Triphosphate
dTTP	deoxythymidine Triphosphate
DNA	Deoxyribonucleic Acid
FAD	Riboflavin-Adenosine Dinucleotide
Linker variation	Variation at linker regions
MW test	Mann-Whitney U test
nsSNV	non-synonymous Single Nucleotide Variation
Nucleosome variation	Variation within nucleosome core regions
NuPoP	Nucleosome Positioning Prediction engine
rs IDs	reference SNV identifiers
RNA	Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variation
SVM	Support Vector Machines

1. INTRODUCTION

In genetics, variation refers to the change of nucleotide sequence DNA and RNA of an organism caused by the unrepaired damage to its genome, such change can either be an insertion, deletion or inversion of a segment of DNA/RNA sequences on the chromosome. A single base variation is the replacement of a single nucleotide base by another base in DNA (A, C, G and T) and RNA (A, C, G and U) sequences. Variations most often arise during the DNA replication stage of meiosis (cell division process of gametes) and thereby can be inherited to offspring or by mutagens such as chemicals, radiations, etc. which are mostly of inheritance. However, not necessarily all variations are harmful. Alternatively, very few benign variations are beneficial like those variations that increase the fitness of organisms and thereby promoting traits that are desirable. For instance, a specific 32 bp deletion in human CCR5 confers HIV resistance to homozygotes and delays AIDS onset in heterozygotes (Sullivan *et al*., 2001).

Single nucleotide variation, abbreviated as SNV, involves the swapping of a nucleotide to another. SNV can either be located in protein-coding regions (coding SNV) or non-coding regions (non-coding SNV). The latter one may not have effects as it is not involved in the process of protein coding. The synonymous SNV is a coding SNV that do not alter the encoded amino acid due to degeneracy of the genetic code (Barreiro *et al*., 2008; Stenson *et al*., 2009; Varela *et al*., 2010). The non-synonymous SNV is the one causing the change of amino acid and thereby affecting protein functions, structures and may even cause structure instability, wrong folding and protein aggregation (Thusberg and Vihinen, 2009; Olatubosun *et al*., 2012).

Eukaryotic genomes are thought to encode an intrinsic nucleosome organization and this nucleosome positioning code may favor several specific chromosome functions such as transcription factor binding, transcription initiation and remodeling of the nucleosomes themselves (Segal *et al*., 2006). Nucleosome is formed by the two helical DNA strands of length 147 bp and these fragments wrap on histone proteins which are composed of two copies of each H2A, H2B, H3 and H4 (Luger *et al*., 1997). This special structure ensures the folding and compaction of chromatins and thereby allowing the carriage of massive genetic information in cells. Nucleosomes are found to be organized by multiple factors including processes of chromatin remodeling, competition with site-specific DNA-binding proteins and the DNA sequence preference of themselves. Segal *et al.* (2006) detected low nucleosome occupancy at functional binding sites and transcriptional start sites. They explain this phenomenon in a way that genomes use their intrinsic nucleosome organization either by encoding stable nucleosomes over non-functional sites to decrease their accessibility to functional sites or by encoding unstable nucleosomes over possible functional sites to increase accessibility of transcriptional binding factors to these sites. However, variations in genomic DNA may disrupt nucleosome-positioning signals encoded in DNA, hence altering the binding sites of transcription factors in the linker DNA and thereby leading to unfaithful gene regulations (Harbison *et al*., 2004; Segal *et al*., 2006; Tolstorukov *et al*., 2011).

DNA methylation, a process of adding a methyl group to cytosine or adenine DNA nucleotides, plays a major role in gene expression. CpG dinucleotides are common variation positions due to the methylation-induced deamination of 5-methyl cytosine and thereby causing the substitution of CpG to TpG/CpA. CpG islands, a higher concentration of CpG sites, are found in promoter regions of multiple genes from mammalian genomes (Saxonov *et al.*, 2006; Appanah *et al.*, 2007). Methylation of cytosines in CpG sites within gene promoter is associated with the cause of gene silencing and such feature is found in a variety of cancerous cells. In the contrary, the hypo-methylation of CpG sites is implicated in the over-expression of oncogenes within cancerous cells (Jones and Laird, 1999).

The ultimate goal of this thesis work was to determine whether pathogenic and neutral variations have different impact on the formation of nucleosome in terms of the nucleosome binding affinities and occupancy levels by comparing thousands of variations of both neutral and pathogenic types collected from VariBench database (Nair and Vihinen, 2013). Comparisons were carried out for variations within nucleosome core regions and at linker regions. The special structure of nucleosome has been thought to have relevance to variation frequencies and thereby affecting gene expression; hence nucleotide substitution rates and the dinucleotide compositions of both types of variations were also of great interest in figuring out the possibility of predicting pathogenicity of novel variants and thereby possibility providing some information for the development of pathogenicity predictor.

2. REVIEW OF LITERATURE

2.1 Nucleotides

Nucleotides are basic structural units of nucleic acids, Deoxyribonucleic acid (DNA) and Ribonucleic acid (RNA) which control the synthesis of proteins in cells. Each nucleotide is composed of a five-carbon sugar (ribose or deoxyribose), nucleobase (nitrogenous base) and at least one phosphate group. There are four types of nucleotides in DNA, abbreviated as dATP, dCTP, dGTP and dTTP while for RNA, they are ATP, CTP, GTP and UTP. Nucleotides play various roles in physiological activities. For instance, nucleotides act as carriers of chemical energy in cells (e.g. ATP, GTP). Nucleotides intermediate in cellular communication and signal transduction (e.g. cGMP and cAMP). Furthermore, nucleotides are integrated to cofactors in enzymatic reactions, e.g. coenzyme A, FAD, etc. (Alberts *et al*., 2002). The general structure of a nucleotide is presented in Figure 2.1.

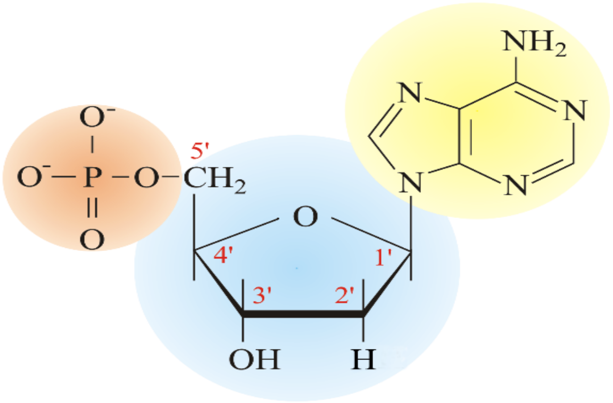

Figure 2.1 General structure of a nucleotide consisting of a phosphate group, a sugar (deoxyribose) and one nitrogenous base (adenine) (Source: Generalic, 2013).

2.1.1 Classification of nucleobases

Nucleobases are classified according to certain heterocyclic aromatic compounds called purines and pyrimidines. Adenine (A) and guanine (G) belong to the double-ringed class of molecules called purines while cytosine (C), thymine (T) and uracil (U) are all pyrimidines. Bases form pairs between the two helical strands of DNA: A pairs with T while C with G. The purine pyrimidine combination favors dimensional structure of DNA. Hydrogen bonding of nucleobases ensures the paring stability. There are two hydrogen bonds between A and T, while three hydrogen bonds between C and G, therefore DNA with high GC content is more stable than DNA with low GC content. These pairing rules are also known as Watson-Crick base pairing.

2.1.2 Transition and transversion variations

In molecular biology, there are two types of DNA substitution variations, transitions and transversions. Transition variations involve base changes of similar shape. Specifically, a transition is a single nucleotide variation which changes a purine (two rings) nucleotide to another purine (A ↔ G) or interchanges between one-ring pyrimidines (C ↔ T). In contrast, transversion variations refer to the substitution of a purine to a pyrimidine or vice versa. Figure 2.2 provides information on nucleotide substitutions.

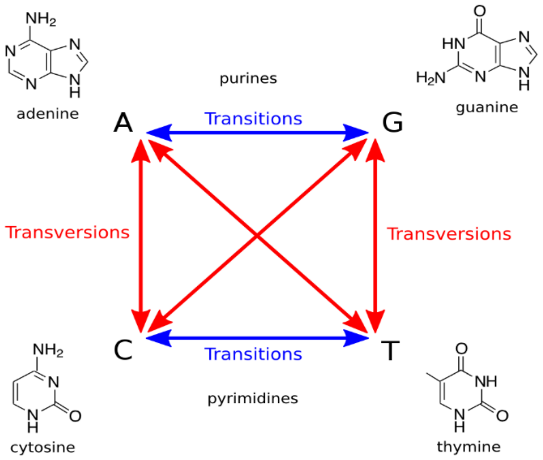

Figure 2.2 Transition *versus* Transversion variations (Source: Petulda, 2012).

A single nucleotide variation, abbreviated as SNV, is a single nucleotide substitution of one base to another in the same position of DNA sequence. There are two major types of SNVs, the non-coding SNV and the coding SNV. Coding SNV can be subdivided into two groups, the synonymous and non-synonymous SNVs (nsSNV) which are located in protein-coding regions of the DNA. A synonymous SNV does not change the encoded amino acid while an nsSNV alters the protein sequence.

Although there is more number of possibility for transversion variations as shown in Figure 2.2, a universal bias is in favor of transition variations over transversions due to the underlying chemistry of variations. Transition variations are less likely to result in amino acid substitutions due to "wobble", and therefore are more likely to persist as “silent substitutions” which are also known as synonymous SNVs (Collins and Jukes, 1994; Yang and Nielsen, 2000; Ebersberger *et al.*, 2002). However, by considering the fact of natural selection, Keller *et al.* (2007) proposed that the transition bias is not universal based on the study of variations that have accumulated in regions of the genome (grasshopper) which are free from selection. They found no evidence of a transition bias after the exclusion of variations associated with DNA methylation effect.

2.1.3 Nucleotide composition bias and substitution pattern and rate

DNA is composed of four kinds of nucleotides and they are abbreviated as dATP, dCTP, dGTP, and dTTP. These nucleotides are not distributed equally in genome giving a frequency of 25% for each. Rather, the nucleotide composition is biased. It was found that the overall nucleotide composition in human genome is 29.55% A, 20.44% C, 20.46% G and 29.54% T by estimating a total number of 2.86* 10^9 bases from genomic sequences downloaded from NCBI (Zhao and Boerwinkle, 2002). Moreover, substitution proportions for A↔G and C↔T were found up to 32.77% and 32.81% respectively whereas proportions were considerably lower for A↔T (7.46%) and C↔G (8.92%).

2.2 Evolution

Evolution is a process that results in the change of inherited characteristics of biological populations over successive generations. Evolutionary processes give rise to diversity at every level of biological organizations, ranging from species, individual organisms to molecules such as DNA and proteins (Hall and Hallgrímsson, 2008). An evolutionary process includes many general principles, such as the inherited variations, natural selection, the adaption to environment as well as the speciation due to the isolation of sub-populations and the adaption to diverse environment (Maynard-Smith and Szathmáry, 1997). It is crucial to note that the ontogeny of an individual is not considered as evolution because individual organisms do not evolve and that the changes in population must be passed on to the next generation. In addition, it is recognizable that "natural selection" is not synonymous with "evolution". Precisely, evolution can occur by processes other than natural selection such as genetic drift, a change in the frequency of a gene allele in a population (Masel, 2011). Natural selection can occur without any evolutionary change, as when natural selection maintains the status quo by eliminating deviants from the optimal phenotype (Futuyma, 2009). Evolution can be observed by detecting a change in gene frequency in a population.

2.2.1 Natural selection theory

Natural selection refers to a phenomenon of the survival of the fitness and elimination of the weak during the survival competition of biological organisms. Precisely, it is a process where organisms which get adapted to their environmental changes tend to survive and leave more offspring, hence eventually contributing to the appearance and elimination of certain genotypes in populations. This theory was originally proposed by Charles Darwin in 1859. There are four components in Darwin's natural selection theory:

1) Overproduction: most populations have more offspring each year than local resources can support, hence leading to a struggle for resources.
2) Struggle to survive: a result from the overproduction of organisms. Each species has to struggle for the survival, e.g. the competition of food, mate and habitat, etc.
3) Inherited variation: Some traits, consistently passed on from parent to offspring, are heritable. Organisms within populations exhibit individual variation in appearance and behavior. The accumulation of such variations through generations and generations cause more diversity between individuals.
4) Successful reproduction: Individuals possessing traits well suited for the struggle for local resources will contribute more offspring to the next generation.

According to Darwin's opinion, natural selection is resulted from the interaction between organisms and environment. From the evolutionary point of view, individuals with successful survival are not necessarily the fittest. Rather, only those individuals survived and consistently leave more offspring are considered as the fittest. Considering the fact that evolution alters the inherited characteristics at population level rather than at individual level, the modern evolutionary synthesis revised Darwin's opinion from the angle of Population Genetics and suggest that the genetic diversity existing in natural populations is a key factor in evolution and also it is a process of promoting the beneficial alleles among population (Darwin, 1872; Fisher, 1930; Mayr, 2002; Huxley, 2010).

2.2.1.1 Fitness

Natural selection is regarded as one of the most important milestones of modern biology. However, fitness is regarded as the central concept in natural selection. The definition for fitness from modern evolutionary theory was not determined by how long an organism survives, rather by how successful the organism is at reproducing. For instance, suppose an organism only lives half as long as others of the same species, but has twice more offspring surviving to adulthood, hence its genes will become more common in the adult population of the next generation and thus is considered as fitness. Precisely, fitness is the success of one's reproduction and averagely contributes to the accumulation of genotype or phenotype through generations (Darwin, 1872; Hartl, 1981; Maynard-Smith, 1989; Orr, 2009).

Obviously, natural selection is not equivalent to evolution rather; natural selection is one of several mechanisms contributing to the evolution of organisms, which further alters frequencies of genotypes of individuals in population due to their fitness. Figure 2.3 presents the natural selection process and the essence of fitness.

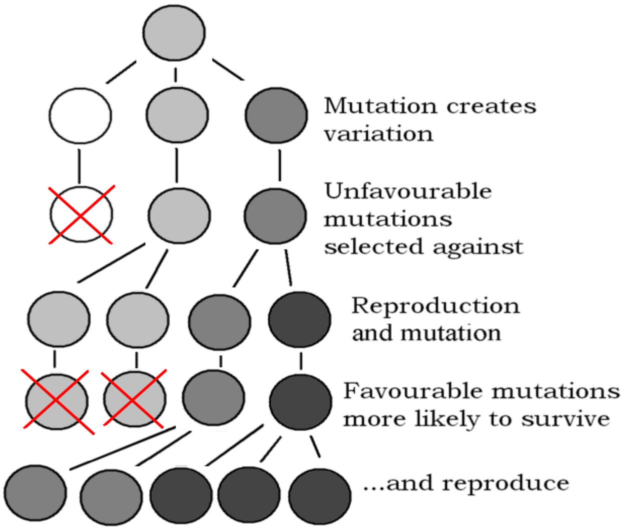
Mutation creates variation
Unfavourable mutations selected against
Reproduction and mutation
Favourable mutations more likely to survive
...and reproduce

Figure 2.3 A representation of natural selection process in which beneficial variations tend to survive while unfavorable ones are eliminated due to lower fitness to certain environment. (Source: Elembis, 2007)

2.2.2 Neutral theory of molecular evolution

The neutral theory of molecular evolution is the theory that at the molecular level evolutionary changes and polymorphisms are mainly due to mutations that are nearly enough neutral with respect to natural selection that their behavior and fate are mainly determined by mutation and genetic drift (Kimura, 1983). Genetic drift refers to the change in frequency of a genetic variant or allele in a population due to random sampling (Masel, 2011). The neutral theory of molecular evolution was mainly based on the substitution rate of nucleotides in nucleic acids and amino acids in proteins and the fact that the changes of nucleic acids and protein molecules caused by the substitution do not affect the function of biological macromolecules.

The main difference between this theory and the Darwin's evolution theory is that the evolution of organisms is mainly because of the random genetic drift of neutral mutations among populations, rather than selection. Precisely, the neutral theory of molecular evolution suggests most of mutations are neutral which means there are no advantages or disadvantages, hence natural selection and the survival of fitness do not apply to these neutral mutations. Nevertheless, the neutral theory is not antagonistic to Darwinian selection; rather, it produces another facet of the evolutionary process by emphasizing the much greater role of mutation pressure and random drift (Kimura, 1968; King and Jukes, 1969; Ohta, 1973; Kimura, 1983; Ohta, 1992; Ohta and Gillespie, 1996; Ohta, 2002; Nei, 2005; Hughes, 2007).

2.3 Nucleosome

Nucleosomes are the basic repeating units of eukaryotic genomic DNA and around 75-90% of genomic DNA is wrapped in nucleosomes, which enables the storage of massive genetic information in compact space. Each nucleosome contains a 147 bp stretch of DNA sequence and a histone protein octamer which contains two copies each of the core histones H2A, H2B, H3 and H4 (Luger *et al.*, 1997). The octamer is wrapped by the DNA sequence fragment and adjacent nucleosomes are linked by a stretch of free DNA called "linker DNA" which is normally of length of 10-80 bp varying from different species and tissues. A series of successively higher order structures are folded through nucleosomes and eventually form a chromosome (chromatin). Precisely, chromosomes are compacted in a way of forming higher order structures by connecting nucleosomes with linker regions of the DNA and linker histones such as H1 and its isoforms (e.g. H5) (Kornberg, 1974; Zhou *et al.*, 1998; Kornberg and Lorch, 1999).

The nucleosome structure which was obtained with ID "1AOI" from Protein Data Bank (PDB) was edited in Chimera 1.8 (Pettersen *et al.*, 2004) for a better view of the interactions between histones and nucleosomal DNA. The structure is presented in Figure 2.4.

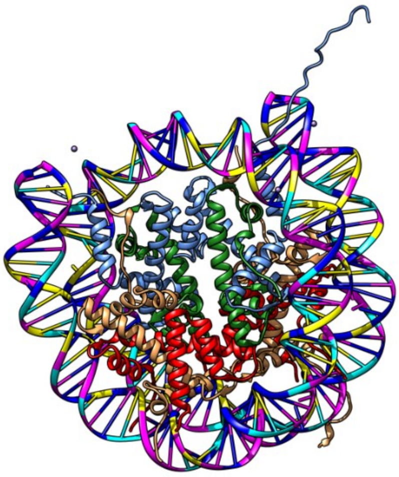

Figure 2.4 The structure of nucleosome. Nucleotide A, C, G and T are colored by blue, cyan, yellow and magenta, respectively. Histone proteins are highlighted as follows: H2A (brown), H2B (red), H3 (blue) and H4 (green).

The special structure of the nucleosome prevents the nucleosomal DNA from being accessed by various complexes which ensures the faithful gene regulation. Nucleosome organization is crucial for gene regulation. In living cells the nucleosome organization is determined by multiple factors, including the action of chromatin remodellers, competition with site-specific DNA-binding proteins as well as the DNA sequence preference of nucleosomes themselves (Satchwell *et al.*, 1986; Vignali *et al.*, 2000; Korber *et al*., 2004; Ioshikhes *et al.*, 2006; Segal *et al.*, 2006; Lee *et al.*, 2007; Yuan and Liu, 2008; Kaplan *et al*., 2009).

Chromatin remodeling allows the access of condensed DNA to regulatory transcription machinery proteins by dynamically modifying chromatin architecture and thereby controlling the gene expression. Chromatin remodeling is mainly carried out by two factors, one is the covalent modification of core histones of nucleosomes and the other is the nucleosome movement, ejection or restructuration by ATP-dependent chromatin remodeling complexes (Schulze and Wallrath, 2007; Whitehouse *et al.,* 2007; Teif and Rippe, 2009). The enzymatic modification of nucleosome histones, such as histone acetyltransferases, deacetylases and methyltransferases, affects the binding affinity between histones and DNA by loosening or tightening the condensed DNA on histones (Wang *et al.,* 2007). Due to the structure of nucleosomes (DNA wrapping on histones), there is competition between histone proteins and DNA-binding proteins for the DNA occupancy. Nucleosome shows higher affinity for some particular DNA sequences reflecting the sharp bending ability of DNA sequences as is required by the nucleosome structure (Segal *et al.*, 2006). As a consequence, there is difficulty to determine the relative importance of each of these mechanisms *in vivo* due to the combined actions of all influencing factors discussed above.

In order to determine the significant impact of DNA sequences on nucleosome positioning *in vivo*, Segal *et al.* built a nucleosome-DNA interaction model by using purified yeast nucleosome-combined sequences. Their results demonstrated that genomes encode an intrinsic nucleosome organization and this intrinsic organization can explain approximately 50% of the *in vivo* nucleosome positions. They proposed that this nucleosome positioning code may facilitate specific chromosome functions including transcription factor binding, transcription initiation and even remodeling of the nucleosomes themselves (Segal *et al.*, 2006). Precisely, nucleosomes facilitate their own remodeling by encoding intrinsically low nucleosome occupancy at sites destined for remodeling. Low nucleosome occupancies were found at functional binding sites. This is thought to be because genome use their own intrinsic nucleosome organization to encode stable nucleosomes over non-functional sites and thereby decreasing the accessibility of nucleosomes to transcription factors. As a consequence, the intrinsic nucleosome organization may contribute to the direction of transcription factors to their proper target sites while excluding them from irrelevant sites (e.g. sites occupied by nucleosomes). Analogously, nucleosomes are found to have low occupancies at transcription sites and this is thought to be because genome direct transcriptional machinery to functional sites by encoding nucleosomes with low occupancies, and thus enhancing their accessibility (Widom, 2001; Richmond and Davey, 2003; Sekinger *et al.*, 2005; Segal *et al.*, 2006; Kaplan *et al.*, 2009; Tolstorukov *et al.*, 2011).

Nucleosomes play a key role in gene regulation and the overwhelming majority of regulatory events occur at the transcription level. The genetic defects in transcription factors are regarded as reason of causing diseases because transcription factors control the expression of many genes, e.g. gene activation and gene silencing. In most cases, mutations in transcription factors lead to pleiotropic effects (Villard, 2004). Furthermore, variations or alterations in factors involved in nucleosome assembly have been connected to the cause of cancer and other human diseases (Groth *et al.*, 2007; Burgess and Zhang, 2013). Thus, the study of nucleosome positioning is of importance and might give insight to the diagnosis and treatment of related diseases. In this thesis work, nucleosome binding affinity and nucleosome occupancy were studied by comparing two large datasets containing both pathogenic and neutral variations. Further, the impact of both types of variations on the formation of nucleosome was investigated.

2.4 Nucleosome positioning prediction tools

There are various software available in predicting preferential nucleosome positions from DNA sequences. To select an appropriate tool for preforming this task was also one of the concerns. About the first successful nucleosome positioning prediction tool was developed by the Segal group. They built a probabilistic nucleosome-DNA interaction model by aligning nucleosome DNA sequences and their reverse complements about their centers. They associated a dinucleotide distribution with each position defined as ‘i’ which was estimated from the combined dinucleotide counts at three neighboring positions, such that the probability assigned by the model to a 147-bp sequence S is:

$$P(s) = P_1(S_1)\ {}_{i=2}^{147}\prod P_i\,(S_i \mid S_{i-1})$$

They made position weight matrices which characterize periodic patterns of specific dinucleotides and Boltzmann distribution to compute the probability of every configuration. In addition, they applied a dynamic programming method which efficiently computes the probability whether each base pair of S starts a nucleosome or is occupied by a nucleosome (Segal *et al.*, 2006; Field *et al.*, 2008; Kaplan *et al.*, 2008).

In addition to Segal’s model, there are several other outstanding nucleosome positioning prediction tools, e.g. the Mielle’s model, Peckham and Gupta’s Support Vector Machines (SVMs). Miele *et al.* (2008) had constructed a physical model of DNA bending around the histone octamer. This method calculates the free energy of a DNA fragment required to form the ideal curved structure without any training procedure (Anselmi *et al.*, 2002; Tolstorukov *et al.*, 2007). In addition, a model called Support Vector Machines (SVMs) was introduced to determine the nucleosomal and non-nucleosomal DNA, and this model is mainly based on the statistic oligomer frequency (Peckham *et al.*, 2007; Gupta *et al.*, 2008).

Tanaka and Nakai (2009) made an assessment over these three models by evaluating their prediction accuracy by using the genome-scale *in vivo* nucleosome maps in human, medaka fish, nematode, candida yeast and budding yeast. They came to a conclusion that Miele's model did not work well in all organisms from their evaluation test and regarded Gupta's SVM with the RBF kernel as the best predictor. However, due to the requirement of a variety of nucleosomal and non-nucleosomal DNA sequences for model training and the occurrence of the deterioration of prediction accuracy when training SVM with data from different organisms, Segal's method was recommended because of its stable performance (Segal *et al.*, 2006; Peckham *et al.*, 2007; Gupta *et al.*, 2008; Miele *et al.*, 2008; Tolstorukov *et al.*, 2008; Tanaka and Nakai, 2009).

In addition to these three tools mentioned above, there are other outstanding tools with accurate prediction under different conditions. Thus, to select an appropriate tool under a suitable condition is of great importance.

2.4.1 Nucleosome positioning prediction engine

Nucleosome positioning prediction engine, abbreviated as NuPoP (Wang *et al*., 2008; Xi *et al*., 2010), is an R package and is built upon a duration hidden Markov model for both Watson and Crick strands, in which the linker DNA length is explicitly modeled. Owing to the flexible and command-driven user interface and some features which suit the thesis task (e.g. R-based, output content, etc.), NuPoP was eventually selected among various outstanding nucleosome positioning prediction tools for performing the thesis task.

NuPoP has integrated two models, the nucleosome or linker DNA state model can be chosen as either a 4th order or 1st order Markov chain. Precisely, the 1st order Markov chain is meant for both nucleosome and linker DNA states while the 4th order (default) distinguishes nucleosome/linker in up to 5-mer usage and thus is slightly more effective in prediction, but runs slower. According to author's manual the time used by 4th order model is about 2.5 times of the 1st order model. Wang *et al.* modeled each chromosomal DNA sequence with a duration hidden Markov model of two oscillating states: nucleosome (N) and linker DNA (L). The nucleosome state has a fixed length of 147 bp ($\mathbf{e}$=e_1, ...,e_{147}) and the linker state has a variable length $F_L(k)$ (k = 1, ..., τ_L, L denotes for the maximum length they allow) with the assumption of a fixed state at each position and the starting, ending linker state of a complete chromatin sequence. G_L $(\mathbf{e}|k)$ denotes the homogeneous Markov chain model for the linker DNA. The probability for observing **e** as a linker DNA is given as follows:

$$P_L(\mathbf{e}) = G_L(\mathbf{e}|k)F_L \text{ (k)}$$

Additionally, Wang *et al.* defined the nucleosome occupancy at a specific position i and denoted o_i as the posterior probability that z_i = 1, i.e.,

$$o_i := P(z_i = 1|x)$$

The group also defined the histone binding affinity score at position i as the log likelihood ratio for the region x_{i-73}, ... x_i, ..., x_{i+73} to be a nucleosome vs. a linker, i.e.,

$$a_i := \log[\frac{P_N\, x_{i-73}, .. x_i, ... x_{i+73}}{G_L\, x_{i-73}, .. x_i, ... x_{i+73}{}^{|147}}]$$

The optimal path z can be found by the standard Viterbi algorithm and the nucleosome occupancy score can be estimated by using forward and backward algorithms with models P_N, G_L and F_L.

Three built-in functions including predNuPoP, readNuPoP and plotNuPoP are provided for nucleosome positioning prediction, prediction results read-in and prediction results visualization respectively. NuPoP takes a file of DNA sequence of any length in FASTA format as input. However, due to boundary effects, it was recommended to add at least 5000 bp of flanking sequence around the sequence of interest for the prediction accuracy. NuPoP outputs the Viterbi prediction of optimal nucleosome position map and a file in plain text format which includes five variables:

- Position: position in the input DNA sequence
- P-start: probability that the current position is the start of a nucleosome
- Occup: nucleosome occupancy score (from backward and forward algorithms)
- N/L: nucleosome (1) or linker (0) for each position based on Viterbi prediction
- Affinity: nucleosome binding affinity score

A typical Viterbi prediction of optimal nucleosome position map generated in the course of performing the thesis task is presented in Figure 2.5.

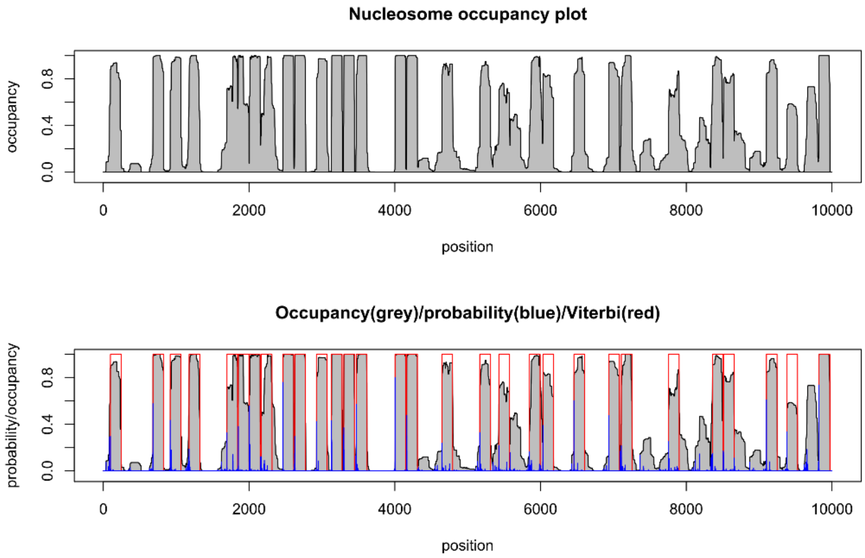

Figure 2.5 Typical Viterbi prediction of optical nucleosome position map generated by NuPoP. Nucleosome occupancy is marked as grey bars. Blue lines indicate the probability of a specific position being the start site of a nucleosome. Red boxes outline the Viterbi optimal prediction for nucleosomes.

In addition to the R package, NuPoP has two other formats including a web server prediction engine and a stand-alone FORTRAN program. The R package version (2.0.0) has been selected in this thesis study for convenience (Wang *et al.,* 2008; Xi *et al.,* 2010).

2.5 Statistical aspects

2.5.1 Non-parametric statistics

The Non-parametric statistics is a statistical method wherein the premise of the normality of the data is exempted. Precisely, non-parametric statistics neither relies on a predefined distribution of the data nor assumes the fix of model structure. Rather, nonparametric statistics is based on ranking or order of sorts (Gibbons and Chakraborti, 2003; Wasserman, 2006; Corder and Foreman, 2009; Hettmansperger and McKean, 2010; Bagdonavicius *et al*., 2011). As the demand for parameters are relived, nonparametric statistics have gained appreciation due to their ease of use.

2.5.2 The Mann-Whitney U test

The Mann-Whitney U test (MW test), also known as two sample Wilcoxon's test, is one of the most powerful nonparametric tests for comparing differences between two populations.

The null hypothesis for the Mann-Whitney U test is usually assumed as identical distribution functions between two populations against the alternative hypothesis that the two distribution functions differ only with respect to location. The MW test, a common nonparametric alternative for two sample t-test, does not require the assumption of the normality of sample distributions. Rather, it is based on the calculation of sum of ranks. Specifically, all the observations are arranged into a single ranked series and are ranked from lowest to highest, tied rank values were included where appropriate. These rankings are then resorted into two separate samples and sums of ranks T1 and T2 are calculated. The MW test has two approaches in evaluating the comparison depending on sample size (Mann and Whitney, 1947; Fay and Proschan, 2010).

1. For moderate size samples ($8 < \max(n_1, n_2) < 20$), the calculation is provided as follows:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1$$

$$U' = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_2$$

Where $n_1$and n_2 are the sample sizes for sample 1 and 2, T_1 and T_2 are sums of the ranks for sample 1 and 2, respectively. Value of U' is compared to the critical value and smaller U' value results in rejection of null hypothesis.

2. For larger samples ($\max(n_1, n_2) > 20$), the MW test use z values for testing.

$$z = \frac{U - u_U}{\sigma_U} \sim N(0,1)$$

Where $u_U = \frac{n_1 n_2}{2}$ and $u_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$.

Comparison between obtained z value and the critical z value yields either acceptance or rejection of the null hypothesis.

The probability value (abbreviated as p-value) is obtained in MW test performed in R and the decision whether to accept or reject the null hypothesis is based on the p-value and the significance level. For instance, a p-value less than 0.05 (5% significance level) means the rejection of the null hypothesis. In many applications, the MW test is used in place of the two sample t-test when the normality assumption is questionable.

This thesis work has dealt with two large datasets containing pathogenic and neutral variations. The MW test was applied to check whether there are differences between pathogenic and neutral SNVs in terms of nucleosome binding affinity and nucleosome occupancy levels.

2.6 CpG dinucleotide

A dinucleotide is a single piece of DNA or RNA that is of two nucleotides long. Alternatively, it is a single molecule composed of two linked nucleotides. For instance, a thymidine dinucleotide contains two thymidine nucleotides attaching by a phosphate bridge. In particular, the 5'-phosphate of one thymidine bonds to the 3'-hydroxyl group of the other thymidine, similar to the bonding seen in complete DNA and RNA molecules. Dinucleotide is often abbreviated as NpN, where "N" is a nucleotide of either A, C, G or T and "p" indicates the phosphate bridge. For instance, thymidine dinucleotide would be abbreviated as TpT. This abbreviation is crucial because it distinguishes dinucleotides from base pairs in double-stranded DNA. For instance, CG is an interacting pair of bases on opposite strands while CpG is a dinucleotide within one strand. However, in addition to the NpN kind dinucleotides, it is worth mentioning that there is another group of dinucleotides which are essential for energy transfer. It binds phosphate-to-phosphate and thereby creating a 5'-5' diphosphate bridge. The most common examples belonging to this type are NAD+ (niacin-adenosine dinucleotide) and FAD **(**riboflavin-adenosine dinucleotide**)** which are involved in metabolically-crucial redox reactions.

DNA methylation is a process involving the addition of a methyl group to cytosine or adenine nucleotides. DNA methylation plays a central role in gene expression, e.g. X chromosome inactivation (Yen *et al.*, 1984), genetic imprinting (Ferguson-Smith *et al.*, 1993), gene-expression regulation (Jones and Takai, 2002; Suzuki *et al.*, 2007; Weber *et al.*, 2007) and the defense mechanisms against parasitic DNA and transposons (Wilson and Murray, 1991; Barlow, 1993). More and more researches have shown that DNA methylation may cause genomic instability (Chen *et al.*, 1998) and is implicated in pathological processes such as cancer (Laird *et al.*, 1996) which is closely related to histone modification and RNA-associated silencing.

DNA methylation in mammals is carried out by three methyltransferases (DnmtT1, Dnmt3A and Dnmt3B) (Chen and Riggs, 2005) which target the cytosine in CpG dinucleotide. CpG dinucleotide is thought to be common variation position due to a high frequency of the methylation-induced deamination of 5-methyl cytosine. This process causes the variation from CpG to TpG and its complementary pair CpA, and thereby leading to the deficiency of CpG dinucleotide in human genome (Li and Chen, 2011). Meanwhile, CpG dinucleotide plays an essential role in many cellular functions, such as the gene expression which is controlled by the cytosine methylation status. Thus, there are conflicts between these two processes for instance the high variation frequency caused by the cytosine methylation damages CpG dinucleotide while functional processes require the preservation of CpG dinucleotide. Confused by such problem, Li and Chen (2011) conducted a research by analyzing the variation and frequency spectrum of newly derived alleles from the human genome. They found that there is a trend towards generating more CpGs, which was mainly contributed by high frequency variations from CpA/TpG to CpG. In other words, CpGs which suffer an enormous amount of decrease due to the cytosine methylation tend to be recreated from TpG and CpA rather than other dinucleotides (Kangaspeska *et al.,* 2008; Li and Chen, 2011).

Consequently, the study of dinucleotide CpG is of importance in investigating human genetic diseases which are significantly contributed by DNA methylation. Data source used in this thesis work contains massive human genetic variations caused by single nucleotide variations within gene coding regions. In order to find out how dinucleotide CpG of pathogenic and neutral types behave among all other dinucleotides, frequencies of CpG dinucleotide as well as all other dinucleotides were analyzed and compared for both neutral and pathogenic variations. Dinucleotide compositions in human genome were included for reference, and the location of variations has been taken into account in the study.

3. OBJECTIVES

The key objectives of this thesis project are to investigate whether pathogenic SNVs and neutral SNVs have different impact on the formation of nucleosome and therefore possibly could contribute to the development of pathogenicity predictors for novel variants. Key objectives of this project are:

- To investigate whether pathogenic and neutral variations have different localization along the DNA.
- To figure out whether pathogenic and neutral variations would cause different nucleosome binding affinities.
- To find out if pathogenic and non-disease causing variations have different nucleosome occupancy levels.
- To identify the nucleotide composition bias and substitution rates of pathogenic and neutral variants of large datasets.
- To measure dinucleotide variability and the degree of pathogenicity.

4. MATERIALS AND METHODS

4.1 Materials

4.1.1 Data source

Variation datasets used in the thesis work were downloaded from VariBench (Nair and Vihinen, 2013), a benchmark database for human variations, created and maintained by Institute of Biomedical Technology, University of Tampere, Finland. At present, it is maintained by Department of Experimental Medical Science, Lund University, Sweden. VariBench contains information for experimentally verified effects and datasets that have been used for developing, testing the performance of prediction tools and for training novel predictors in this field. Currently, VariBench datasets are capable of testing and training four different variations affecting:

– (a) protein tolerance

– (b) protein stability

– (c) transcription factor binding sites

– (d) splice sites

The neutral dataset comprising 21170 human non-synonymous coding SNVs was extracted from the dbSNP database build 131, while the pathogenic dataset containing 19335 missense SNVs was obtained from the PhenCode database (Nair and Vihinen, 2013). Due to the existence of data redundancy and empty entries, there were altogether 20973 unique entries selected from the neutral dataset whereas 19335 pieces of non-overlapping information were chosen from the pathogenic dataset. In consequence, there were altogether 20793 neutral and 18412 pathogenic sequences downloaded based on corresponding identifiers in FASTA format for further analysis. An overview of the data is presented in Table 4.1

Table 4.1 Summary of amount of SNVs in Varibench datasets and selected datasets

Datasets	**Neutral SNVs**	**Pathogenic SNVs**
VariBench datasets	21770	19335
Selected datasets	20973	18412

4.1.2 Tool used for nucleosome positioning prediction

Nucleosome positioning prediction engine (NuPoP) (bioconductor version 2.0.0) (Wang *et al.*, 2008; Xi *et al.*, 2010) was considered as the ideal tool among various outstanding software owing to its flexible and command-driven user interface and environment, and for which it was selected to predict preferential nucleosome positions from DNA sequences. NuPoP is built upon a duration hidden Markov model for both Watson and Crick strands, thus the results produced by NuPoP for both Watson and Crick strands are exactly same but in a reverse order. The 4th order Markov chain rather than the 1st order was chosen as the nucleosome or linker DNA state model due to its better performance. Three built-in functions including predNuPoP, readNuPoP and plotNuPoP are provided for nucleosome positioning prediction, prediction results read-in and prediction results visualization respectively.

NuPoP is capable of taking DNA sequence of any length in FASTA format as input, however due to boundary effects; a flanking sequence of 5000 bp was added around the sequence (site) of interest in this thesis project for prediction accuracy. NuPoP outputs the Viterbi prediction of optimal nucleosome position map and a file in plain text format. The text file includes five variables: position, P-start (probability for a position being the start of a nucleosome), nucleosome/liner state, nucleosome occupancy and nucleosome binding affinity scores. Information of specific nucleosome was extracted from predicted files by a Python script. For convenience, the R package version (2.0.0) was selected instead of the web server and FORTRAN in this thesis study.

4.1.3 Statistical analysis

R statistical computing environment (version 3.0.1) (R Core Team, 2013) was chosen here to perform tasks such as data prediction (with the integration of NuPoP), statistical analysis and data visualization.

4.2 Methods

4.2.1 Datasets preparation

4.2.1.1 Data filtration

The datasets preparation work was mainly conducted by Python scripts which extensively used Biopython modules (Cock *et al.*, 2009). There are altogether 21170 neutral SNVs and 19335 pathogenic SNVs in both variation datasets VariBench (Nair and Vihinen, 2013), out of which 20973 and 18412 entries were selected, respectively. In other words, 197 neutral SNVs and 923 pathogenic SNVs were filtered out due to the existence of empty entries and several overlapping entries. This was done semi-automatically as follows:

1. the removal of empty entries:

   First of all, a Python script was compiled to detect empty entries. All empty entries in both original neutral and pathogenic files were deleted manually based on line numbers of empty entries returned by the script.

2. the removal of overlapping entries:

   Secondly, a function called *RemovalOfOverlappingEntries* was created in the same script aiming to detect overlapping entries in both pathogenic and neutral datasets. In the neutral file, each set of information was relisted in an alphabetical order based on reference SNV identifiers (rs IDs). Function *RemovalOfOverlappingEntries* took a file storing rs IDs as parameter, and a file containing line numbers of overlapping entries was returned. Overlapping entries were deleted manually according to line numbers returned by the function. Due to the absence of rs IDs in the pathogenic file, the removal work was conducted semi-manually by invoking a Python script which compares the identity of each set of information in the entire dataset. Likewise, line numbers of overlapping entries were obtained and corresponding entries were removed accordingly.

4.2.1.2 The retrieval of strand information for neutral dataset

Given that the strand information is not provided in the neutral dataset and the need of which in the thesis work is indispensable, a Python script was compiled aiming to detect the information automatically. A workflow describing the basic idea of the algorithm is provided in Figure 4.2.

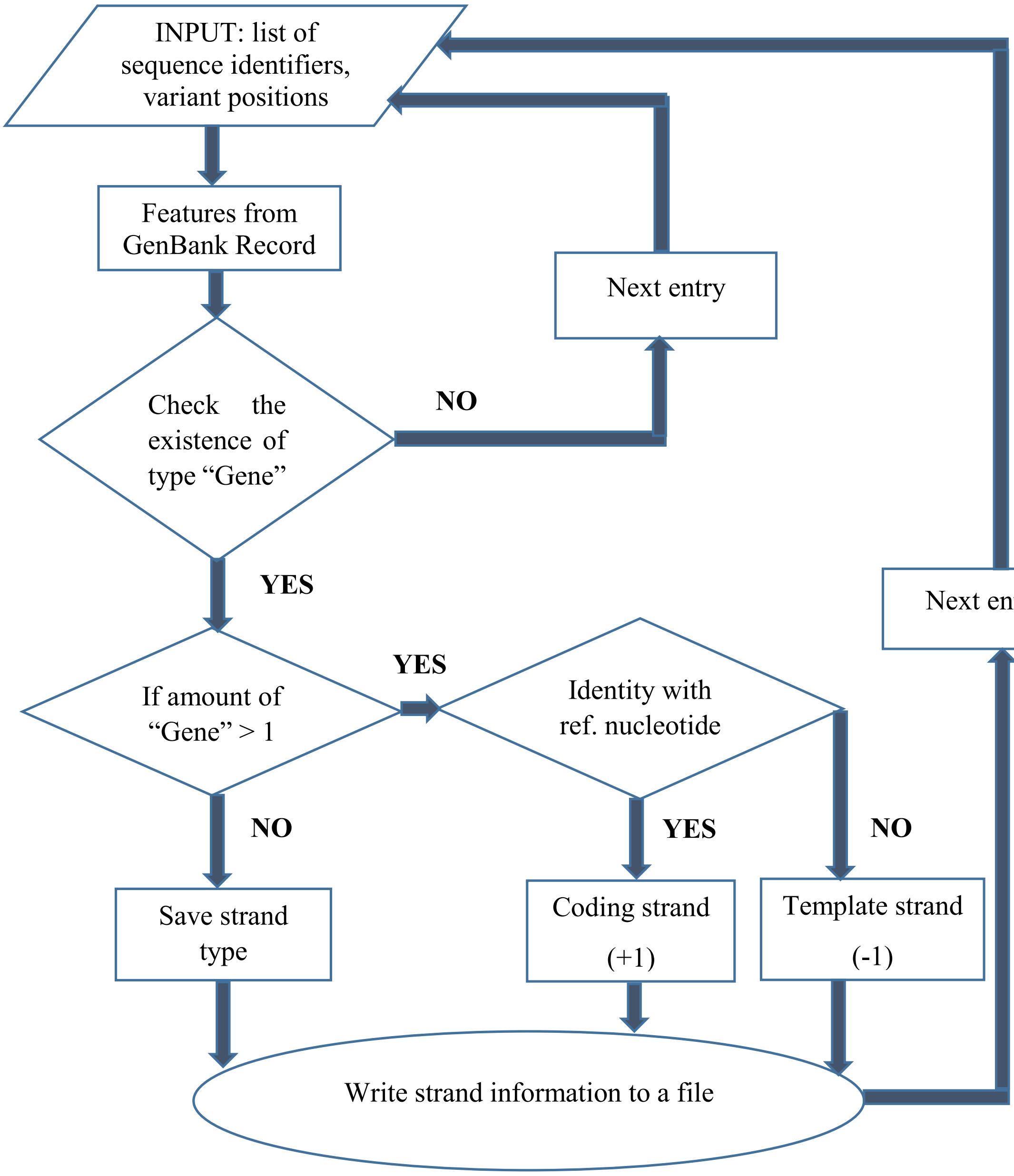
INPUT: list of sequence identifiers, variant positions
Features from GenBank Record
Next entry
Check the existence of type "Gene"
NO
YES
Next entry
If amount of "Gene" > 1
YES
Identity with ref. nucleotide
NO
YES
NO
Save strand type
Coding strand (+1)
Template strand (-1)
Write strand information to a file

Figure 4.2 A flowchart checking strand information for neutral variations according to provided sequence identifiers, variant positions and reference codons.

The algorithm consists of five steps:

1. Provide two text files in plain text format as input; one containing sequence identifiers, while the other having corresponding variant positions.
2. Download information of the site-specific variant in a GenBank record format by Biopython inbuilt function *Entrez.efetch* iteratively, under function *Entrez.efetch* the variant position was specified as parameters of *seq_start* and *seq_end.*

3. Check features of each variant record, and search type "Gene" under which strand type is given and marked as either 1 or -1.
4. Delete variants that are not located within genes, in other words, type "Gene" was not found.
5. Nucleotide can be within many genes (Sanna *et al.*, 2008); alternatively, many "Gene" types could be found, strand information was further checked by comparing the extracted variant to reference nucleotide provided in the neutral dataset. Identity with the reference nucleotide refers to coding strand whereas difference refers to template strand. Document retrieved information to a file.

4.2.1.3 The identification of reading frames and missense codons for pathogenic dataset

The pathogenic dataset was not as informative as the neutral dataset. Specifically, information for reading frames, the site-specific variant positions as well as missense codons was not provided. In other words, for information related to DNA only data for genomic IDs, reference codons, reference codon positions in three genomic coordinates and genomic strands were given. The information of the site-specific variation in reference codon and the missense codon is necessary to complete this thesis project. The detection of this information was carried out by a Python script based on several provided data, such as the protein variations (in HGVS format) and reference codons, etc.

First of all, reference codons were converted to strand-specific codons as the reference codons provided were on the coding strand. For cases of variations occurring on the template strand reference codons were given in a reverse order but not complemented. The Python script took two text files as input, one containing reference codons while the other comprising information of reference amino acids, variant positions as well as missense amino acids in HGVS format. Reading frame of each codon was detected automatically by the script according to reference codons, missense amino acids as well as information from DNA codon table which is listed in Table 9.7 in appendix. The general idea for the algorithm is given as follows:

1. Provide two files mentioned above as input; read all entries to a list with sub lists embedded. Each sub list is in a format of "reference codon, missense amino acid".
2. Parse through the list once at a time; check codons encoding specified missense amino acid from Table 9.7 in appendix. Pack the value (missense codon candidates) to a list.

3. Set a counting parameter *C* to check the amount of matches from the reference codon to retrieved missense codons for each entry. For instance, there are cases of no match, exactly one match and several matches. Examples for each case are listed in Table 4.3. A three-iteration loop was performed for each entry to check the reading frame and the counting parameter *C*. The 1st loop is to check whether frame is one, the two last nucleotides were extracted from reference codon and retrieved missense codon list to check whether the two last nucleotides of reference codon can be found in the extracted list. If this conditional statement is reached, *C* is increased by one. Likewise, 2nd loop is to check frame case of two by taking the 1st and 3rd nucleotides from both the reference codon and the missense codon list. If the 1st and 3rd nucleotides from the reference codon can be found in the extracted list, *C* is increased by one. Case of frame three was checked in a similar way.
4. Check the value of *C*. Value of zero indicates that more than one nucleotide has been altered (case of "no match" in the Table 4.3). However, value of one indicates a single nucleotide variation and thus the frame can be identified. Similarly, value greater than one (case of "several matches") means the frame is not able to be identified.
5. Write those entries for which frames can be identified (*C* equals to one) to a file. In addition, document corresponding missense codon(s) to another file. Furthermore, record line numbers of those entries which are not able to identify frames and delete them from the dataset. Parse the next entry iteratively.

Table 4.3 Special examples of checking reading frames based on reference codons and missense amino acids. Case "no match" indicates a non-SNV alteration, while case "one match" indicates a single nucleotide variation. Case "several matches" means reading frames are unable to be identified based only on information from reference codons and missense amino acids. "Ref", "Mis." and "AA" represent "reference", "missense" and "amino acid", respectively.

Cases	Ref. AA	Ref. codon	Mis. AA	Mis. codon	Conclusion
No match	Leucine (L)	CTG	Histidine (H)	CAT, CAC	both frame 2 and 3 are changed
One match	Glycine (G)	GGG	Glutamic acid (E)	GAA, GAG	GGG→GAG: only frame 2 has changed
Several matches	Cysteine (C)	TGC	Serine (S)	TCT, TCC, TCA, TCG, AGT, AGC	TGC→TCC: frame 2 TGC→AGC: frame 1

After the application of the Python script, 106 entries of case “no match” and 455 entries of case “several matches” were found. As a consequence, a total number of 561 entries were excluded from the pathogenic dataset because of the inability to identify reading frames and missense codons. The Python script in pseudocode is given as follows:

```
function retrieveFrame (referenceCodon, all_Missense_Codons_list)
        initialize readingFrame to 0; initialize count to 0
        set Missense_DiNucleotides_List to an empty list; set Ref_Dinucleotides_List to an empty list
        for i= 1 to 3 do
                initialize temp to an empty list
                if i==1 then
                        extract=referenceCodon[1:3]
                        for each item in all_Missense_Codons_list do
                                temp.append (item [1:3])
                        next
                else if i==2 then
                        extract= referenceCodon[0]+referenceCodon[2]
                        for each item in all_Missense_Codons_list do
                                temp.append (item [0] + item [2])
                        next
                else if i==3 then
                        extract=referenceCodon[:2]
                        for each item in all_Missense_Codons_list do
                                temp.append (item [:2])
                        next
                end if
                if extract can be found in list temp then
                        set readingFrame to i
                        increase count by 1
                end if
                Missense_DiNucleotides_List.append (temp)
                Ref_Dinucleotides_List.append (extract)
        next
        return count, readingFrame, Missense_DiNucleotides_List, Ref_Dinucleotides_List
end retrieveFrame function

function retrieveMissenseCodon (Missense_DiNucleotides_List, Ref_Dinucleotides_List, frame)
        initialize selected_Missense_Codon to an empty list
        list =Missense_Nucleotides_List [frame-1]
        extract =Ref_nucleotides_List [frame-1]
        for each index in list do
                if list [index] ==extract then
                        selected_Missense_Codon.append (all_Missense_Codons_list [index])
                end if
        next
        return selected_Missense_Codon
end retrieveMissenseCodon function
```

4.2.1.4 The identification of site-specific variation positions for pathogenic SNVs

Reference codon positions in the pathogenic dataset were given a three genomic coordinates format, thus the identification of the site-specific variation in the reference codon was necessary in conducting the comparison work between neutral and pathogenic variants.

This task was performed by a Python script. The program took a text file comprising variation positions of reference codons in a three genomic coordinates. Another file consisting of reading frames was also provided. A repetition loop was performed to extract the site-specific variation positions based on specified reading frames. The script in pseudocode is given as follows.

```
function retrieveUniquePositions (variationCoordinatesList, readingFrameList)
        define uniquePositions as an empty list
                for i=1 to length (variationCoordinatesList) do
                          coordinateList=variationCoordinatesList[i-1]
                          frame=readingFrameList[i-1]
                          uniquePositions.append (coordinateList [frame-1])
                  next
        return uniquePositions
end retrieveUniquePositions function
```

4.2.1.5 The identification of reference and missense nucleotides for pathogenic dataset

In this project, the comparison between neutral and pathogenic variations was subdivided into four nucleotides A, C, G and T. Due to the lack of this information in the pathogenic dataset, the identification of reference nucleotides and missense nucleotides was much needed. Based on information of reading frames and missense codons retrieved from previous studies, the task of identifying reference and missense nucleotides was simple. The Python script took three files containing reference codons, missense codons and reading frames as input. Reference nucleotides and missense nucleotides were extracted based on codons and corresponding reading frames. The retrieved reference and missense nucleotides were written to a file, and all previously retrieved data were collected to the pathogenic dataset. The Python script in pseudocode is given as follows.

```
function retrieveRefMissenseNucleotides (referenceCodonList, missenseCodonList, readingFrameList)
          initialize referenceNucleotide and missenseNucleotide as two empty lists
         for i=1 to length (readingFrameList) do
                 referenceNucleotide.append (referenceCodonList[i] [readingFrameList[i]-1])
                 missenseNucleotide.append (missenseCodonList[i] [readingFrameList[i]-1])
          next
         return referenceNucleotide, missenseNucleotide
end retrieveRefMissenseNucleotides function
```

4.2.2 The retrieval of DNA sequences and extraction of dinucleotides

All neutral and pathogenic sequences were downloaded based on their identifiers from the Nucleotide database, NCBI (National Center for Biotechnology Information) by a Python script. Although NuPoP is capable of taking DNA sequence of any length, 5000 bp flanking sequence around the variant site was added due to boundary effects and prediction efficiency. As a consequence, DNA sequences of length 10001 bp in FASTA format were submitted to NuPoP for nucleosome positioning prediction. As genes are located both on the coding strand (aka Crick strand, strand +1) and the template strand (aka Watson strand, strand −1), DNA sequences were downloaded based on the strand where variations have occurred to make sure the SNVs studied are within the genes. Dinucleotides, the variant nucleotide followed by one nucleotide after, were extracted simultaneously for further studies. Each file was named with format "identifier_position", where "identifier" refers to the DNA identifier and "position" refers to the variation position (bp) on DNA. The workflow is given in the Figure 4.4.

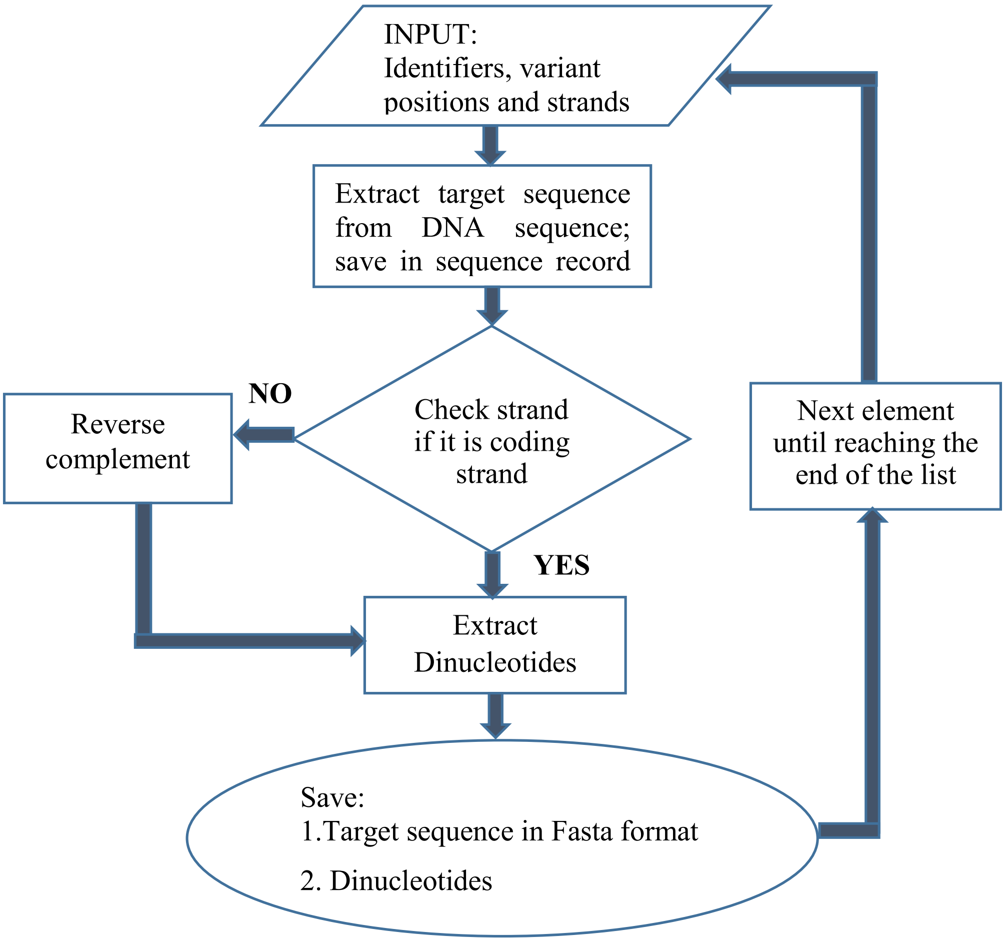
INPUT:
Identifiers, variant positions and strands
Extract target sequence from DNA sequence; save in sequence record
Check strand if it is coding strand
NO
Reverse complement
YES
Extract Dinucleotides
Save:
1.Target sequence in Fasta format
2. Dinucleotides
Next element until reaching the end of the list

Figure 4.4 The workflow of the Python script extracting dinucleotides and input sequences needed for NuPoP to predict nucleosome positions.

4.2.3 Pipelines for nucleosome positioning prediction and information extraction

This task was performed by both R and Python scripts. It includes two steps, the first step was performed by R while the other by Python. In particular, in the first step, a file containing identifiers of DNA sequences were submitted to NuPoP for prediction, and in the second step, information from files produced by NuPoP were extracted.

4.2.3.1 NuPoP in predicting nucleosome positions

A plain-text format file which stores identifiers of DNA sequences was submitted to R. A repetitive execution compliance with the invocation of library NuPoP were performed to predict nucleosome positions, binding affinities, occupancy scores, etc.

4.2.3.2 Extraction of information from predicted files

SNVs were studied based on the location of variations, variations within nucleosome core regions and at linker regions. NuPoP marks '1' for nucleosomal DNA while '0' for linker DNA. A Python script was compiled to extract information from files produced by NuPoP, and the information was written to an excel file for the final statistical analysis. Specifically, the basic algorithm was designed as follows:

1. Input a plain text file containing filenames of files produced by NuPoP to the script; iteratively process one entry at a time.
2. If variant is within nucleosome core regions, spread the search from both upstream and downstream sides until the 1st nucleotide at linker regions (symbol "0") is found.
3. If the variant is at linker regions, search neighboring nucleosomes from upstream and downstream sequences. In particular, expand the search from both sides until 1st nucleotide within nucleosome core regions (symbol "1") is reached; continue the search until the 1st nucleotide at linker regions (symbol "0") is found.
4. Calculate variant positions (bp), nucleosome binding affinity and occupancy scores.
5. Write each set of information produced in step 4 to an Excel file
6. Iteratively repeat steps mentioned above.

The workflow of the algorithm is given in Figure 4.5.

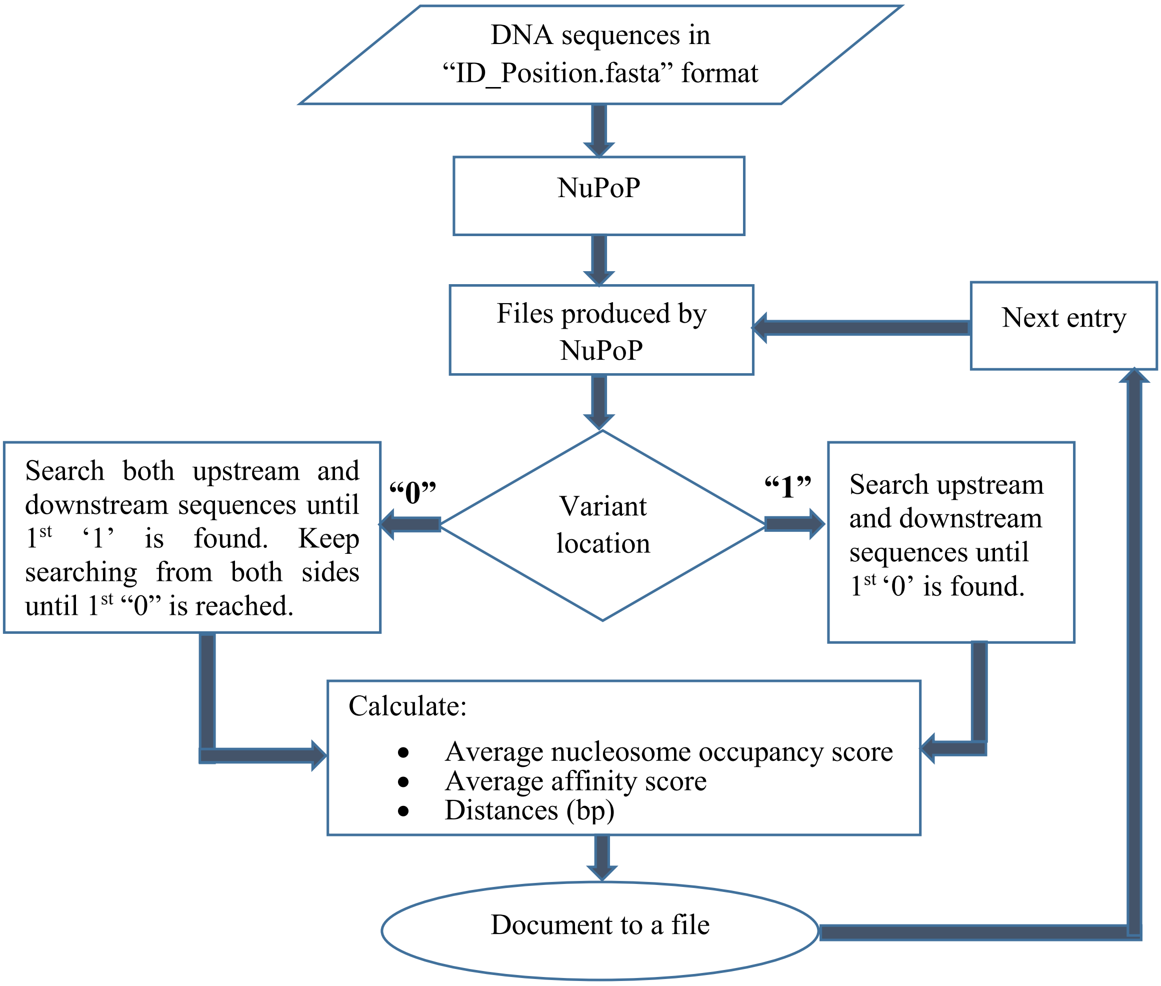
DNA sequences in
"ID_Position.fasta" format
NuPoP
Files produced by NuPoP
Next entry
Variant location
"0"
"1"
Search both upstream and downstream sequences until 1st '1' is found. Keep searching from both sides until 1st "0" is reached.
Search upstream and downstream sequences until 1st '0' is found.
Calculate:
Average nucleosome occupancy score
Average affinity score
Distances (bp)
Document to a file

Figure 4.5 A flowchart showing processes of nucleosome positioning predictions and the extraction of information from predicted files. Symbols "1" and "0" in the workflow indicate variants within nucleosome core regions and at linker regions, respectively.

4.2.4 Statistical analysis

Distributions of variations along DNA, nucleosome affinity and occupancy levels were statistically analyzed in order to explore if neutral and pathogenic SNVs have different impact on these aspects and if any, how they differ from each other. Several graphs, e.g. boxplots, bar charts, line charts, etc. in R were drawn to present, visualize and compare both types of data not only in an overall view but also at nucleotide classes (purine and pyrimidine) and individual nucleotide levels. In addition, distributions grouped by nucleotide substitutions and substitution types (transitions and transversions) were also considered. In order to interpret data statistically, the Mann-Whitney U test (MW test) was selected and applied to perform statistical tests for these comparisons.

5. RESULTS

5.1 Visualization of variant positions within nucleosome core regions and at linker regions

The study of variant positions within nucleosome core regions and at linker regions was performed to investigate whether neutral and pathogenic SNVs have different positioning distribution. According to statistical analysis, approximately 85.32% neutral variants were observed within nucleosome core regions whereas 14.68% variants were positioned at linker regions. The proportions of pathogenic variants within nucleosome core regions and at linker regions were 89.02% and 10.98%, respectively. For variants within nucleosome core regions, distances from themselves to the nucleosome start site were calculated, while distances from variants at linker regions to end sites of their neighboring upstream and downstream nucleosomes were counted.

As shown in Figure 5.1, there was no obvious positioning difference between neutral and pathogenic variants irrespective of the location of variations (within nucleosome core regions and at linker regions). The Mann-Whitney U test for variations within nucleosome core regions also showed a p-value of up to 0.952 indicating an identical distribution between pathogenic and neutral variations within nucleosome core regions. Moreover, the figure reveals that both types of variants distributed relatively equal within nucleosome core regions (147 bp). Hence, variant positioning difference between pathogenic and neutral types were not observed.

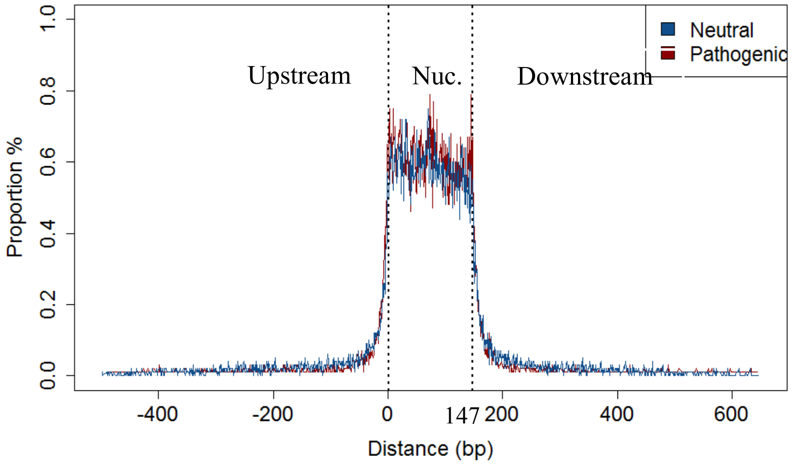

Figure 5.1 Visualization of variant positions (bp) within nucleosome core regions and at linker regions. Position 0 and 147 in the plot (black dotted vertical lines) correspond to nucleosome start and end sites, respectively. Variation distances to neighboring downstream nucleosomes include length of nucleosomal DNA themselves (147 bp).

5.2 Nucleosome binding affinity

5.2.1 Overall comparison of nucleosome binding affinity scores

Nucleosome binding affinity scores of all selected neutral and pathogenic variants were systematically calculated and statistically plotted. As demonstrated in Figure 5.2, average affinity score was higher for pathogenic variations than that of neutral type despite the location of variations (within nucleosome core regions and at linker regions). Inter-quartile distance showed a lower trend in pathogenic variations in Fig.5.2.a. However, in Fig.5.2.b and Fig.5.2.c, inter-quartile distances were higher for pathogenic variants. In particular, In Fig.5.2.a, first quartile, median and third quartile affinity scores for pathogenic variations were 8.16, 15.06, and 19.36; for neutral variations 5.04, 12.53 and 18.74, respectively. In Fig.5.2.b, these values were 1.10, 7.53 and 14.91 for pathogenic variations while -1.87, 3.12 and 9.27 for neutral variations. The similar trend in Fig.5.2.b can be also found in Fig.5.2.c, with values of 0.87, 7.38, 14.81for pathogenic variations whereas -1.77, 3.42 and 9.69 for neutral variations. Notably, for both pathogenic and neutral variants within nucleosome core regions in Fig.5.2.a, the nucleosome binding affinity was evidently higher than variants at linker regions in Fig.5.2.b and Fig.5.2.c. Further, a very similar distribution pattern was found for variations at upstream and downstream linker regions (Fig.5.2.b and Fig.5.2.c).

5.2.1.1 The Mann-Whitney U test

The Mann-Whitney U test (also known as the two sample Wilcoxon's test) was applied to examine the overall distribution of nucleosome binding affinity scores between pathogenic and neutral types of variations within nucleosome core regions and at linker regions. For variations within nucleosome core regions (Fig.5.2.a), the null hypothesis was assumed as identical affinity mean values between two groups. The alternative hypothesis was considered as distinct affinity mean values between neutral and pathogenic variations. The Mann-Whitney U test showed a p-value of less than $2.2*10^{-16}$ which indicates the rejection of the null hypothesis. Thus, significant difference in binding affinity scores between pathogenic and neutral variations were observed. Same test and hypothesis were set for variations at linker regions. Similar p-values (p-value< $2.2*10^{-16}$) were found for both variants at upstream and downstream linker regions when performing the MW test and a p-value less than 0.05 suggests the acceptance of the alternative hypothesis. Boxplots in Figure 5.2 depicts a higher affinity score for pathogenic type of variations than the neutral type.

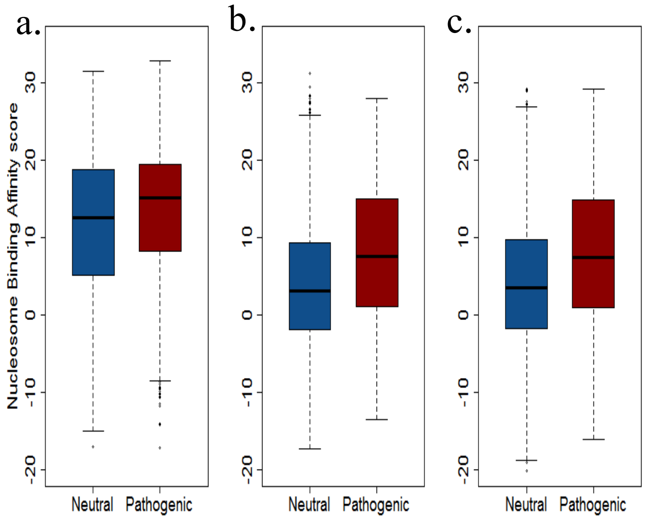

Figure 5.2 Distribution of overall nucleosome binding affinity scores between neutral and pathogenic variations. (a) Variations within nucleosome core regions (b) Variations at upstream linker regions. (c) Variations at downstream linker regions.

5.2.2 Comparison of nucleosome binding affinity scores for individual nucleotides

As the overall comparison of nucleosome binding affinity scores was different between pathogenic and neutral types of variants, affinity comparisons for individual nucleotides were performed and plotted to observe a deeper insight, if any. Figure 5.3 illustrates distribution of affinity scores between neutral and pathogenic variations grouped by individual nucleotides. Notches in box plots suggest a rough guide of having significant difference in medians. Specifically, if notches of two plots do not overlap then this is “strong evidence” that the two medians differ. As can be clearly seen from the figure, no overlapping notches were observed between each listed group, thus we could roughly conclude that there was affinity difference between pathogenic and neutral variations for each nucleotide group. Consistently, first quartile, median and third quartile of pathogenic variations all scored higher values than those of neutral type. In other words, nucleosome binding affinity of pathogenic variations showed a considerably higher intensity than that of neutral type in all listed groups in Figure 5.3. Notably, both pathogenic and neutral variants within nucleosome core regions displayed stronger binding affinity than those of linker variants. A detailed view of these values is provided in Table 9.8 in appendix.

5.2.2.1 The Mann-Whitney U test for individual nucleotides

MV tests were performed to analyze and compare nucleosome biding affinity between pathogenic and neutral types of variations in individual nucleotide groups. The null hypothesis was considered as no significant affinity difference between pathogenic and neutral variations. In contrast, the alternative hypothesis was set as distinct affinity between pathogenic and neutral types. The results from all MW tests showed p-values of less than 0.05 which indicates the rejection of null hypothesis and proves that pathogenic variations have a higher binding affinity than neutral variations (Figure 5.3). Moreover, it is notable that nucleotide C and G showed relatively higher affinity scores than those of A and T in all observations listed in Figure 5.3.

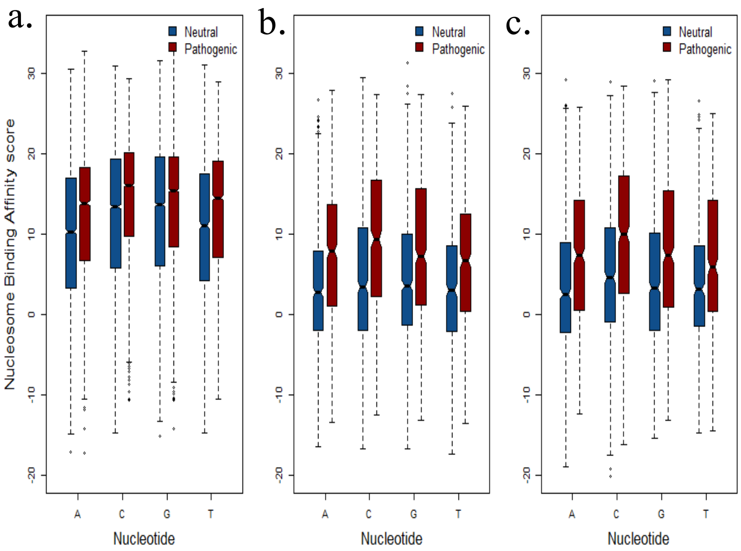

Figure 5.3 Distribution of nucleosome binding affinity scores between neutral and pathogenic variations grouped by individual nucleotides. (a) Variations within nucleosome core regions (b) Variations at upstream linker regions. (c) Variations at downstream linker regions.

5.2.3 Comparison of nucleosome binding affinity scores in other aspects

For a deeper view of the data, nucleosome binding affinity scores of pathogenic and neutral variations were compared according to nucleotide classes (purine and pyrimidine), substitution patterns and substitution types (transitions and transversions) respectively. Features shown in both overall and individual comparisons (Figure 5.2 and Figure 5.3) have been found in these aspects respectively. Distributions of affinity scores grouped by nucleotide classes, substitution types and substitution patterns were plotted in Figure 9.1, 9.2 and 9.3 in appendix, respectively.

Figure 9.1, 9.2 and 9.3 demonstrate a same trend as previous studies that binding affinity of neutral variations showed a lower trend than that of pathogenic type irrespective of the location of variations; both pathogenic and neutral variants within nucleosome core regions (Fig. 9.1.a, Fig.9.2.a and Fig.9.3.a) revealed a higher binding affinity than variants at linker regions. In addition, for variations within nucleosome core regions, inter-quartile ranges of pathogenic type were obviously narrower in comparison with those of neutral type (Fig.9.1.a, Fig.9.2.a and Fig.9.3.a), while a reverse trend was observed for variations at linker regions. Detailed 1st quartile, median and 3rd quartile values of both pathogenic and neutral variants are presented in Figure 9.9.

Notably, there was no prominent difference in the distribution patterns of both neutral and pathogenic variations between purine and pyrimidine groups in Figure 9.1. In other words, groups of purine and pyrimidine showed a similar affinity distribution pattern irrespective of the type and location of variations. Figure 9.2 presents comparisons of nucleosome binding affinity scores between neutral and pathogenic variations classified by substitution types (transition and transversions). Analogously, same trend shown in previous studies has also been found here. The non-overlapping notches in each boxplot suggest strong evidence that difference between medians of transition and transversion groups were significant. Similar to Figure 9.1, transition and transversion groups displayed a similar affinity distribution pattern irrespective of the type and location of variations. Figure 9.3 illustrates the distribution and comparison of nucleosome binding affinities between neutral and pathogenic types of variations classified by nucleotide substitutions. In addition to the same features found in previous studies, larger difference in the binding affinity was observed in substitution patterns A→T, C→A and T →A in Fig.9.3.a; A →C, C →A and G →T in Fig.9.3.b and Fig.9.3.c. MW tests were applied individually to aspects of nucleotide classes, nucleotide substitutions and substitution types. The MW tests showed p-values of less than 0.05 for all of these aspects, which indicates the significant difference in binding affinity between pathogenic and neutral variations.

5.3 Nucleosome occupancy level

5.3.1 Overall comparison of nucleosome occupancy scores

Nucleosome occupancy scores were systematically studied in this thesis work to investigate whether there is significant difference in the mean values of nucleosome occupancy scores between neutral and pathogenic variants. According to statistical analysis, occupancy scores of neutral variations vary from 0.110 for lowest to 1.000 for highest, while values of pathogenic variations range from 0.206 to 1.000. A variety of outliers were observed for both pathogenic and neutral variations.

As illustrated in Figure 5.4, pathogenic variants showed higher occupancy scores than those of neutral type irrespective of the location of variations. Inter-quartile ranges have been found fairly wider for neutral variations. In particular, inter-quartile distance of neutral variants within nucleosome core regions was 0.049 whereas it was 0.043 for pathogenic type (Fig.5.4.a). In Fig.5.4.b and Fig.5.4.c, values of neutral and pathogenic types were 0.09 and 0.07; 0.09 and 0.06, respectively. Precisely, in Fig 5.4.a, first quartile, median and third quartile occupancy scores for pathogenic variations were 0.939 0.965 and 0.983 whereas 0.933, 0.964 and 0.982 for neutral variations. These values were generally scored smaller in Fig 5.4.b, with values of 0.911, 0.957 and 0.977 for pathogenic variations whereas 0.880, 0.940 and 0.973 for neutral variations. Similar values could be found in Fig.5.4.b and Fig.5.4.c. In particular, these values were 0.911, 0.952 and 0.974 for pathogenic variations whereas 0.886, 0.943 and 0.973 for neutral variations. Notably, both types of variations within nucleosome core regions (Fig 5.4.a) showed an evidently higher occupancy distribution in comparison with that of linker variations (Fig.5.4.b and Fig.5.4.c).

5.3.1.1 The Mann-Whitney U test

MW tests were performed to check the overall distribution of nucleosome occupancy scores of both pathogenic and neutral variations inside and outside of nucleosomes. The null hypothesis was assumed as identical occupancy mean values between neutral and pathogenic groups, while the alternative hypothesis was considered as significant difference in the distribution of occupancy mean values between these two groups. All MW tests applied for comparisons shown in Fig.5.4.a, .5.4.b and .5.4.c returned p-values below 0.05 which indicates the rejection of the null hypothesis. Thus, significant difference in occupancy mean values between pathogenic and neutral variations was observed. Boxplots in Fig 5.4 depicts a higher occupancy score for pathogenic variations than neutral type.

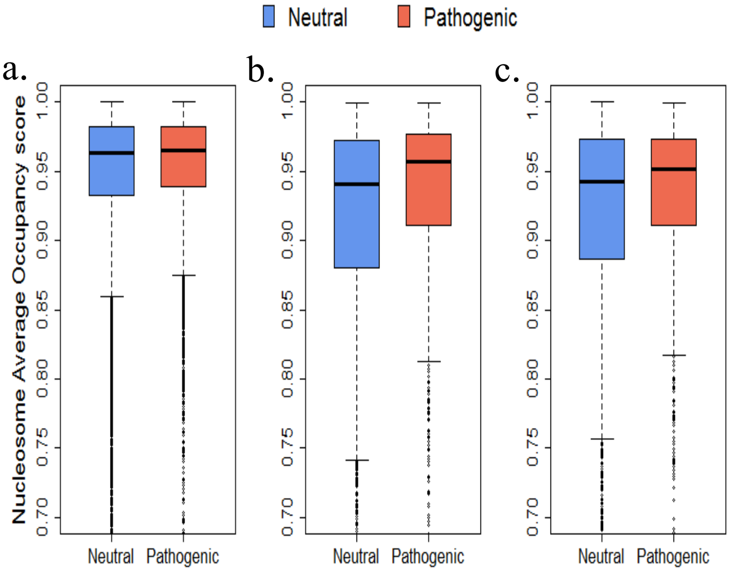

Figure 5.4 Distribution of overall nucleosome occupancy scores between neutral and pathogenic variations. (a) Variations within nucleosome core regions (b) Variations at upstream linker regions. (c) Variations at downstream linker regions.

5.3.2 Comparison of nucleosome occupancy scores for individual nucleotides

As the overall comparison of nucleosome occupancy scores were different between pathogenic and neutral types, occupancy comparisons at individual nucleotide level were statistically applied for a deeper view. As shown in Figure 5.5, pathogenic variants showed a general trend to gain higher occupancy scores than those of neutral type in each nucleotide group irrespective of the location of variations, and this trend was especially apparent for variations at linker regions (Fig.5.5.b and Fig.5.5.c). Moreover, variations within nucleosome regions displayed a considerable higher occupancy distribution than those of linker variations. Consistently, 1st quartile, median and 3rd quartile occupancy scores were dramatically higher for variations within nucleosome core regions, and these values were higher for pathogenic variations than neutral type in each nucleotide group. A detailed view of values of 1st quartile, median and 3rd quartile scores is presented in Table 9.11.

5.3.2.1 The Mann-Whitney U test for individual nucleotides

As figures only present general view of the data distribution, Mann-Whitney U tests were applied to examine if difference shown in figures are statistically significant. The null hypothesis was considered as no significant difference in occupancy mean values between pathogenic and neutral types in each nucleotide group. In contrast, the alternative hypothesis was assumed as significant occupancy mean values between the two types in each of four nucleotide groups. Results from MW tests are presented in Table 5.6. For variations within nucleosome core regions (Fig.5.5.a), results from all MW tests showed p-values below 0.05 except for nucleotide C (p-value=0.123). In other words, there is no significant difference in nucleosome occupancy level between pathogenic and neutral variations at nucleotide level C. For variations at upstream linker regions, all p-values scored below 0.05 which indicates the rejection of the null hypothesis, hence pathogenic variations at upstream linker regions express higher nucleosome occupancy than neutral type for each nucleotide. In Fig.5.5.c, both nucleotides C and T were observed with relatively high p-values of 0.1578 and 0.3372, respectively, which means no significant occupancy difference between pathogenic and neutral variations for nucleotide C and G was observed.

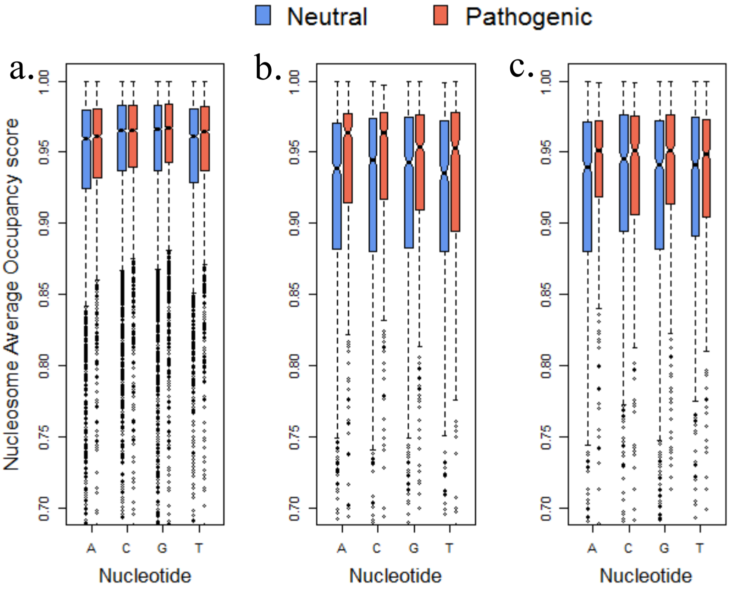

Figure 5.5 Distribution of nucleosome occupancy scores between neutral and pathogenic variations grouped by individual nucleotides. (a) Variations within nucleosome core regions (b) Variations at upstream linker regions. (c) Variations at downstream linker regions.

Table 5.6 Mann-Whitney U test results for pathogenic and neutral variations grouped by individual nucleotides. Cells shaded by yellow color indicate significant occupancy difference (p-value <0.05) between pathogenic and neutral types of variations.

Nucleotide	Nuc. variations p-value	Linker variations p-value (upstream)	Linker variations p-value (downstream)
A	$7.164*10^{-4}$	$2.959*10^{-8}$	$7.597*10^{-4}$
C	0.1230	$9.050*10^{-7}$	0.1578
G	$1.363*10^{-4}$	0.004553	$5.053*10^{-5}$
T	$3.373*10^{-4}$	0.005994	0.3372

5.3.3 Comparison of nucleosome occupancy scores in other aspects

For a better understanding of the data, nucleosome occupancy scores of pathogenic and neutral variations were compared in aspects such as nucleotide classes (Figure 9.4), substitution types (transitions and transversions) (Figure 9.5) and nucleotide substitutions (Figure 9.6), respectively. The distribution of occupancy scores in these aspects showed similar features to the trend observed in overall (Figure 5.4) and individual nucleotide comparisons (Figure 5.5). In particular, variations inside nucleosomes generally showed a higher occupancy distribution than variants at linker regions irrespective of the type of variations. Moreover, pathogenic variants were found to tend to gain higher occupancy scores than neutral type irrespective of the location of variations except for few cases that no significant difference were found. For instance, when performing MW tests for variations categorized by nucleotide classes, no significant occupancy difference was found between pathogenic and neutral variations (at downstream linker regions) of pyrimidine group (Fig.9.4.c). The distribution of occupancy scores for nucleotide classes, substitution types and nucleotide substitutions are presented in Figure 9.4, 9.5 and 9.6 in appendix, respectively.

Mann-Whitney U tests were performed for all these aspects, results of which were shown in Table 9.12 (nucleotide classes), 9.13 (substitution types) and 5.7 (nucleotide substitutions). According to Table 9.12, only the pyrimidine group for variations at downstream linker regions showed a non-significant difference (p-value= 0.1329) in mean values of nucleosome occupancy between pathogenic and neutral types. As demonstrated in Table 9.13, all p-values for groups of transitions and transversions were less than 0.05 which indicates significant occupancy difference in the observation. In Table 5.7, all p-values lower than 0.05 are highlighted in color yellow. For variations within nucleosome core regions, five substitution patterns were observed with p-values less than 0.05 which indicates significant nucleosome occupancy difference between pathogenic and neutral variations for substitutions A→T, C→A, G→C, T→A and T→C. Analogously, for variations at upstream linker regions, six patterns were observed with p-values of lower than 0.05 depicting the existence of significant difference whereas only three patterns were found for variations at downstream linker regions.

Table 5.7 Mann–Whitney U test results for pathogenic and neutral variations grouped by substitution patterns. Cells shaded by yellow color indicate significant occupancy difference (p-value <0.05) between pathogenic and neutral types of variations.

Substitution patterns	**Variants within Nuc. core regions P-value**	**Variants at upstream linker regions P-value**	**Variants at downstream linker regions P-value**
A →C	0.3416	$1.878 * 10^{-6}$	0.06964
A →G	0.2751	0.004167	0.01322
A →T	0.003168	0.02357	0.1161
C →A	0.001204	0.009996	0.5518
C →G	0.2609	$2.555 * 10^{-5}$	0.07717
C →T	0.06158	0.1758	0.6258
G →A	0.05391	0.3532	0.09914
G →C	0.004629	0.04987	0.005012
G →T	0.1291	0.1147	0.02041
T →A	0.002401	0.2337	0.9258
T →C	0.005975	0.08534	0.6794
T →G	0.8475	0.06401	0.5615

5.4 Nucleotide composition bias and nucleotide substitution rates

19335 and 18412 entries of SNVs were eventually selected for neutral and pathogenic datasets to conduct the analysis work. Among these entries, proportions of variations grouped by individual nucleotides were statistically analyzed and plotted in R.

To examine the nucleotide composition bias from both neutral and pathogenic variants, Figure 5.8 was drawn to show the observation. The overall nucleotide compositions in the human genome are 29.55% A, 20.44% C, 20.46% G and 29.54% T (Zhao and Boerwinkle, 2002). Nucleotide proportions in human genome were drawn as background for reference.

Compositions of the four nucleotides from neutral dataset were 22.55% A, 28.29% C, 35.09% G and 14.07% T, while proportions were 17.29% A, 26.37% C, 36.29% G and 20.06% T from pathogenic dataset. In Figure 5.8, it is obvious that proportions of nucleotide A and T of type neutral and pathogenic are much lower when compared to their expected proportions in human genome. In contrast, proportions of C and G from both datasets were found much higher than expected proportions. Thus, nucleotide C and G, especially G, might have a higher probability upon variations. In other words, nucleotide A and T showed lower degree of variability.

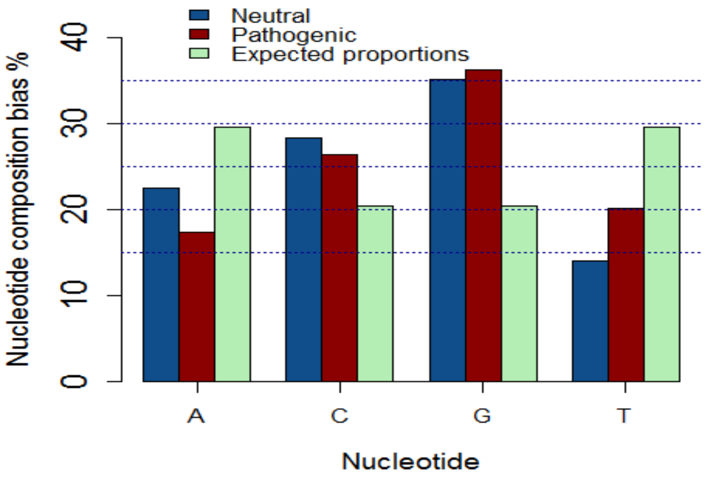

Figure 5.8 Distribution of individual nucleotide compositions of neutral and pathogenic variations as well as the expected proportions in human genome.

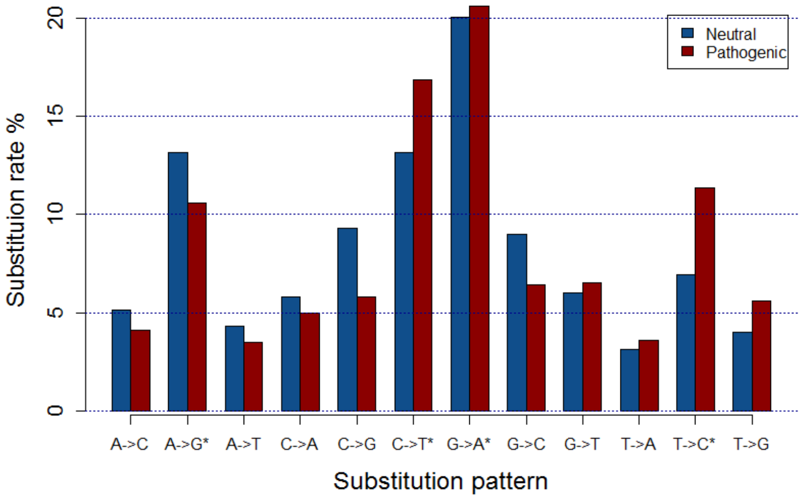

Figure 5.9 Distribution of nucleotide substitution rate of each substitution pattern of pathogenic and neutral types. Transition variations are indicated by asterisk (*), and unmarked ones are transversion variations.

For a deeper view, Figure 5.9 was plotted to observe the substitution rate of each substitution pattern of pathogenic and neutral types. As shown in the figure, both types of variations were found extremely frequent for several substitution patterns, e.g. substitution C→T and G→A. Although nucleotides C and G have been observed with high variation frequencies (Figure 5.8), substitution frequencies of C→A, C→G, G→C and G→T were dramatically lower when compared to C→T and G→A. Similar trend was observed for substitutions derived from nucleotides A and T. In particular, substitution frequencies of A→G and T→C were much higher than those of e.g. A→T and T→A. Interestingly, substitution patterns C→T, G→A, A→G and A→G all belong to transition variations. Hence, transition variations were observed with a considerably higher variation frequency than that of transversion variations.

Table 5.10 illustrates the detailed substitution rate of each substitution pattern of neutral and pathogenic types. Substitution patterns belonging to transition variations are highlighted in red asterisk (*), and substitution patterns showing higher variation frequency for pathogenic type are shaded with color yellow. Three out of four types of transition variations displayed higher variation tendency to be converted into pathogenic type, except for pattern A→G. Moreover, three out of eight types of transversion variations, e.g. G→T, T→A and T→G, expressed higher probability to become pathogenic variations.

Table 5.10 Illustration of substitution rate of each substitution pattern of pathogenic and neutral types. Transition variations are indicated by red asterisks (*) whereas unmarked ones are transversion variations. Cells highlighted with color yellow indicate higher substitution rate for pathogenic variations than neutral type. "N." and "P." represent "Neutral" and "Pathogenic", respectively.

Substitution pattern	Substitution rate (%) N.	Substitution rate (%) P.	Substitution pattern	Substitution rate (%) N.	Substitution rate (%) P.
A→ C	5.11	4.09	**G→ A** *	20.09	20.65
A→ G *	13.15	10.61	**G→ C**	8.99	6.42
A→ T	4.30	3.46	**G→ T**	6.02	6.51
C→ A	5.81	5.00	**T→ A**	3.14	3.60
C→ G	9.32	5.82	**T→ C** *	6.94	11.38
C→ T *	13.16	16.86	**T→ G**	3.99	5.59

5.5 Dinucleotide variability and the degree of pathogenicity

DNA methylation at the 5’ cytosine has been found to cause the reduction of gene expressions, and DNA methylation typically occurs at the CpG dinucleotide content. The hypermethylation of CpG islands in gene promoter region may cause gene silencing. For instance, the silencing of oncogene suppressor contributes to a higher probability of causing cancer. In contrast, the hypomethylation has been thought to cause chromosomal instability and the loss of imprinting (Daura-Oller *et al.,* 2009). As a consequence, the study of CpG dinucleotide content is of significance to investigate upon how these large genome scale disorders affect DNA regulation.

In this thesis work, variations of both pathogenic and neutral types were analyzed in all sixteen different dinucleotide contents. Variation proportion in CpG dinucleotide content was compared to non-CpG dinucleotides together with expected dinucleotide proportions in human genome for reference. In addition, comparisons were conducted both for variations within nucleosome core regions and at linker regions.

Fig.5.11.a shows the comparison of dinucleotide proportions between pathogenic and neutral variations within nucleosome core regions. As the number of pathogenic and neutral variations is different in provided datasets, changes were calculated according to their respective proportions (%). As exhibited in Fig.5.11.a, dinucleotides CpC, CpG and their complementary pattern GpG, GpC of both pathogenic and neutral types showed a considerably higher distribution than other dinuclotides in comparison with their expected proportions in human genome. Alternatively, these dinucleotides expressed a higher variation tendency. In particular, as illustrated in Table 5.12, compositions (%) of CpG of pathogenic and neutral types were 10.99 and 9.19 respectively whereas its expected proportion in human genome is only 4.18. Likewise, the composition of pathogenic GpG was higher than its neutral type, with a percentage of 11.34 and 8.09, respectively.

In contrast, dinucleotides ApA and ApT and their complementary pattern TpT and TpA of both pathogenic and neutral types displayed significant lower proportions in comparison with their expected proportions in human genome, hence these dinucleotides are less likely to be mutated. For instance, expected proportion of TpT in human genome was found almost three times more than the proportions of its pathogenic and neutral types. Some dinucleotides (CpC, GpC) showed almost equal proportions. For instance, compositions of CpC of pathogenic and neutral types remained at the same level, steadily remaining at 7.38 and 7.43, respectively. In other words, these kinds of dinucleotides display an equal variation tendency to be changed into either pathogenic or neutral types. A detailed view of these dinucleotide compositions is presented in Table 5.12. Dinucleotides which show higher composition for pathogenic variations are highlighted in yellow.

The same study was repeated for variants destined at linker regions (Fig.5.11.b). Most of the features shown in Fig.5.11.a was also found in Fig.5.11.b but slightly differs in the degree of variability. Several dinucleotides e.g. CpC, CpG, GpG, TpG, etc., which apparently showed higher variability in Fig.11.a, showed a similar trend also in Fig.5.11.b. Dinucleotide GpAs of pathogenic and neutral types at linker regions were observed with highest proportions, showing almost a twofold increase in comparison with its expected proportion in human genome. In addition, both types of CpC and GpC were observed with a very similar distribution which means they show equal variability, and this feature is consistent with trends observed in Fig.5.11.a. However, major differences between Fig.5.11.a and Fig.5.11.b were also observed. In particular, dinucleotide ApTs of both pathogenic and neutral types within nucleosome core regions (Fig.5.11.a) showed lower distributions in comparison with its expected proportion in human genome, which means ApTs within nucleosome core regions are less abundant to be mutated. However, the opposite point of this view has been found in Fig.5.11.b. Specifically, proportions of ApTs of pathogenic and neutral types were observed higher than the expected proportion in human genome, hence showing a higher variation tendency, provided variations occur at linker regions.

Fig.5.11.c demonstrates the degree of pathogenicity of each dinucleotide within nucleosome core regions and at linker regions. Logarithm ratios of dinucleotides were calculated by dividing proportions of neutral variations with pathogenic type. Line with asterisk (-*-) in blue indicates distribution of variants within nucleosome core regions, while line with triangle (-Δ-) represents distribution of variations at linker regions. A negative logarithm ratio means a higher tendency to be converted into pathogenic variations whereas positive value depicts a lower degree of variability in becoming pathogenic type. A dotted line highlighted in color orange was positioned at zero which means an equal pathogenicity upon variations. It can be seen clearly from the figure that the distribution of logarithm ratios of majority of dinucleotides were consistent for variants within nucleosome core regions and at linker regions. Precisely, dinucleotides CpG, GpG, TpA, TpC, TpG and TpT within nucleosome core regions and at linker regions are more likely to become pathogenic type (with negative ratios) while dinucleotides ApC, ApG, ApT, CpA, CpT and GpT within nucleosome core regions and at linker regions tend to be converted into neutral type (with positive ratios). In addition, distribution of CpC and GpC located very close to the dotted line, which shows almost equal tendency to be changed into either pathogenic or neutral types upon variations. In other words, they showed equal pathogenicity upon variations.

Dinucleotides showing same tendency upon variations might differ in the degree of pathogenicity based on the location of variations. For instance, logarithm ratio of dinucleotide TpG within nucleosome core regions bottomed almost at value -1, which means the probability of becoming pathogenic type is two times more than that of neutral type. Although TpG at linker regions also expressed a same trend, the tendency to be converted into pathogenic type is comparably lower than variations within nucleosome core regions. Likewise, dinucleotides CpG, GpG, TpC and TpT at linker regions all showed higher degree of pathogenicity than those within nucleosome core regions. Unlike those dinucleotides having higher degree of pathogenicity, CpA showed the least pathogenicity. Alternatively, CpA is most likely to become neutral type, and this tendency was observed much higher for variations at linker regions than within nucleosome core regions.

Noticeably, logarithm ratios of dinucleotides ApA and GpA were distributed at different sides of the dotted line (0), and they were the only two dinucleotides showing different variation tendency based on the location of variations. In particular, ApA is more likely to be converted into neutral type when variations occur within nucleosome core regions, while it shows a higher tendency to become pathogenic type for variations occurring at linker regions. Similarly, GpA within nucleosome core regions tends to become pathogenic type whereas it shows a higher trend to be changed into neutral type for variations at linker regions. As a consequence, the location where variations have occurred had impact on dinucleotide variability and degree of pathogenicity, and thus should be taken into account when predicting pathogenicity of novel variants.

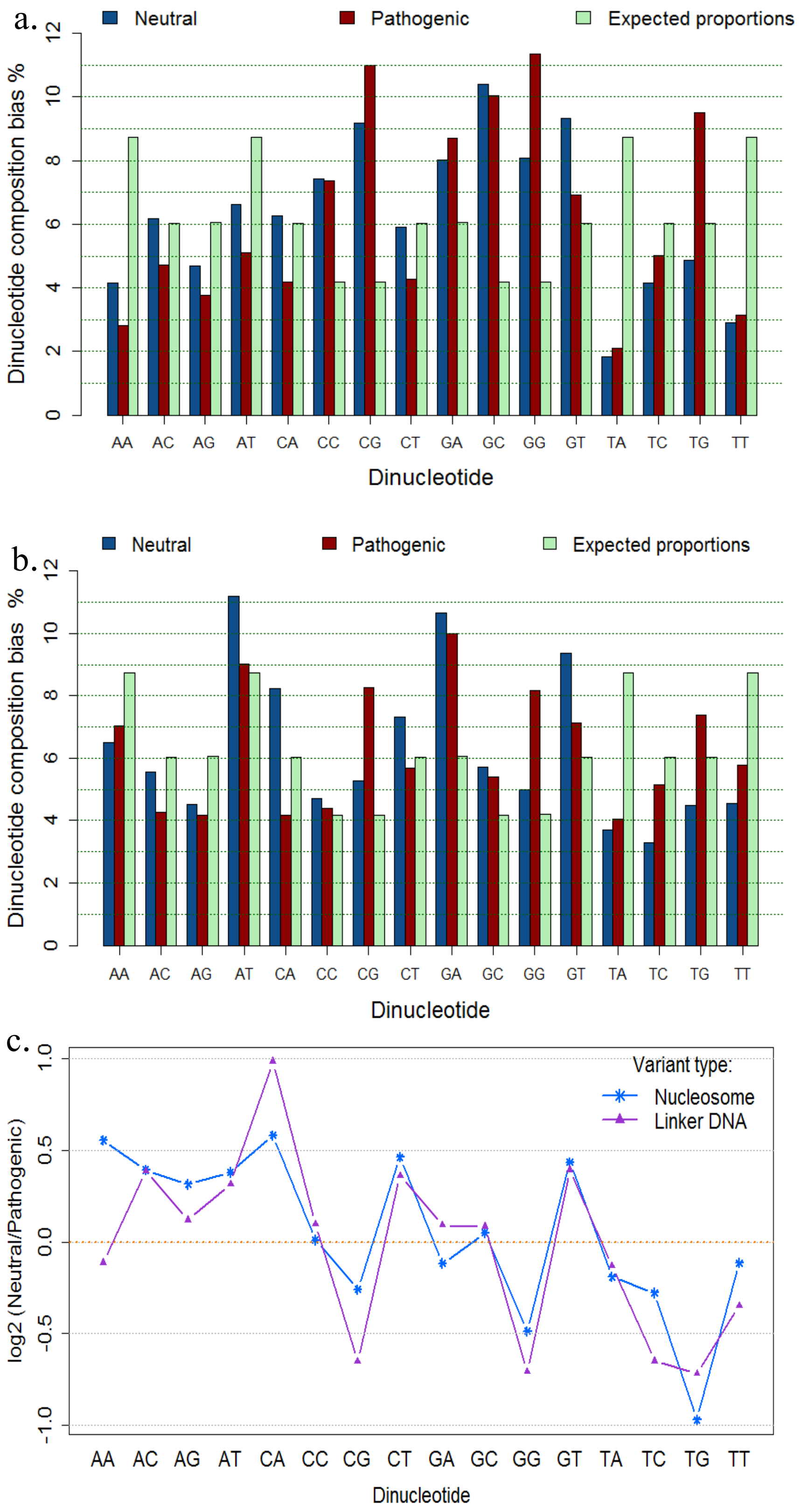

Figure 5.11 Distribution of dinucleotide compositions of pathogenic and neutral variations. Dinucleotide proportions in human genome are included for reference (green bars). (a) dinucleotides within nucleosome core regions. (b) dinucleotides at linker regions. (c) variation tendency and degree of pathogenicity of each dinucleotide within nucleosome core regions (blue) and at linker regions (purple).

Table 5.12 Dinucleotide compositions. Expected composition refers to dinucleotide proportions in human genome. Cells highlighted with color yellow indicate higher dinucleotide composition for pathogenic variations than neutral type. "N." and "P." represent "Neutral" and "Pathogenic", respectively.

Dinucleotides	Composition N. (Nuc.) %	Composition P. (Nuc.) %	Composition N. (Linker) %	Composition P. (Linker) %	Composition expected %
ApA	4.15	2.82	6.5	7.03	8.73
ApC	6.19	4.72	5.55	4.26	6.04
ApG	4.69	3.78	4.51	4.16	6.05
ApT	6.62	5.09	11.2	9.01	8.73
CpA	6.27	4.19	8.22	4.16	6.04
CpC	7.43	7.38	4.71	4.4	4.18
CpG	9.19	10.99	5.26	8.26	4.18
CpT	5.90	4.28	7.31	5.69	6.04
GpA	8.02	8.69	10.65	10	6.05
GpC	10.40	10.04	5.72	5.39	4.18
GpG	8.09	11.34	5.00	8.16	4.19
GpT	9.33	6.91	9.35	7.13	6.04
TpA	1.83	2.09	3.7	4.06	8.73
TpC	4.14	5.02	3.28	5.15	6.04
TpT	4.86	9.52	4.48	7.37	6.04
TpG	2.91	3.15	4.55	5.79	8.73

6. DISCUSSION

Estimation of nucleosome stability based on the type and location of variations was the main aspect of this research. Nucleosome regulates gene expression in various aspects, for instance, either by intrinsically encoding stable nucleosomes over non-functional sites or organizing unstable nucleosomes over functional sites, thereby either decreasing their accessibility to these sites or enhancing the accessibility of transcriptional factors towards these functional binding sites. Nucleosome expresses special preference to some particular DNA sequences, which indicates possible participation in establishing their positioning patterns. As a consequence, analysis of impact of DNA sequence variations on nucleosome stability may provide possible information for their correlation. In addition, study of (di) nucleotide variability and degree of pathogenicity, thereby possibly providing information for the prediction of pathogenicity of novel variants was another concern in this research.

Human single nucleotide variations (SNVs) of both neutral and pathogenic types were obtained from VariBench database. DNA sequences of specified length were downloaded with respective identifiers and variant positions from NCBI, and these sequences were subsequently subjected to NuPoP for predicting nucleosome positions.

6.1 Visualization of variant positions within nucleosome core regions and at linker regions

Variant positions (bp) within nucleosome core regions and at linker regions were calculated and compared. There were approximately 83.32% and 89.02% pathogenic and neutral variations within nucleosome core regions, respectively. In contrast, only 14.68% and 10.98% variants placed along linker DNA for each of the type. This is in agreement with the finding from other studies, showing that 75-90% of genomic DNA is wrapped in nucleosomes (Felsenfeld and Groudine, 2003; Segal *et al.*, 2006). In Figure 5.1, variants within nucleosome core regions showed a considerably higher distribution than variants occurring at linker regions, this is possibly because the majority of genomic DNA is packed in nucleosomes, thereby providing a higher variation incidence for variants within nucleosomes, and considering the length of nucleosomal DNA (147 bp), each position gains a higher observation. No apparent difference in the distribution of variation positions between pathogenic and neutral variants within nucleosome core regions and at linker regions were observed.

6.2 Nucleosome stability aspects

6.2.1 Regarding nucleosome binding affinity

Overall nucleosome binding affinity scores of both neutral and pathogenic variants were systematically calculated and statistically plotted. As demonstrated in Figure 5.2, distribution of binding affinity scores was found higher for pathogenic variants than neutral type. Notably, 1st quartile, median and 3rd quartile of affinity scores were larger for pathogenic variants irrespective of the location of variations (within nucleosome core regions and at linker regions). An interesting phenomenon found here is that for variations within nucleosome core regions, the inter-quartile range was found wider for neutral type (Fig.5.2.a), while it was found wider for pathogenic variants occurring at linker regions (Fig.5.2.b and Fig.5.2.c). Apparently, variants within nucleosome core regions displayed a considerable higher binding affinity than variants at linker regions irrespective of the type of variations.

Mann-Whitney U test was performed to examine the actual difference in affinity scores between pathogenic and neutral variations. Null hypothesis was assumed as identical affinity mean values between these two types of variations, while alternative hypothesis was considered as distinct affinity mean values. Results obtained from all MW tests showed p-values of less than 0.05 (p-value < $2.2*10^{-16}$, 5% significance level), which indicates a significant difference in affinity mean values between pathogenic and neutral types. Research results from other studies might provide possible information for the association between DNA sequence variations and nucleosome organization in the human genome. Genome are thought to encode an intrinsic nucleosome organization (Segal *et al.,* 2006), variations in genomic DNA can disrupt nucleosome-positioning signals encoded in DNA and alter the binding sites of transcription factors in the linkers (Tolstorukov *et al.,* 2011). In particular, the modifications on histone by adding or removing various chemical elements affect the binding affinity between histones and DNA, and thus loosening and tightening the condensed DNA wrapped around histones, which further leads to gene repression or the increase of gene expression (Wang *et al.,* 2007). Moreover, nucleosomes show higher affinity for some particular DNA sequences, reflecting the sharp bending ability of DNA sequences as is required by the nucleosome structure (Segal *et al*., 2006). According to these views, the bendability of DNA sequences and modifications on histone proteins affect nucleosome binding affinities. Hence, pathogenic variants might favor nucleosome binding affinity by causing a sharper bendability for DNA sequences, and some mechanisms which affect the histone modifications causing the tightness of the DNA sequence wrapped around histones.

These features have also been observed in the study of affinity scores for individual nucleotides of both types (Fig.5.3). Non-overlapping notches in boxes indicate a strong evidence of significant difference in affinity scores between pathogenic and neural variants in each nucleotide group. Notably, for variants occurring inside nucleosomes (Fig.5.3.a) median difference between pathogenic and neutral variants of each nucleotide group were not considerably as large as the difference of variants at linker regions. In particular, in Fig.5.3.b and Fig.5.3.c, binding affinities of neutral variants in each nucleotide group were distributed noticeably lower than those of pathogenic type. Nucleotide C was found to have highest median difference than other nucleotide groups. Moreover, nucleotides C and G expressed higher binding affinity than nucleotides A and T, which possibly means C and G might favor more nucleosome stability when compared to A and T. Researches have shown that poly (dA-dT) particularly disfavor nucleosome formation (Anderson and Widom, 2001; Prytkova et al., 2011).

For a deeper insight, analysis went further to examine affinity scores based on nucleotide classes (Figure 9.1), substitution types (Figure 9.2) and nucleotide substitutions (Figure 9.3), respectively. Analysis of these three aspects revealed all same features as previous studies, that pathogenic variants showed higher binding affinity irrespective of the location of variations (within nucleosome core regions and at linker regions); binding affinity was stronger for variants occurring inside nucleosomes irrespective of the type of variations. In Fig.9.1.a, variants of both pathogenic and neutral types displayed similar affinity distribution pattern between purine and pyrimidine groups. Similar features can also be found in Fig.9.1.b and Fig.9.1.c. This trend was also observed when comparing binding affinity between pathogenic and neutral variants grouped by substitution types (transitions and transversions) (Figure 9.2). Although significant difference was observed in all substitution patterns, several substitution patterns were found with a relatively larger difference in affinity median values, e.g. substitution C→A.

6.2.2 Regarding nucleosome occupancy level

Nucleosome occupancy level between pathogenic and neutral variations were studied in a same way as nucleosome binding affinity. Figure 5.4 presents the overall comparison of occupancy scores between pathogenic and neutral variations. The same behavior as the analysis of binding affinity was found, that higher occupancy level was shown for pathogenic variants irrespective of the location of variations; occupancy level was stronger for variations occurring within nucleosome core regions irrespective of the type of variations. In addition to the observation obtained from Figure 5.4, the results obtained from MW tests provided some more information which supports the above mentioned features. Segal *et al.* (2006) have found low nucleosome occupancy at functional binding sites and transcription starting sites, they explain this phenomenon in a way that genome encode unstable nucleosomes over these sites to increase the accessibility of transcriptional machinery to these sites.

Occupancy level between pathogenic and neutral variations were compared for individual nucleotides for further exploration, as shown in Figure 5.5. A similar trend as the study in Figure 5.4 was found for most of nucleotides that variations within nucleosome core regions have scored higher occupancy level than those at linker regions, and pathogenic variations tend to have higher occupancy level than the neutral type. However, few exceptions were noticed by performing MW tests to examine the actual difference. In particular, p-values for nucleotide C (Fig.5.5.a) and C, T (Fig.5.5.c) all scored greater than 0.05 (5% significance level) which indicates identical occupancy mean values between these two types of variations (details in Table 5.6).

In addition, occupancy level were also analyzed based on nucleotide classes (Figure 9.4), substitution types (Figure 9.5) and substitution patterns (Figure 9.6), respectively. The distribution of occupancy scores between pathogenic and neutral variations of these groups showed a similar trend as previous studies, except for few cases that, no significant occupancy difference was found. For instance, in Fig.9.3.c, MW test scored a p-value of 0.1329 for the pyrimidine group, which indicates identical occupancy level between pathogenic and neutral variants at downstream linker regions. However, this observation might be inherited from previous study of individual nucleotides that nucleotide C and T at downstream linker regions both have scored p-values greater than 0.05, which result in a p-value greater than 0.05 for the corresponding pyrimidine group. Significant occupancy difference between pathogenic and neutral variations was found in each transition and transversion group. Not all substitution patterns were observed with significant occupancy difference between pathogenic and neutral variations, such as pattern C→T, G→A, T→G, etc.

6.3 Nucleotide variability aspects

6.3.1 Variability of individual nucleotides

The study of the variability of individual nucleotides revealed that nucleotides C and G are more likely to be mutated into either pathogenic or neutral variants although their proportions in human genome are relatively less (Figure 5.8). In contrast, nucleotides A and T were observed with a considerably lower variation frequency in spite of their high proportions in human genome.

Figure 5.9 demonstrates the substitution rage of each substitution pattern. Obviously, substitution C→T and G→A predominate among all other patterns, this observation is similar to Hershberg and Petrov's statement that the most common variation is always G:C to A:T transition (Hershberg and Petrov, 2010). In addition to this view, Zhao and Boerwinkle (2002) also have found a considerably high substitution proportions for C/T (32.81%) and A/G (32.77%) by studying the substitution rates of a large amount of SNVs. This phenomenon might be explained by findings that around 60% ~90% of all cytosines in CpGs are methylated to thymine in mammals (Ehrlich *et al.,* 1982; Tucker, 2001). In addition to the variability, the degree of conversion into pathogenic and neutral types of variations was found different. In particular, variation C→T and its complementary pattern G→A displayed a stronger tendency to become pathogenic variations whereas C→A, C→G and G→C are more likely to be converted into neutral type, and their substitution rates are considerably lower when compared to pattern C→T and G→A. Although nucleotides A and T have been observed with low variability, substitution pattern A→G and T→C displayed fairly higher variation frequency than their other substitution patterns such as A→T and T→A. Notably, substitution of C→T and T→C showed a highest pathogenicity among all others whereas A→G and C→G displayed the least.

Table 5.10 illustrates an observation that transition variations are generated at a fairly higher frequency than transversions. This observation is similar to other literature findings that although there is twice the number of possible transversion variations than transition variations, transition variations appear more often in genome (Ebersberger *et al.,* 2002). The high frequency for transition variations might be because of the property of nucleotide chemistry that similar structure favors their substitution variations as mentioned by Keller *et al.* (2007) in the study of transition and transversion bias. Yang and Nielsen (2000) mentioned that transitions are more likely to be synonymous variations at third positions than transversions, although three out of four types of transition variations observed here showed a higher tendency to cause pathogenicity (Table 5.10). However, the frame of variations in codons has not been taken into account in this study.

6.3.2 Variability of different dinucleotides

It has been known that nucleotide variations are not random but highly related to neighboring-nucleotide effects (Zhao and Boerwinkle, 2002). Variation frequency of CpG dinucleotide content was compared to non-CpG dinucleotides as well as their expected proportions in human genome which was calculated based on single nucleotide composition proportions (Figure 5.11). For variations within nucleosome core regions (Fig.5.11.a), dinucleotides CpC, CpG and complementary pattern GpG, GpC displayed a significant higher variation frequency in comparison with their expected proportions in human genome. On the other hand, some dinucleotides were found less frequent to be mutated although their expected proportions in human genome are fairly high. Dinucleotides ApA, ApT and their complementary pattern TpT, TpA are typical dinucleotides of this kind where variations were found much less frequent than others.

The majority of features shown in Fig.5.11.a can be also found in the same study but for variations at linker regions, although there are little difference in the degree of variability. In particular, in Fig.5.11.b, GpA and CpG expressed highest variability, while TpA is least likely to be mutated when compared to their expected proportions in human genome. Thus, dinucleotide CpG is a common variation site. The frequency of CpG dinucleotides in human genome is 1% rather than the expected 4.41%. The deficiency of CpG could be because most CpGs in mammalian genome are methylated on the C residue, which undergoes spontaneous deamination to T (Bird, 1980; Yoder *et al*., 1997). Further, Li and chen (2011) have found a fixation preference for CpG to be regenerated from TpG and CpA than other dinucleotides.

Fig.5.11.c represents the degree of pathogenicity of each dinucleotide. Log ratio below zero shows higher pathogenicity whereas above zero means a tendency to become neutral type. Several dinucleotides displayed higher variability in degree of pathogenicity. For instance, CpG, GpG and TpG expressed highest variability to become pathogenic type whereas CpA and GpT are more likely to be changed into neutral type irrespective of the location of variations (within nucleosome core regions and at linker regions). Consistently, literature findings have also suggested that the methylation-induced deamination of 5-methyl cytosine in CpG content may contribute significantly to the high incidence of human genetic diseases (Cooper and Youssoufian, 1988). Notably, the log ratio of dinucleotide CpA at linker regions peaked at 1 which indicates a twice higher tendency to become neutral variations rather than pathogenic type. In other words, CpA at linker regions showed the least pathogenicity. Analogously, TpG at nucleosome core regions bottomed at ratio -1 which reveals almost twice the preference in changing into pathogenic variations rather than neutral type, hence shows the highest pathogenicity. Several dinucleotides showed almost equal pathogenicity. They were equally changing into either pathogenic or neutral type upon variations within nucleosome core regions and at linker regions. Dinucleotides which fall into this category were CpC and GpC. Further, ApA and GpA were the only two dinucleotides showing different variation tendency based on the location of variations. In other words, their variation tendency are location-dependent.

6.4 Future perspectives

This study was initiated to investigative whether neutral and pathogenic variants have different impact on nucleosome stability in terms of nucleosome binding affinity and occupancy levels; analysis of the pathogenicity of variants was another concern in this research. Positive results have been found from this study, hence this study will possibly help in understanding the effect of nucleotide changes to the formation of nucleosome, and benefit the research group in further bioinformatics related research towards determining roles of nucleosomes in regulating gene expressions. In addition, results from this study might provide information for distinguishing neutral and pathogenic variants, thereby contributing to the development of a pathogenicity prediction tool for predicting pathogenicity of novel variations at DNA level.

7. CONCLUSION

The ultimate goal of this study was to find whether neutral and pathogenic SNVs have different impact on nucleosome stability in terms of nucleosome binding affinities and occupancy levels; analyses of variant localization, variability and degree of pathogenicity were also of interest in this research.

There was no obvious positioning difference between neutral and pathogenic variations along the DNA. Pathogenic SNVs scored higher nucleosome binding affinity whereas neutral SNVs scored lower binding affinity. In addition, higher binding affinity was observed for variations within nucleosome core regions than variations at linker regions. Same features were found when expanding the analysis to levels of individual nucleotides, nucleotide classes, substitution patterns and substitution types.

Similar features were observed in the study of nucleosome occupancy level, that variations within nucleosome core regions appeared to have higher occupancy scores than linker ones irrespective of the type of variations; pathogenic variations showed higher occupancy level than neutral type irrespective of the location of variations (within nucleosome core regions and at linker regions). However, few exceptions were found that substitution C→T, G→A and T→G have showed identical occupancy level between pathogenic and neutral types of variants within nucleosome core regions and at linker regions.

The location of variations were found to have impact on the variability and degree of pathogenicity of nucleotides and dinucleotides. Nucleotides C and G were more abundant in variations when compared to their proportions in human genome. Specifically, substitution C→T and its complementary pattern G→A showed the highest tendency to be mutated, and they are more likely to be changed into pathogenic type. Transition variations are considerably more frequent and tend to cause pathogenicity. CpG is one of the dinucleotides that showed highest variability among all others whereas TpA and TpT expressed the least. Dinucleotides CpG, GpG and TpG all displayed high tendency to cause pathogenicity whereas CpA and GpT behaved contrarily in their tendency to cause pathogenicity. CpC and GpA showed equal variation tendency irrespective of the location of variations. The variation tendency of ApA and GpA to be converted into pathogenic and neutral types is location-dependent.

8. REFERENCES

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Wlater P. (2002). *Molecular biology of the cell* (4th ed.). New York, NY: Garland Science.

Anderson JD and Widom J. (2001). Poly (dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. *Mol Cell Biol*, **21**: 3830-3839.

Anselmi C, De Santis P, Paparcone R, Savino M, Scipioni A. (2002). From the sequence to the superstructural properties of DNAs. *Biophys Chem*, **95**: 23-47.

Appanah R, Dickerson DR, Goyal P, Groudine M, Lorincz MC. (2007). An unmethylated 3' promoter-proximal region is required for efficient transcription initiation. *PLoS Genet*, **3**: e27.

Bagdonavicius V, Kruopis J, Nikulin MS. (2011*). Non-parametric tests for complete data*. London & Hoboken: Iste & Wiley.

Barlow DP. (1993). Methylation and imprinting: from host defense to gene regulation? *Science*, **260**: 309-310.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. (2008). Natural selection has driven population differentiation in modern humans. *Nature Genet*, **40**: 340-345.

Bird AP. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res*, **8**: 1499-1504

Burch CL and Chao L. (1999). Evolution by small steps and rugged landscapes in the RNA virus phi6. *Genetics*, **151**: 921-927.

Burgess RJ and Zhang ZG. (2013). Histone chaperones in nucleosome assembly and human disease. *Nat Struct Mol Biol*, **20**: 14-22.

Chen CY, Liou J, Forman LW, Faller DV. (1998). Correlation of genetic instability and apoptosis in the presence of oncogenic Ki-Ras. *Cell Death Differ*, **5**: 984-995.

Chen RZ, Pettersson U, Beard C, Jackson-Grusby L, Jaenisch R. (1998). DNA hypomethylation leads to elevated mutation rates. *Nature*, **395**: 89-93.

Chen ZX and Riggs AD. (2005). Maintenance and regulation of DNA methylation patterns in mammals. *Biochem Cell Biol*, **83**: 438-448

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**: 1422-1423.

Collins DW and Jukes TH. (1994). Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics*, **20**: 386-396.

Cooper DN and Youssoufian H. (1988). The CpG dinucleotide and human genetic disease. *Hum Genet*, **78**: 151-155.

Corder GW and Foreman DI. (2009). *Nonparametric statistics for non-statisticians: a step-by-step approach*. Hoboken, NJ: Wiley.

Darwin C. (1872). *On the origin of species* (6th ed.). London: John Murray.

Ebersberger I, Metzler D, Schwarz C, Pääbo S. (2002). Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet*, **70**: 1490-1497.

Ehrlich M, Gama Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA, Gehrke C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Res*, **10**: 2709-2721.

Elembis. (2007). *Diagram of mutation and selection in evolution.* [Graph]. Retrieved from http://643px-Explanation of Evolution v2.1.PNG

Fay MP and Proschan MA. (2010). Wilcoxon–Mann–Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, **4**: 1-39.

Felsenfeld G and Groudine M. (2003). Controlling the double helix. *Nature,* **421**: 448-453.

Ferguson-Smith AC, Sasaki H, Cattanach BM, Surani MA. (1993). Parental-origin-specific epigenetic modifications of the mouse H19 gene. *Nature*, **362**: 751-755.

Fisher RA. (1930). *The genetical theory of natural selection*. Oxford, MS: Clarendon Press.

Futuyma DJ. (2009). *Evolution* (2nd ed.). Sunderland, MA: Sinauer Associates.

Generalic. (2013). *Structure of nucleotide*. [Graph]. Retrieved from
http://glossary.periodni.com/glossary.php?en=nucleotide

Gibbons DJ and Chakraborti S. (2003). *Nonparametric statistical inference* (4th ed.). Boca Raton, FL: CRC Press.

Groth A, Rocha W, Verreault A, Almouzni G. (2007). Chromatin challenges during DNA replication and repair. *Cell*, **128**: 721-733.

Gupta S, Dennis J, Thurman RE, Kingston R, Stamatoyannopoulos JA, Noble WS. (2008). Predicting human nucleosome occupancy from primary sequence. *PLoS Comput Biol*, **4**: e1000134.

Hall BK and Hallgrímsson B. (2008). *Strickberger's evolution* (4th ed.). Sudbury, MA: Jones and Bartlett Publishers.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**: 99-104.

Hartl DL. (1981). A Primer of population genetics. *Am J Med Genet*, **17**: 869.

Hershberg R and Petrov DA. (2010). Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet*, **6**: e1001115.

Hettmansperger TP and McKean JW. (2010). *Robust nonparametric statistical methods* (2nd ed.). Boca Raton, FL: *CRC*.

Hughes AL. (2007). Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity*, **99**: 364-373.

Huxley JS. (2010). *Evolution: the modern synthesis*. Cambridge, MA: *The MIT Press*.

Ioshikhes IP, Albert I, Zanton SJ, Pugh BF. (2006). Nucleosome positions predicted through comparative genomics. *Nature Genet*, **38**: 1210-1215.

Jones PA and Laird PW. (1999). Cancer epigenetics comes of age. *Nat. Genet*, **21**: 163-167.

Jones PA and Takai D. (2001). The role of DNA methylation in mammalian epigenetics. *Science*, **293**: 1068-1070.

Kangaspeska S, Stride B, Metivier R, Polycarpou-Schwarz M, Ibberson D, Carmouche RP, Benes V, Gannon F, Reid G. (2008). Transient cyclical methylation of promoter DNA. *Nature*, **452**: 112-115.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, Segal E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **45**: 362-366.

Keller I, Bensasson D, Nichols RA. (2007). Transition-transversion bias is not universal: a counter example from Grasshopper pseudogenes. *PLoS Genet*, **3**: e22.

Kimura M. (1968). Evolutionary rate at the molecular level. *Nature*, **217**: 624-626.

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

King JL and Jukes TH. (1969). Non-Darwinian evolution. *Science*, **164**: 788-797.

Korber P, Luckenbach T, Blaschke D, Horz W. (2004). Evidence for histone eviction in trans upon induction of the yeast PHO5 promoter. *Mol Cell Biol*, **24**: 10965-10974.

Kornberg RD and Lorch Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, **98**: 285-294.

Kornberg RD. (1974). Chromatin structure: a repeating unit of histones and DNA. *Science*, **184**: 868-871.

Laird PW and Jaenisch R. (1996). The role of DNA methylation in cancer genetic and epigenetics. *Annu Rev Genet*, **30**: 441-464.

Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C. (2007). A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genet*, **39**: 1235–1244.

Li MK and Chen SS. (2011). The tendency to recreate ancestral CG dinucleotides in the human genome. *BMC Evol Biol*, **11**: 3.

Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, **389**: 251-260.

Mann HB; Whitney DR. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat*, **18**: 50-60.

Masel J. (2011). Genetic drift. *Curr Biol*, **21**: 837-838.

Maynard-Smith J and Szathmáry E. (1997). *The major transitions in evolution*. Oxford, MS: Oxford University Press.

Maynard-Smith J. (1989). *Evolutionary genetics*. Oxford, MS: Oxford University Press.

Mayr E. (2002). *What evolution is*. London: Weidenfeld & Nicolson.

Miele V, Vaillant C, d'Aubenton-Carafa Y, Thermes C, Grange T. (2008). DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res*, **36**: 3746-3756.

Nair PS and Vihinen M. (2013). VariBench: a benchmark database for variations. *Hum Mutat*, **34**: 42-49.

Nei M. (2005). Selectionism and neutralism in molecular evolution. *Mol Biol Evol*, **22**: 2318-2342.

Ohta T and Gillespie JH. (1996). Development of neutral and nearly neutral theories. *Theor Popul Biol*, **49**: 128-142.

Ohta T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, **246**: 96-98.

Ohta T. (1992). The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst*, **23**: 263-286.

Ohta T. (2002). Near-neutrality in evolution of genes and gene regulation. *PNAS*, **99**: 16134-16137.

Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M. (2012). PON-P: integrated predictor for pathogenicity of missense variants. *Hum Mutat*, **33**: 1166-1174.

Orr HA. (2009). Fitness and its role in evolutionary genetics. *Nat Rev Genet*, **10**: 531-539.

Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z. (2007). Nucleosome positioning signals in genomic DNA. *Genome Res*, **17**: 1170-1177.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. (2004). UCSF Chimera-a visualization system for exploratory research and analysis. *J Comput Chem*, **25**:1605-1612.

Petulda. (2012). *Definition of transitions and transversions.* [Graph]. Retrieved from http://en.wikipedia.org/wiki/File:Transitions-transversions-v3.png

Prytkova T, Zhu X, Widom J, Schatz GC. (2011). Modeling DNA-bending in the nucleosome: role of AA periodicity. *J Phys Chem B*, **115**: 8638-8644.

Richmond TJ and Davey CA. (2003). The structure of DNA in the nucleosome core. *Nature*, **423**: 145-150.

R Core Team. (2013). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sanna CR, Li WH, Zhang. (2008). Overlapping genes in the human and mouse genomes, *BMC Genomics*, **9**: 169.

Satchwell SC, Drew HR, Travers AA (1986). Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol*, **191**: 659-675.

Saxonov S, Berg P, Brutlag DL. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *PNAS*, **103**: 1412-1417.

Schulze SR and Wallrath LL. (2007). Gene regulation by chromatin structure: paradigms established in *Drosophila melanogaster*. *Annu Rev Entomol*, **52**: 171-192.

Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang JP, Widom J. (2006). A genomic code for nucleosome positioning. *Nature*, **442**: 772-778.

Sekinger EA, Moqtaderi Z, Struhl K. (2005). Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol Cell*, **18**: 735-748.

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. (2009). The human gene variation database: 2008 update. *Genome med*, **1**: 13.

Sullivan AD, Wigginton J, Kirschner D. (2001). The coreceptor mutation CCR5Δ32 influences the dynamics of HIV epidemics and is selected for by HIV. *PNAS*, **95**: 10214-10219.

Suzuki MM, Kerr ARW, De Sousa D, Bird A. (2007). CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res*, **17**: 625-631.

Tanaka Y and Nakai K. (2009). An assessment of prediction algorithms for nucleosome positioning. *Genome Inform*, **23**: 169-178.

Teif VB and Rippe K. (2009). Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities. *Nucleic Acids Res*, **37**: 5641-5655.

Thusberg J and Vihinen M. (2009). Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum Mutat*, **30**: 703-714.

Tolstorukov MY, Choudhary V, Olson WK., Zhurkin VB, Park PJ. (2008). nuScore: a web-interface for nucleosome positioning predictions. *Bioinformatics*, **24**: 1456-1458.

Tolstorukov MY, Colasanti AW, McCandlish DM, Olson WK, Zhurkin VB. (2007). A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J Mol Biol*, **371**: 725-738.

Tolstorukov MY, Volfovsky N, Stephens RM, Park P J. (2011). Impact of chromatin structure on sequence variability in the human genome. *Nat Struct Mol Biol*, **18**: 510-515.

Tucker KL. (2001). Methylated cytosine and the brain: a new base for neuroscience. *Neuron*, **30**: 649-652.

Varela MA and Amos W. (2010). Heterogeneous distribution of SNPs in the human genome: microsatellites as predictors of nucleotide diversity and divergence. *Genomics,* **95**: 151-159.

Vignali M, Hassan AH, Neely KE, Workman JL. (2000). ATP-dependent chromatin-remodeling complexes. *Mol Cell Biol*, **20**: 1899-1910.

Villard J. (2004). Transcription regulation and human diseases. *Swiss Med Wkly*, **134**: 571-579.

Wang GG, Allis CD, Chi P. (2007). Chromatin remodeling and cancer, part II: ATP-dependent chromatin remodeling. *Trends Mol Med*, **13**: 363-372

Wang JP, Fondufe-Mittendorf Y, Xi L, Tsai GF, Segal E, Widom J. (2008). Preferentially quantized linker DNA lengths in *Saccharomyces cerevisiae*. *PLoS Comput Biol*, **4**: e1000175.

Wasserman L (2006). *All of nonparametric statistics*. New York, NY: Springer.

Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, Schübeler D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet*, **39**: 457-466.

Whitehouse I, Rando OJ, Delrow J, Tsukiyama T. (2007). Chromatin remodelling at promoters suppresses antisense transcription. *Nature*, **450**: 1031-1035.

Widom J. (2001). Role of DNA sequence in nucleosome stability and dynamics. *Q Rev Biophys*, **34**: 269-324.

Wilson GG, Murray NE. (1991). Restriction and modification systems. *Annu Rev Genet*, **25**: 585-627.

Xi L, Fondufe-Mittendorf Y, Xia L, Flatow J, Widom J, Wang JP. (2010). Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics*, **11**: 346.

Yang ZH and Nielsen R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*, **17**: 32-43.

Yen PH, Patel P, Chinault AC, Mohandas T, Shapiro LJ. (1984). Differential methylation of hypoxanthine phosphoribosyltransferase genes on active and inactive human X chromosomes. *PNAS*, **81**: 1759-1763.

Yoder JA, Walsh CP, Bestor TH, Walsh CP, Bestor TH, Bestor TH. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet*, **13**: 335-340.

Yuan GC and Liu JS. Genomic sequence is highly predictive of local nucleosome depletion. (2008). *PLoS Comput Biol*, **4**: e13.

Zhao ZM and Boerwinkle E. (2002). Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res*, **12**: 1679-1686.

Zhou YB, Gerchman SE, Ramakrishnan V, Travers A and Muyldermans S. (1998). Position and orientation of the globular domain of linker histone H5 on the nucleosome. *Nature*, **395**: 402-405.

9. APPENDIX

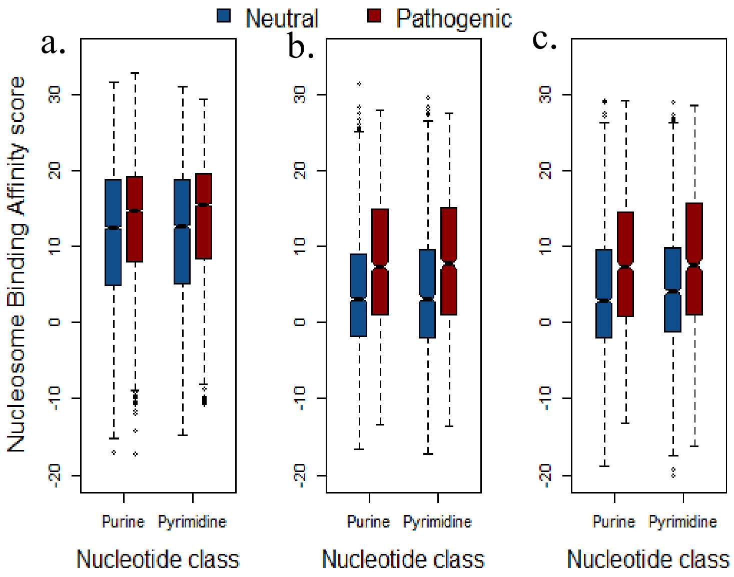

Figure 9.1 Distribution of nucleosome binding affinity scores between neutral and pathogenic variations grouped by nucleotide classes. (a) Variations within nucleosome core regions (b) Variations at upstream linker regions. (c) Variations at downstream linker regions.

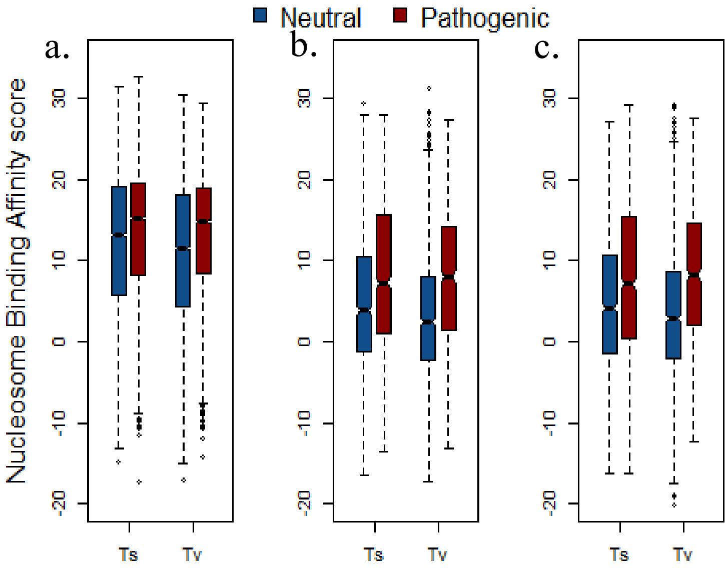

Figure 9.2 Distribution of nucleosome binding affinity scores between neutral and pathogenic variations grouped by substitution types. (a) Variations within nucleosome core regions (b) Variations at upstream linker regions. (c) Variations at downstream linker regions. "Ts" and "Tv" represent "transitions" and "transversions", respectively.

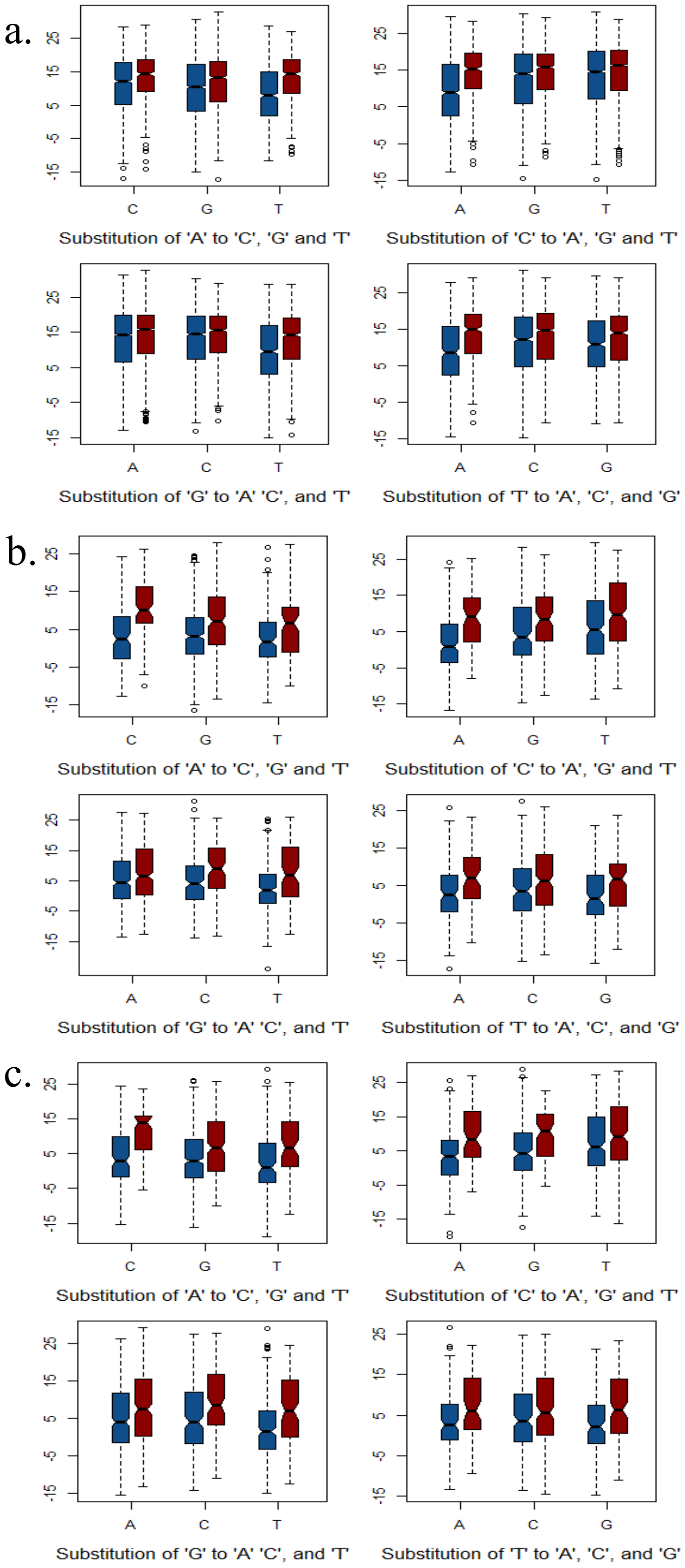

Figure 9.3 Distribution of nucleosome binding affinity scores between neutral and pathogenic variations grouped by nucleotide substitutions. (a) Variations within nucleosome core regions (b). Variations at upstream linker regions. (c) Variations at downstream linker regions.

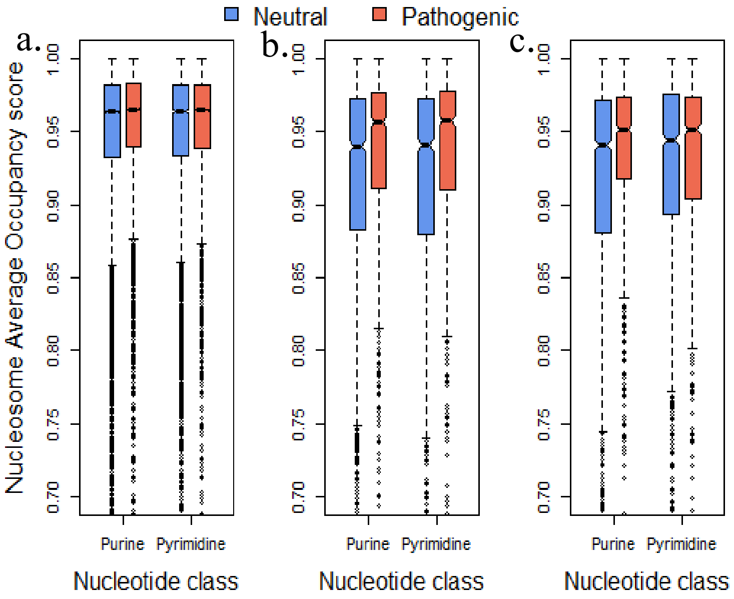

Figure 9.4 Distribution of nucleosome occupancy scores between neutral and pathogenic variations grouped by nucleotide classes. (a) Variations within nucleosome core regions (b) Variations at upstream linker regions. (c) Variations at downstream linker regions.

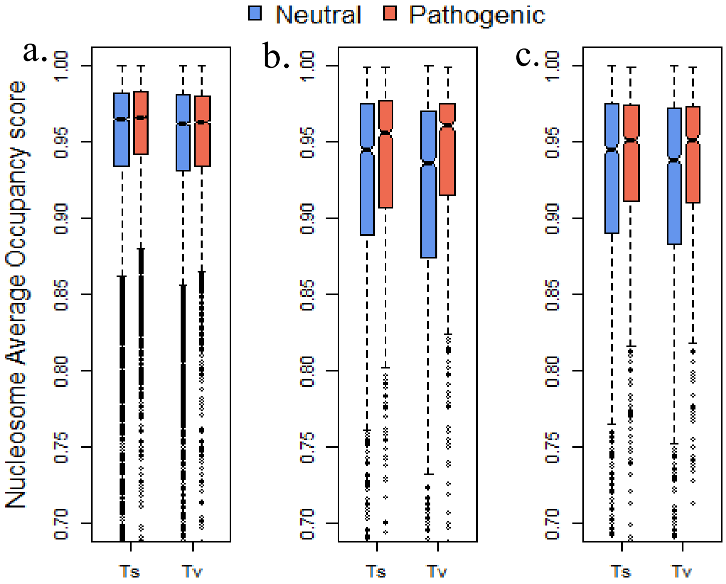

Figure 9.5 Distribution of nucleosome occupancy scores between neutral and pathogenic variations grouped by substitution types. (a) Variations within nucleosome core regions (b) Variations at upstream linker regions. (c) Variations at downstream linker regions. “Ts” and “Tv” represent “transitions” and “transversions”, respectively.

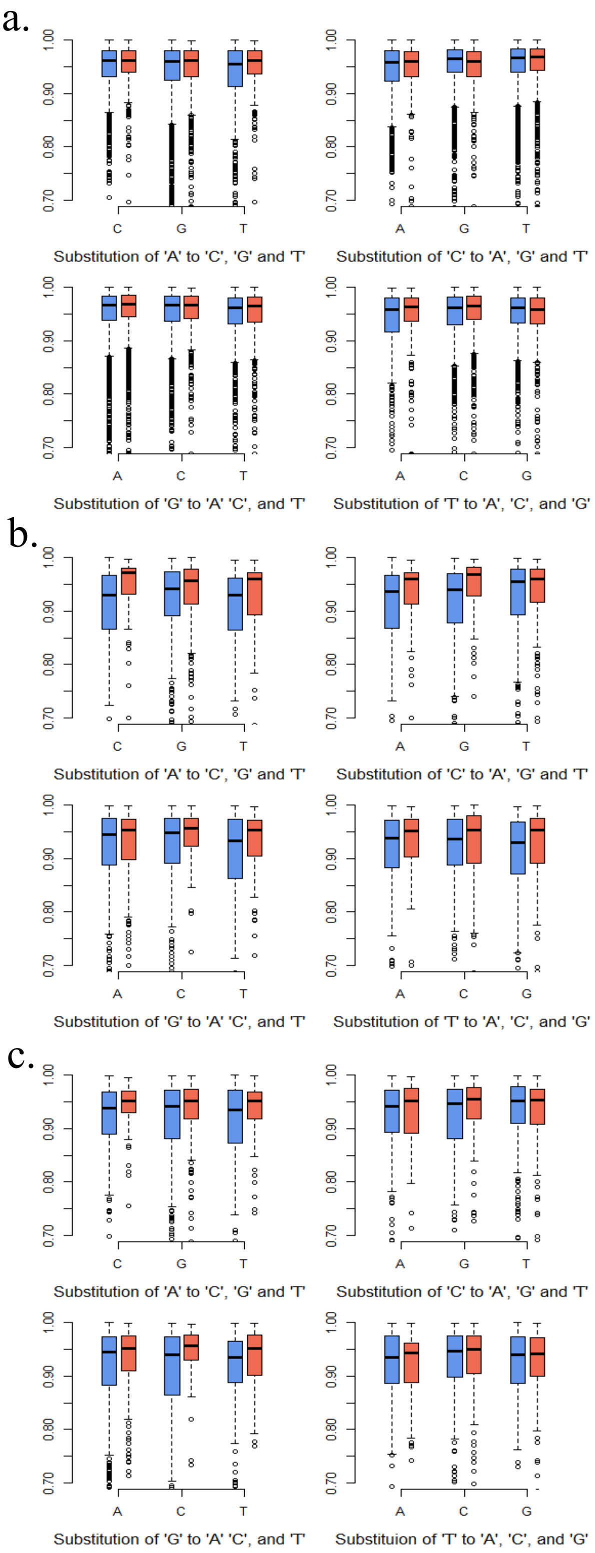

Figure 9.6 Distribution of nucleosome occupancy scores between neutral and pathogenic variations grouped by nucleotide substitutions. (a) Variations within nucleosome core regions (b) Variations at upstream linker regions. (c) Variations at downstream linker regions.

Table 9.7 Amino acid standard genetic codon table. Each amino acid is listed with corresponding coding codons.

Amino Acid	Codons	Amino Acid	Codons
A	GCT, GCC, GCA, GCG	**M**	ATG
C	TGT, TGC	**N**	AAT, AAC
D	GAT, GAC	**P**	CCT, CCC, CCA, CCG
E	GAA, GAG	**Q**	CAA, CAG
F	TTT, TTC	**R**	CGT, CGC, CGA, CGG, AGA, AGG
G	GGT, GGC, GGA, GGG	**S**	TCT, TCC, TCA, TCG, AGT, AGC
H	CAT, CAC	**T**	ACT, ACC, ACA, ACG
I	ATT, ATC, ATA	**V**	GTT, GTC, GTA, GTG
K	AAA, AAG	**W**	TGG
L	TTA, TTG, CTT, CTC, CTA, CTG	**Y**	TAT, TAC

Table 9.8 Nucleosome binding affinity scores (1st quartile, median, 3rd quartile) of both pathogenic and neutral variations divided by individual nucleotides. "P." and "N." represent "Pathogenic and "Neutral", respectively.

Variants within Nuc. core regions	**1st quartile P.**	**1st quartile N.**	**Median P.**	**Median N.**	**3rd quartile P.**	**3rd quartile N.**
A	6.63	3.34	13.83	10.27	18.30	16.88
C	9.65	5.70	16.08	13.40	20.09	19.35
G	8.40	6.06	15.41	13.70	19.60	19.55
T	7.03	4.20	14.45	11.09	18.98	17.41
Variants at upstream linker regions	**1st quartile P.**	**1st quartile N.**	**Median P.**	**Median N.**	**3rd quartile P.**	**3rd quartile N.**
A	0.98	-2.03	7.82	2.77	13.63	7.83
C	2.24	-2.02	9.37	3.41	16.73	10.78
G	1.17	-1.33	7.16	3.60	15.59	9.99
T	0.45	-2.11	6.72	2.98	15.59	8.51
Variants at downstream linker regions	**1st quartile P.**	**1st quartile N.**	**Median P.**	**Median N.**	**3rd quartile P.**	**3rd quartile N.**
A	0.54	-2.21	7.36	2.50	14.20	8.98
C	2.67	-0.90	9.98	4.64	17.21	10.83
G	0.86	-1.92	7.35	3.31	15.40	10.13
T	0.42	-1.52	5.91	3.09	14.20	8.59

Table 9.9 Nucleosome binding affinity scores (1st quartile, median, 3rd quartile) of both pathogenic and neutral variations divided by nucleotide classes. "P." and "N." represent "Pathogenic" and "Neutral", respectively.

Variants within Nuc. core regions	**1st quartile P.**	**1st quartile N.**	**Median P.**	**Median N.**	**3rd quartile P.**	**3rd quartile N.**
Purine	7.90	4.92	14.76	12.47	19.14	18.66
Pyrimidine	8.42	5.13	15.41	12.57	19.57	18.83
Variants at upstream linker regions	**1st quartile P.**	**1st quartile N.**	**Median P.**	**Median N.**	**3rd quartile P.**	**3rd quartile N.**
Purine	1.10	-1.68	7.41	3.11	14.81	9.03
Pyrimidine	1.11	-2.08	7.73	3.17	15.12	9.63
Variants at downstream linker regions	**1st quartile P.**	**1st quartile N.**	**Median P.**	**Median N.**	**3rd quartile P.**	**3rd quartile N.**
Purine	0.77	-2.07	7.35	2.91	14.53	9.56
Pyrimidine	1.13	-1.15	7.62	4.07	15.67	9.85

Table 9.10 Nucleosome binding affinity scores (1st quartile, median, 3rd quartile) of both pathogenic and neutral variations classified by substitution types. "P." and "N." represent "Pathogenic" and "Neutral", respectively.

Variants within Nuc. core regions	**1st quartile P.**	**1st quartile N.**	**Median P.**	**Median N.**	**3rd quartile P.**	**3rd quartile N.**
Transitions	8.16	5.70	15.23	13.17	19.57	19.17
Transvertions	8.37	4.32	14.81	11.58	19.01	18.16
Variants at upstream linker regions	**1st quartile P.**	**1st quartile N.**	**Median P.**	**Median N.**	**3rd quartile P.**	**3rd quartile N.**
Transitions	0.98	-1.24	7.24	3.86	15.65	10.52
Transvertions	1.34	-2.33	8.01	2.39	14.19	8.12
Variants at downstream linker regions	**1st quartile P.**	**1st quartile N.**	**Median P.**	**Median N.**	**3rd quartile P.**	**3rd quartile N.**
Transitions	0.36	-1.36	7.16	4.16	15.49	10.65
Transvertions	2.06	-2.08	8.21	2.80	14.53	8.65

Table 9.11 Nucleosome occupancy scores (1st quartile, median, 3rd quartile) of pathogenic and neutral variations divided by individual nucleotides. “P.” and “N.” represent “Pathogenic and “Neutral”, respectively.

Variants within Nuc. core regions	**1st quartile P.**	**1st quartile N.**	**Median P.**	**Median N.**	**3rd quartile P.**	**3rd quartile N.**
A	0.932	0.924	0.961	0.959	0.980	0.980
C	0.940	0.936	0.965	0.965	0.982	0.983
G	0.943	0.937	0.967	0.966	0.984	0.961
T	0.937	0.928	0.964	0.961	0.982	0.981
Variants at upstream linker regions	**1st quartile P.**	**1st quartile N.**	**Median P.**	**Median N.**	**3rd quartile P.**	**3rd quartile N.**
A	0.915	0.882	0.964	0.938	0.977	0.971
C	0.917	0.880	0.963	0.944	0.978	0.974
G	0.909	0.882	0.954	0.943	0.976	0.975
T	0.894	0.880	0.953	0.935	0.978	0.972
Variants at downstream linker regions	**1st quartile P.**	**1st quartile N.**	**Median P.**	**Median N.**	**3rd quartile P.**	**3rd quartile N.**
A	0.919	0.880	0.952	0.940	0.972	0.971
C	0.906	0.894	0.952	0.946	0.975	0.976
G	0.913	0.882	0.952	0.941	0.976	0.972
T	0.904	0.891	0.949	0.942	0.973	0.974

Table 9.12 Mann–Whitney U test results for both pathogenic and neutral variations grouped by nucleotide classes. Cells shaded by yellow color indicate significant nucleosome occupancy difference (p-value <0.05) between pathogenic and neutral variants.

Nucleotide class	**Variants within Nuc. core regions P-value**	**Variants at upstream linker regions P-value**	**Variants at downstream linker regions P-value**
Purine	6.128e-09	4.455e-09	1.345e-07
Pyrimidine	0.002672	1.039e-07	0.1329

Table 9.13 Mann–Whitney U test results for both pathogenic and neutral variations grouped by substitution types. Cells shaded by yellow color indicate significant nucleosome occupancy difference (p-value <0.05) between pathogenic and neutral variations.

Substitution types	**Variants within Nuc. core regions P-value**	**Variants at upstream linker regions P-value**	**Variants at downstream linker regions P-value**
Transitions	6.653e-08	0.0004303	0.0193
Transversions	0.04073	3.997e-13	1.345e-07