Analysis of nucleotide repeats at variation sites

Master's Thesis
Master's Degree Programme in Bioinformatics
Institute of Biomedical Technology (IBT)
University of Tampere, Finland.
Abidemi Foluke Ogunbayo.
June 2012.

Acknowledgements

I give all thanks to God for keeping me alive and healthy throughout my master's degree programme. By his grace, I went, I saw and I conquered.

My special thanks go to my project supervisors Professor Mauno Vihinen and Ortutay Csaba, Acting Professor, PhD who took their precious time to supervise and review my thesis.

I specially appreciate all my teachers in MDP in bioinformatics, University of Tampere for their academic support and made me worthy in this successful professional career.

Million thanks to my parents, spouse and siblings for their moral and financial supports throughout my studies. To Temidayo Ogunbayo, you are the best daughter in the world. May God continue to be with you all.

Finally, it worth mentioning, that I really appreciate all my friends both in Finland and Nigeria for being there for me at all time. You are all appreciated.

Abidemi Foluke Ogunbayo.
Bioinformatics programme
Institute of Biomedical Technology (IBT)
University of Tampere,
Finland.

MASTER'S THESIS

Place:	UNIVERSITY OF TAMPERE Institute of Medical Technology Faculty of Medicine Tampere, Finland
Author:	Abidemi Foluke Ogunbayo
Title:	Analysis of nucleotide repeats at variation sites
Pages:	73 pages + 17 appendices pages
Supervisors:	Professor Mauno Vihinen, Ph. D. and Docent Csaba Ortutay, Ph. D.
Reviewers:	Professor Mauno Vihinen, Ph. D. and Docent Csaba Ortutay, Ph. D.
Time:	June, 2012

Abstract

Background and aims: Nucleotide repeats are sequences that are repeated two or more times in the genome. They are generally classified into two broad groups as tandem repeats or dispersed repeats. They can be found in both prokaryotic and eukaryotic organisms. Previous studies have shown the distribution of the repeats in the genome, likewise, some studies have focused on the development of repeat analysis tools. Some other studies have also pointed out the biological and medical importance of the nucleotide repeats, with much reference to the microsatellite repeats. This present study aimed at finding different patterns of repeats that are present at some pathogenic and neutral variation sites of the human genome. Also its aim is to report the abundance of these repeat patterns and investigate whether and where differences occur between and within the repeat patterns found in both the pathogenic and neutral dataset.

Methods: Datasets of neutral and pathogenic single nucleotide polymorphisms for human genome were downloaded from VariBench. The given genomic.accession.version for both datasets was then used for genomic sequence retrieval from the NCBI ftp site. The resulting sequences were preprocessed with python scripts, to obtain 21 bp each for the samples in each dataset. The 21 bp includes the variation site, located at the center of the sequence, for the analysis of the repeat. REPuter, a repeat analysis tool, was used to carry out the analysis on the sequences. The results of the repeat analysis were further preprocessed with python scripts. Microsoft Excel and R scripts were used for descriptive statistics in order to obtain the distribution of the different patterns of repeats and their nucleotide counts. Inferential statistics was done with ANOVA and Tukey's HSD test to show if significant differences occur between the repeat patterns in both datasets and where these differences occur.

Results: Forward, complemented, reversed and palindromic repeats were found present in both datasets. Descriptive statistics shows similarity in distribution of the patterns of repeats in both dataset. However ANOVA analysis showed that significant differences occur between the patterns of repeats in both dataset, this lead to further analysis with Tukey's HSD test, which confirms the exact pairs with significant difference.

Conclusions: Significant difference occurs in pairs of different pattern of repeat within and between the datasets but not in same pair of pattern of repeat. Based on this, one can study further the pattern of specific unit length of repeat nucleotide bases in close proximity to variant sites in pathogenic dataset, to see their correlation to disease development.

Abbreviations

ANOVA	Analysis of Variance
BLAST	Basic Local Alignment Search Tool
bp	base pair
dbSNP	SNP database
DNA	Deoxyribonucleic acid
ftp	File transfer protocol
GCO	Gene conversion
GWA	Genome -wide association
HGVbase	Human Genome Variation Database
IDbase	Immunodeficiency Database
LARD	Large retrotransposon derivatives
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
MUSCLE	Multiple Sequence Comparison by Log- Expectation
NCBI	National Center for Biotechnology Information
nsSNP	Non-synonymous SNP
ORF	Open reading frame
PALS	Pairwise alignment of Long Sequences
RTR	Reverse transcription
SEG	Segmental duplications
SINE	Short Interspersed Nuclear Elements
SNP	Single Nucleotide Polymorphism
SSR	Simple Sequence Repeat
STR	Short Tandem Repeats
TE	Transposable Element
TRA	Transposition
tRNA	Transfer Ribonucleic acid
TRIM	Terminal Repeat Retrotransposons in Miniature
Tukey's HSD	Tukey's Honest Significant Difference
VNTR	Variable Number of Tandem Repeat
WGD	Whole Genome Duplication

Table of Contents

1 Introduction

A repeat is recurrence of a pattern whereby DNA exhibits recurrence of many features (Rao et al., 2010). According to the human genetic variation fact sheet, (http://www.nigms.nih.gov /Education/Factsheet_GeneticVariation.htm) variations can be simply defined as differences or deviations; however, this definition becomes more complex when it applies to genetic variations, as simplicity is lost to subsequent changes which range from harmless to harmful and latent type of change. Since the completion of Human Genome Project in 2003, study of genetic variations occurring in the human DNA (Amigo et al., 2011) sequences became a major focus in order to understand different types of changes which ranges from phenotypic to pathologic types. The understanding of these changes down to the molecular level has been a challenging task in the scientific world and several approaches have been used towards solving these problems.

Several genetic variation studies have been in place. These studies have lead to the discoveries of single nucleotide variation (Chepelev et al., 2009 ; Amigo et al., 2011) all of which have been achieved by different advances in technology, diverse approaches and tools (Kwok et al., 2003; Tang et al., 2008; Dereeper et al., 2011; Mark et al., 2011). Some studies have focused on SNPs, a common type of small genetic variation which occurs within a person's DNA sequence (Sherry et al., 2001; Guo and Jamison 2005). Most genome-wide association (GWA) studies have shown that some SNPs are associated with diseases (Knowles et al., 2008; Ho et al., 2011; Predazzi et al 2012) even though they do not cause disease, they can help to predict the likelihood of developing a particular disease. Likewise, studies on genetic variations leading to disease became relevant being that findings from such studies are important to understanding and solving health issues.

Although, studies on SNPs have focused on their discovery and association with diseases, a suggested and promising approach to understanding the role they play as pathogenic or neutral SNPs, would involve looking at the repeated DNA sequences around the variation site. This is because repeats have been studied and found to have some biological and medical significance on the human genome (Hofferbert et al., 1997). Even though, different studies focused on analysis of simple sequence repeats (Li et al., 2002; Richard et al., 2008) because of their implication in genetic diseases and developments of repeat

analysis tools, there are little or no publications on analysis of repeats at variation sites, therefore the need for this analysis makes me focus on this aspect.

The significance of study

This study will help in the understanding the distribution of the different patterns of repeats which are at close proximity to the variation sites, and possibly give room for further studies on the significance of these repeats on variation sites.

2 Literature review

2.1 Single nucleotide polymorphism

Single nucleotide polymorphisms (SNPs) are described as common genetic variation among people, which results from differences in a single DNA building block called nucleotide. A variation can only be considered as SNP when they occur in at least 1% of the population. They are well known as biological markers that help in the location of genes that are associated with diseases. They have been reported to occur once on every 300 nucleotides on the average (Sachidanandam et al., 2001). An overview of SNPs is given in Figure 2.1.

SNPs may fall within coding or non coding sequences of genes as well as regions between the genes (intergenic regions). Those SNPs found outside the gene are usually referred to as linked SNPs, and have no effect on protein production or function. However some that are found in the gene are usually referred to as causative SNPs, which are further classified into non-coding SNP and coding SNP as illustrated in Figure 2.2.. The non coding SNPs have been reported to cause changes in the amount of protein produced, due to their presence in the gene's regulatory sequences; they affect the level of gene expression, while the coding SNPs cause changes in the amino acid sequence. SNPs that are located in the coding region and results in the amino acid variation in the protein products of genes are also referred to as non-synonymous SNPs (nsSNPs) (Ramensky et al.,2002). While those SNPs in which both alleles produce the target protein are called synonymous SNPs (Nei et al., 2000).

SNPs have been widely used in genetic and genomic studies due to their significance and ability to help in the detection of genes associated with complex diseases (Myles et al., 2008; Sirota et al., 2009; Huang et al., 2009). Data generated on SNP studies are available for the entire public on the NCBI dbSNP database (Sherry et al., 2001) and HGVbase of the Human Genome Variation Database (Brookes et al., 2000).

.

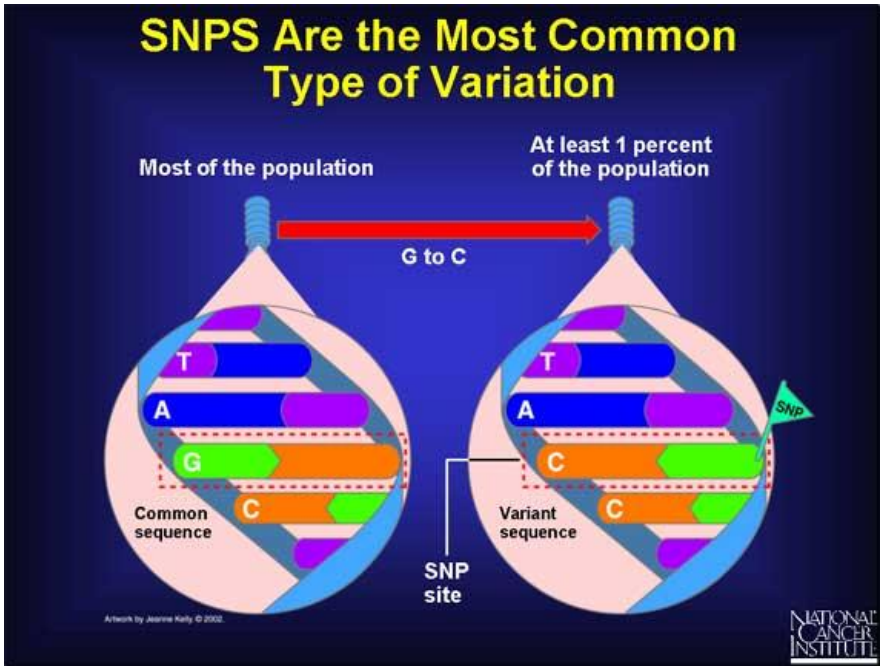


Figure 2.1. Overview of SNPs
Source: http://www.cancer.gov/cancertopics/understandingcancer/geneticvariation/

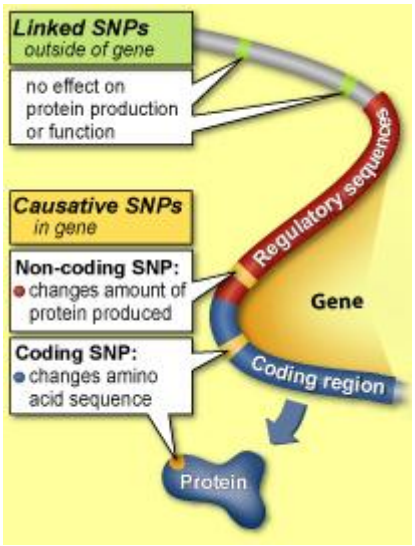


Figure 2.2.Types of SNPs
Source: http://www.learn.genetics.utah.edu/content/health/pharma/snips/

2.2 History of DNA repeats

A repeat is recurrence of a pattern whereby DNA exhibits recurrence of many features. In DNA repeats, the number of occurrences of a pattern is called copy number (Rao et al., 2009). Genome copy number can however be described as number of copies of tandem or interspersed repeats in the genome (Rao et al., 2009).

Far back in the 1960's, repetitive DNA sequences were discovered. This is evident in the study carried out on the organization of eukaryotic genomes by Britten and Kohne in 1968, using renaturation kinetics (Richard et al., 2008). They discovered that a large fraction of the genome is made up of repeated DNA sequences. In the 1980's DNA repeats were grouped into three main categories based on the copy number in the genome (Elder and Tuner, 1995). The classification includes highly repetitive DNA sequences, middle repetitive DNA sequences and unique sequences. This was also reported by two scientists, Long and Dawid in 1980 during their study. They observed that, highly repetitive DNA sequence ranges from several thousands to millions while middle repetitive DNAs number ranges in thousands and unique sequences make up the smallest proportion of DNA (Elder and Tuner, 1995).

Nowadays, advances in sequencing technology have contributed to the availability of complete genome of lots of organisms, thus, more repetitive DNA sequences are being discovered. A vivid example is a study carried out on the sequencing and analysis of the human genome, which led to the discovery of repetitive DNA sequences in the human genome (Lander et al., 2001). Most of these discoveries have thus initiated a better classification of repetitive DNA sequences. Although several classifications are devised, however, a more general and mostly used classification into tandem and dispersed repetitive DNA sequences by (Brown, 2002).

2.3 DNA repeats classification

2.3.1 Tandem Repeats

Tandem repeats are copies of repetitive DNA sequences that lie adjacent to each other in a genomic sequence. They are common in higher eukayotes and account for several percentage of repeats in the whole human genome (Levy et al., 2007). Likewise, previous study have shown that these repeats account for 1.5% of the human genome and occurred in average on every 10kb in the first 2.2 Mb (Riddle et al., 1997). They may be found in both protein coding and non coding regions of the genome (Toth et al., 2000; Katti

et al., 2001; Subramanian et al., 2002). They have also been suggested to evolved as a result of replication slippage or some recombination activity like unequal crossing over or unequal sister chromatin exchange (Kolpakov et al., 2003; Richard et al., 2008).

Tandem repeats have been classified into three sub categories, these are; the satellites, minisatellites and microsatellites.

The satellite repeats have been reported in some studies (Charlesworth et al., 1994; Kulikova et al., 2004; Plohl et al., 2008) as tandemly repeated DNA sequences located in pericentromeric and telomeric regions of the heterochromatin, which are organized in long, megabase-sized arrays. There are various types of satellite DNA in the human genome, these includes: α (alphoid DNA), β, Satellite 1, 2 and 3. α (alphoid DNA) which is located in all chromosomes and has a repeat unit of 171 bp. Also the β, which is found at the centromeres of chromosomes 1, 9, 13, 14, 15, 21, 22 and Y, has 68 bp repeat unit. Satellite 1 can be found in centromeres and other regions in heterochromatin of most chromosomes, having 25-48 bp repeat unit. Satellite 2 and 3 are another type of satellites, located mostly in the chromosomes, having 5 bp repeat unit. Theoretical studies have however shown that, the evolution of these repeats is directed by the mechanisms of concerted evolution which includes unequal crossing over and gene conversion (Ugarkovic and Plohl, 2002). Just like in other studies on repetitive sequences, changes in copy number of repeats in satellite DNA could be accounted for by, biological processes, such as unequal crossing over, replication slippage among several others (Ugarkovic and Plohl, 2002).

The minisatellites, are also of unique characteristics in terms of lengths. They have repeat lengths ranging in 10s, usually between 10-60 bp. They are also called Variable Number of Tandem Repeats (VNTR) and they have been used interchangeably in most texts. They have been found dispersed over the genome and precisely associated with telomere in terms of their location. Like other tandemly occurring repeats, their evolution can be traced back to mechanisms involving, gene conversion and replication slippage (Richard et al., 2008).

The microsatellites, a third subcategory of tandemly repetitive DNA sequence are characterized by their short sequence repeat length which ranges from 2-8 bp and occur in 100s.They are in most text also being referred to as simple sequence repeats (SSRs) and other times short tandem repeats (STRs). There have been recent studies on

microsatellite repeats due to their biological and medical significance (Gulcher, 2012; Yan et al., 2012). For instance, a more common repeat size of the microsatellite is the trinucleotide repeats. These repeats are known to be hypermutable and thus have been implicated in many neurogenetic and other diseases (Benson,1998; Madsen et al., 2008)

Although, previous study on repeats present in the genome, have suggested them as mere junk DNA with no biological significance or function (Ohno, 1972), however, recent studies now suggests that these repeats have functional roles to play (Belkum et al., 1998; Bayliss et al., 2003; Gangwal and Lessnick 2008).

2.3.2 Interspersed Repeat

Interspersed repeat is another class of repetitive DNA sequences in the genome. Sometimes, they are also being referred to as dispersed repeats. They have been reported to constitute about 35-45% of the genome (Batzer et al., 2002; Richard et al., 2008). Thus several scientific contributions have pin pointed their evolution either directly or indirectly from mobile elements or transposons, through the mechanism of transposition (Brown, 2002). The transposable elements or mobile elements were discovered in the 1940's by McClintock during her studies on origin and behavior of mutable loci in maize (McClintock B, 1950; Pray, 2008). Subsequent to this discovery, the classification of the transposable elements into retrotransposon and DNA transposons was done. However, the characteristics of transposon, indicated in their ability to replicate at other locations in the genome (Saha et al., 2008b), made some scientist suggest different classification based on the mode of transposition and insertion mechanism employed in their replication (Saha et al., 2008b; Vukich et al., 2009).

Retrotransposons were further classified as long terminal repeats (LTRs), retrotransposons (Havecher et al., 2004; Kalender et al., 2004), and non-LTR retrotransposons (Ohshima et al., 2005). The LTR retrotransposon includes Gypsy, Copia, LARD and TRIM. The LTR retrotransposon are replicated and mobilized via RNA intermediates and also found to possess some common characteristic features, which includes 5' and 3' untranslated regions (UTRs) containing minus-strand and plus-strand priming sites respectively (Saha et al., 2008b). Gypsy, Copia, large retrotransposons derivatives (LARDs) and terminal-repeat retrotransposon in miniature (TRIMs) are types of LTR retrotransposons, which are characterized by long terminal repeats (LTRs) and

internal open reading frames (ORFs) (Witte et al., 2001; Kalender et al., 2004; Vukich et al., 2009)

The long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) are characterized as non-LTR retrotransposons (non-long terminal repeats) which are transcribed by polymerase II (pol II) and polymerase III (pol III), respectively (Saha et al., 2008b).

DNA transposons were classified as terminal inverted repeat (TIR), miniature-inverted repeat element (MITE), Helitron and Maverick/Polinton (Wicker et al., 2007; Saha et al., 2008b). These elements have various characteristics features, which distinguishes them from one another. The terminal inverted repeat (TIR) is characterised by open reading frame that encodes transposase. Miniature-inverted repeat element (MITE), is characterized by non coding capacity, with presence of non-coding sequence, thus must rely on transposase encoded in trans by autonomous elements. Helitron is characterized by open reading frame, which encodes helicase enzyme and nuclease/ligase activities, while maverick/politon encodes integrase(int), DNA polymerase B(dpolB), and 8 other proteins (Saha et al., 2008b).

In Figure 2.3, Retrotransposons (A–F) are replicated and mobilized through an RNA intermediate via a copy-and-paste mechanism involving the enzyme reverse transcriptase (rt). They possess 5′ and 3′ untranslated regions (UTRs) containing minus-strand (PBS) and plus-strand (PPT) priming sites, respectively. They typically can be divided into long terminal repeat (LTR) retrotransposons (A–D) and non-LTR retrotransposons (E–F). (A) Gypsy elements contain an ORF with gag and pol genes. The gag gene codes for viral capsid proteins while the pol gene codes for proteinase (pr), integrase (int), reverse transcriptase (rt), and RNase H (rh) activities. (B) Copia elements are similar in overall structure to gypsy elements. However, the two groups possess distinctly different reverse transcriptase amino acid sequences. In most instances, they also exhibit variation in the relative position of int. (C) In LARDs (large retrotransposon derivatives), protein-coding regions have been replaced by a relatively long, conserved, noncoding region. (D)

TRIMs, terminal repeat retrotransposons in miniature, contain short LTRs, PBS and PPT sites, and little else. (E) LINEs, long interspersed nuclear elements, are non-LTR retrotransposons that possess 1 or 2 ORFs. One ORF encodes a pol protein with rt and endonuclease (en) activities. If there is a second ORF, it encodes a nucleic acid binding

protein (nabp) with chaperone and esterase activities. The 3′ UTR sometimes contains the canonical polyadenylation sequence (ATAAA) and a tract of poly-A. LINEs are transcribed by RNA polymerase II. (F) SINEs, short interspersed nuclear elements, possess a region with similarity to a tRNA (TR) or other small RNA, a tRNA-unrelated
region (TU), and a region that, in some instances, appears to be LINEderived (LD). SINEs are transcribed by RNA polymerase III. DNA transposons (G–J) can be mobilized through either a cut-and-paste mechanism (G–H) or through other mechanisms that do not involve RNA intermediates (I–J). They multiply via their host's replication machinery. (G) TIR DNA transposons (cut-and-paste) are characterized by terminal inverted repeats (TIRs) and one ORF that encodes a transposase gene. (H) MITEs, miniature inverted-repeat transposable elements, are extremely short, TIR-flanked, cut-and-paste transposons with no coding capacity. (I) Helitrons are DNA sequences that are propagated through a rolling-circle replication mechanism. Autonomous Helitrons possess a helicase gene that encodes an enzyme with 5′-3′ helicase and nuclease/ligase activities. Helitrons may also contain genes for RPA-like (rpal) single-stranded DNA binding proteins. Helitrons do not create target site duplications and lack TIRs. Both autonomous Helitrons and most non-autonomous Helitron-like transposons have conserved 5′-TC and 3′-CTRR sequences at their termini. (J) Mavericks/Polintons are large elements that encode integrase (int), DNA polymerase B (dpolB), and up to 8 other proteins. It has been argued that Mavericks/Polintons contain all of the genes necessary for both self-transposition and self replication (Saha et al., 2008b)

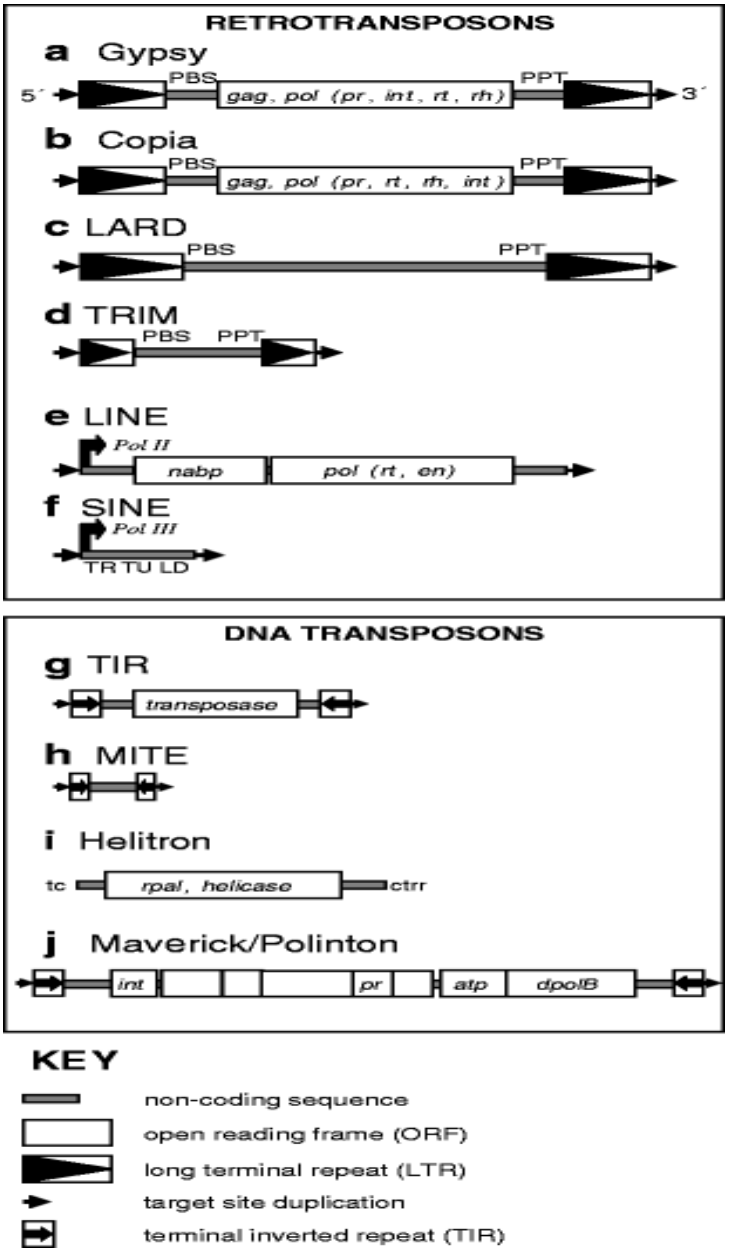


Figure 2.3: Transposons are traditionally classified into retrotransposons and DNA transposons (Saha et al., 2008b)

2.4 Pattern of repeats

DNA repeats could exist in one of the following patterns.

- Forward or direct repeat
- Reverse repeat
- Complement repeat

- Palindromic repeat

Forward repeats are repeats that have direct match direction with original nucleotide sequence. They can also be referred to as direct repeats and they are given in Figure 2.4. In reverse repeats, the nucleotide sequences match each other in a reverse direction which is given in Figure 2.5. In complement repeat, the repeats are complements of the original nucleotide sequence and the direction of the match is direct. The complement repeat is given in Figure 2.6. Palindromic repeat is a reverse complement of the original; this is given in figure 2.7.

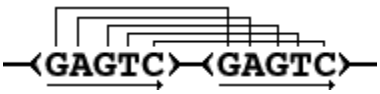


Figure 2.4. Forward(direct) match

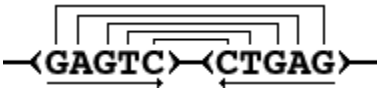


Figure 2.5. Reverse match

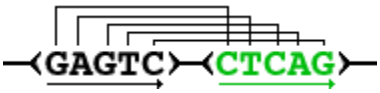


Figure 2.6. Complement match

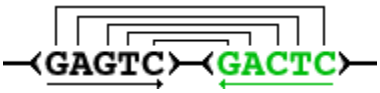


Figure 2.7. Palindromic match

Source: http://bibiserv.techfak.uni-bielefeld.de/reputer/manual.html

Forward repeats are repeats that have direct match direction with original nucleotide sequence and they are referred to as direct repeats. Reverse repeats on the other hand, the nucleotide sequence match each other in a reverse direction. Complement repeats also exist with a different match pattern. The repeats are complements of the original nucleotide sequence and the direction of the match is direct, however in palindromic repeat, the repeat is a reverse complement of the original (Kurtz et al., 2001).

2.5 Distribution of repeats in the human genome

Experimental study has shown that the human genome is the first repeat-rich genome to be sequenced and it consists of at least 50% of repeats (Lander et al., 2001). These include tandem repeats and interspersed (Richard et al., 2008). Following this discovery, several studies on repeats have concentrated on their distribution across the human genome. Studies on distribution of repeats have focused on different types of repeats and the regions where they can be found. Subramainan and his colleagues during their study on triplet repeats discovered that these repeats can be found in both protein coding and non-coding regions (Subramanian et al., 2003), likewise, a study done by Usdin, (Usdin, 2008). Also, during a study by Katti et al., (Katti et al., 2001) on the differential distribution of simple sequence repeats in eukaryotic genome sequences, they pointed out that several such experimental studies have brought about growth of sequence databases.

2.6 Mechanisms of evolution of repeats in the human genome

Owing to the increase in the number of repeat related diseases, a need to closely study repetitive DNA sequences turns out to be an appealing approach, to understand how these repeats are associated with diseases. The knowledge of the evolutionary mechanisms of these repeats is again an important thing to look into. It has been reported by various authors, that genes with expanded repeats either exhibit loss or gain of function (Siwach et al., 2008), thus the underlying mechanisms need to be well understood. The different categories of repeats have been pointed out with different mechanisms of evolution (Richard et al., 2008). These mechanisms range from replication slippage, gene conversion, whole genome duplication, segmental duplication, transposition to reverse transcription (Richard et al., 2008). Tandem repeats often undergo replication slippage and gene conversion while dispersed repeat undergo transposition and reverse transcription, this is shown in figure 2.8.

Replication slippage is a genetic process in which deletions and insertions of small contiguous repeats occur because of misalignment between DNA stands (Micheal, 2004). This causes parts of the template DNA to be copied more than once or missed out during replication. Gene conversion on the other hand, involves non-reciprocal transfer of information between two homologous genes, where one segment replaces nucleotides in its corresponding homolog (Shuging et al., 2008). This usually causes concerted evolution

in gene families through reciprocal exchange of sequence between paralogs (Teshima et al., 2004).

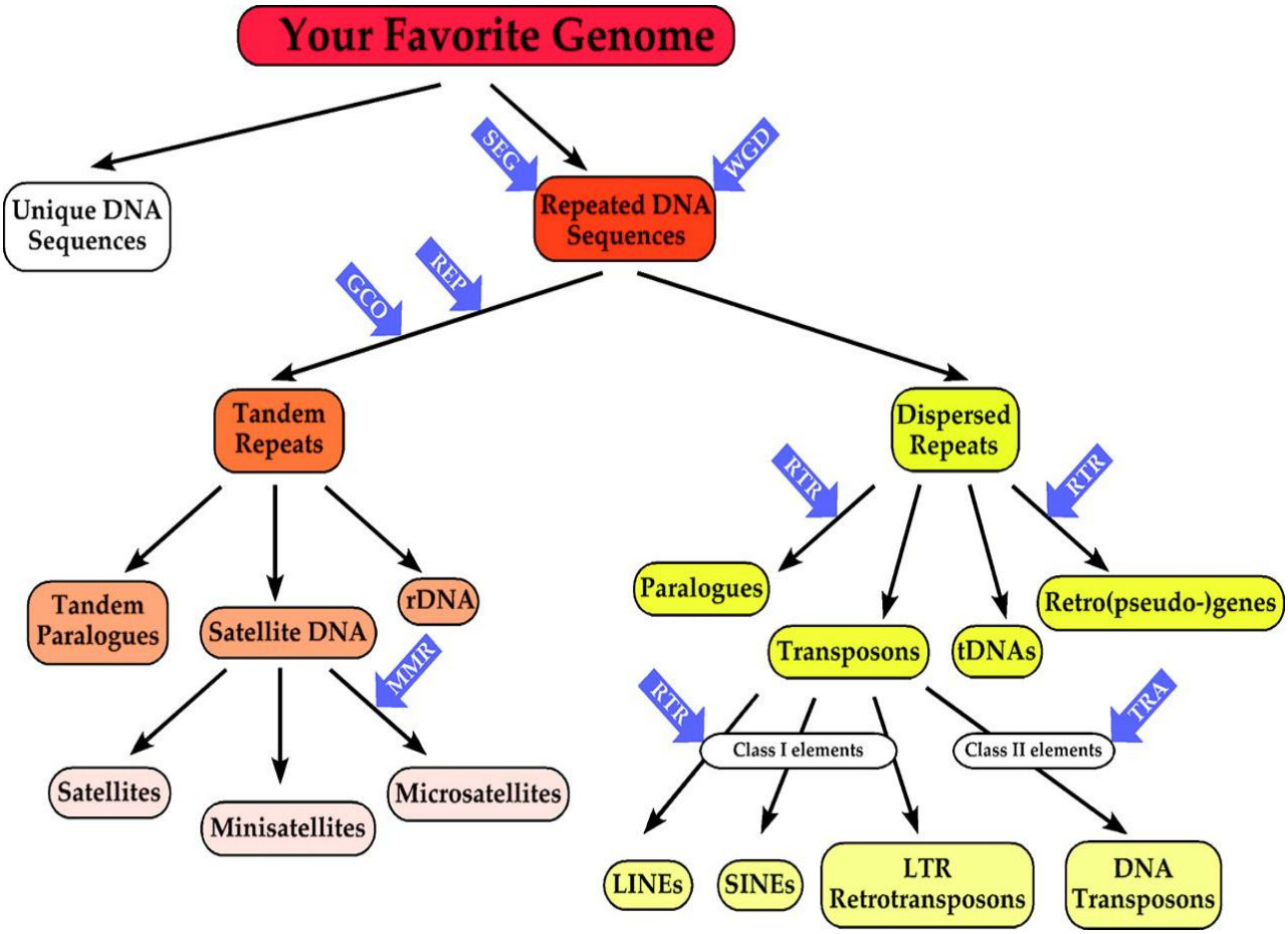


Figure 2.8. Repeated DNA sequences in eukaryotic genomes and mechanisms of evolution. The two main categories of repeated elements (tandem repeats and dispersed repeats/interspersed) are shown, along with subcategories, as described in the text. Blue arrows point to molecular mechanisms that are involved in propagation and evolution of repeated sequences. REP, replication slippage; GCO, gene conversion; WGD, whole-genome duplication; SEG, segmental duplications; RTR, reverse transcription; TRA, transposition (Richard et al., 2008).

2.7 Biological significance of DNA repeats

Some recent studies have reported the biological significance of repetitive DNA sequences (Ugarkovic and Plohl, 2002; Usdin, 2008), contrary to the previous studies which suggests that these DNA sequences are mere “Junk DNA” with unknown function (Ohno, 1972).

Several classes of repetitive DNA sequences have been suggested to have different function and contributions to the genome, thus ascribing them parasitic attributes (Orgel and Crick, 1980) is an old believe.

Study (Ugarkovic and Plohl, 2002) on satellite DNAs, reported their association with complex organizational features necessary for the function of eukaryotic genomes, which includes formation of heterochromatic genomic compartments important for chromosomal behavior in mitosis and meiosis. Also study by Jiang et al (Jiang et al., 1996; Jiang et al.,2003) has suggested the role of α- satellite DNA sequences in the centromere function of eukaryotic chromosomes and other studies which identified SSRs in the centromere of several organisims have indicated their contributions to centromere organization (Centola and Carbon, 1994; Murphy and Karpen 1995; Schmidt and Heslop-Harrison 1996; Brandes et al., 1997; Cambareri et al 1998; Areshchenkova and Ganal 1999) and functional role in sister chromatid cohesion and indirect assistance in kinetochore formation (Murphy and Karpen 1995). The α-satellite DNA sequences have been recognized as the largest tandem repeat family found at all normal human centromere (Komissarov et al., 2001; Rudd et al., 2006). Thus they are well studied and they have provided paradigm for understanding the genomic organization of tandem repeats (Schueler et al., 2001; Rudd et al., 2004).

Likewise several SSRs and minisatellites are known to have some role to play in the regulation of DNA metabolic processes, which includes recombination and replication activities. SSRs have been found to cause recombination directly due to their effects on DNA structure just as repeat number and motif equally could affect recombination (Gendrel et al., 2000; Li et al 2008)

Ijaz and Khan reported that microsatellite markers could characterize and discriminate all genotypes (Ijaz and Khan 2009; Ijaz, 2010), these markers have been reported with high level of polymorphism as well as higher reliability with highest polymorphism content, thus used as molecular marker for fingerprinting (Weising et al., 1995: Diwan and Cregan, 1997; Ashikawa et al., 1999).

Also SSRs have been implicated in gene regulatory activities such as their participation in transcription, expression of gene, protein binding activities, translation (Li et al., 2008). Similar research by Bayliss et al on the mutational mechanism and contributions of simple sequence repeats have that repeats play some roles in the expression of gene (Bayliss et

al., 2003), likewise another study has clearly revealed their role in regulatory function (Belkum et al., 1998). Experimental studies, also suggested that simple sequence repeats are capable of forming a variety of unusual DNA structures with simple and complex loop-folding patterns (Li et al., 2008), such structures may thus have regulatory effects on gene expression (Fabregat et al., 2001)

2.8 Medical implications of repeats

Repetitive DNA sequences in the human genome are of both of biological and medical importance. Studies have shown the role of simple sequence repeats specifically, the trinucleotides in genetic disease development (Sutherland and Richards 1995; Gleicher et al., 2009; Kiliszek et al., 2010). Early studies from 1991 till date have shown that trinucleotide expansion have established relationship with certain human genetic diseases (Bates et al., 1994; Kremer et al., 1991; Hummerich et al., 1995; Arenas-Aranda et al., 1999). Trinucleotide repeat expansions have been implicated as an underlying cause of neurogenetic disorders such as fragile X syndrome of mental retardation and Huntington's disease among several other Mendelian neurological disorders (Paulson et al., 1996; Vincent et al., 2000; Kiliszek et al., 2010). Trinucleotide repeat expansion discovery originally came from the study of the human inheritable fragile X syndrome disorder (Fu et al., 1991; Oberle et al., 1991; Verkerk et al., 1991), which is caused by mutational expansion of untranslated CGG repeats of the first exon of the fragile X mental retardation gene (FMR1) (Salat et al., 2000). Below in Table 2.1 are some examples of genetic diseases caused by trinucleotide repeat expansion.

Another study on microsatellite by Gangwal and Lessnic, 2008 (Gangwal and Lessnick, 2008), showed that, EWS/FLI interacts with GGAA microsatelittes to regulate some of its target genes. EWS/FLI is a known aberrant ETS-type transcription factor that dysregulates a number of genes that are important in the development of Ewing's sarcoma (a solid tumor of bone which occurs in children and young adults).

Table 2.1. Examples of genetic diseases caused by trinucleotide repeat expansion

		Number of copies of repeat	
Disease	Repeated sequence	Normal range	Disease range
Spinal and bulbar muscular atrophy	CAG	11-33	40-62
Fragile-X syndrome	CGG	6-54	50-1500
Jacobsen syndrome	CGG	11	100-1000
Spinocerebellar ataxia	CAG	4-14	21-130
Autosomal dominant cerebellar ataxia	CAG	7-19	37-220
Myotonic dystrophy	CTG	5-37	44-3000
Huntington disease	CAG	9-37	37-121
Friedreich ataxia	GAA	6-29	200-900
Dentarubral-pallidoluysian atrophy	CAG	7-25	49-75
Myoclonus epilepsy of Unverricht-Lunborg type	CCCCGCCCCGCG	2-3	12-13

Source: http://www.nature.com/scitable/content/examples-of-genetic-diseases-caused-by expanding-27926

2.9 Methods and tools for finding DNA repeats

Several computational method and tools have been developed towards finding repeats in the eukaryotic and prokaryotic genomes. Repetitive sequences finding tools could either focus on tandem or dispersed repeats in the genome (Charleswort et al., 1994; Ellegren, 2004; Richard et al., 2008). However, due to the nature of occurrence of tandem repeats, a sequence of consecutive or nearly consecutive copies along a DNA strand, tools for identification of such repeats are in abundance (Saha et al., 2008b), but this is not so for dispersed repetitive DNA sequences like transposons, because their distribution is non-tandem in the human genome. The development of computational approaches and tools towards dispersed repetitive DNA sequences became a major focus in this field and

interestingly, some of these tools, such like RepeatMasker, have been developed and found useful in search for tandem repeats (Smit et al., 2004).

In the hunt for suitable repeat analysis tool for this project, it was discovered that some scientist, Surya and his colleagues have focused on computational approaches and tools used in the identification of dispersed repetitive DNA sequences. They found out that several tools have been developed and thus, they analyzed the approach which most of this tools use in finding dispersed repeats (Saha et al., 2008b; Lerat, 2009). However there are qualities which these analysis tools must fulfill in order to be classified as good enough for systematic study of repeats in the genome. These qualities include efficiency, flexibility and significance, interactive visualization and compositionality (Kurtz et al., 2000).

Kurtz and his colleagues described an efficient tool as one whose algorithm is practically linear in terms of computer memory and execution time. Likewise flexibility and significance of the tool should take into account the recognition of direct, palindromic and some other sequence features with its ability to also evaluate the statistical significance of the repeats. They also pointed out that a good tool should give an interactive visual overview of repetitive regions obtained for effective understanding. Furthermore, they concluded by saying that a repeat analysis tool should provide a suitable platform for further analysis on repeats.

Among the tools that were studied by Saha et al. (Saha et al., 2008b), REPuter (Kurtz et al., 2001) seems to best fit these criteria, thus its use in this analyses becomes important.

There are two general approaches in the identification of dispersed repeats (Saha et al., 2008b). These include reference based method of repeat identification and Ab initio method of repeat identification.

2.9.1 Reference based repeat identification approach

The reference based repeat identification approach requires that a given dataset is compared with reference dataset in a database, such as Repbase database (Jurka et al., 2005). This approach uses two methods, which includes, library based technique and signature based technique. (Saha et al.,2008b; Lerat, 2009).

In library based technique, identification of repeat in a given set of data and of the repeats with a reference repeat sequences in the database is done. Typical examples of

tools using this technique includes, RepeatMasker (Smit et al.,2004), CENSOR (Jurka et al., 1996) and PLOTREP (Toth et al., 2006).

Identification of repeat by RepeatMasker is done in conjunction with RepBase (Smit et al., 2004 ; Saha et al., 2008b). RepBase is a large curated repeat library containing data from numerous eurkaryotes. It can also be used in conjunction with clade specific repeat databases such as TREP (Wicker et al., 2002), STRbase (Ruitberg et al., 2001) and The TIGR plant repeat database (Ouyang et al., 2004).

In signature based techniques, identification of repeats requires a search for nucleotide sequence or amino acid motifs and spatial arrangement characteristics of a particular repeat group (Saha et al., 2008b). This is different from library based technique in that it employs heuristics based on a priori information of particular repeat type (Saha et al., 2008b). Typical examples of tools using this technique includes, FINDMITE (TU, 2001), Inverted Repeat Finder (Warburton et al., 2004), LTR_STRUC (McCarthy et al., 2003) and TE-HMM (Andrieu et al., 2004).

2.9.2 Ab initio Repeat identification approach

Studies (Saha et al., 2008a; Lerat, 2009) have described this type of repeat identification approach as ab initio repeat identification method, which is also being referred to as a non-reference based repeat identification approach. This approach of repeat identification requires no prior knowledge or availability of reference repetitive sequence or motif in the identification of repeats from a given dataset. Several ab initio repeat identification tools are available (Saha et al., 2008a; Lerat, 2009), they also have different approach in the identification of repetitive DNA sequences, usually two stages are involved in the identification of repeats. The first stage involves the identification of repeat sequences while the second stage involves identification of repeat families that are present in the repeats. These two stages thus involve several approaches which have also been adopted by these tools. Identification of repeats using ab initio approach could be classified into groups, these are:

2.9.2.1 Self- comparison approach

In self comparison approach, uncharacterized DNA sequences are aligned with itself in order to identify clusters of similar sequences (Saha et al., 2008a). Tools such as RECON (Bao and Eddy, 2002), PILER (Edgar and Myers, 2005) and PILER–CR (Edgar, 2006) use this approach.

RECON tool is implemented in C and Perl and Its first approach is to make a all-to-all BLAST search using WU-BLASTN and then use an extended approach of single linkage clustering of local pair wise alignments to align results obtained (Bao and Eddy, 2002). Thus, can be used to identify and classify repeat sequences from genomic sequences, however it has been reported to be poor with processing of short-period tandem repeat (Bao and Eddy, 2002).

PILER **(**Edgar and Myers, 2005) tool uses pair-wise alignment of long sequences (PALS). PAL identifies repeats in assembled genomic regions and searches for repeat families with similar profile characteristics to known repeat types (Edgar and Myers, 2005).

2.9.2.2 K-mer approach

In k-mer approach, exact substrings of given length is used in the identification of repeats. Tools such as ReAS (Li et al., 2005), RepeatScout (Price et al., 2005), RAP (Campagna et al., 2005), RepeatFinder (Volfovsky et al., 2001), RepeatGluer (Pevzner et al., 2004) and REPuter (Kurtz et al., 2001) are tools that use this approach.

REPuter (Kurtz et al., 2001) this tool was used for this analysis because of it satisfies repeat analysis tool criteria and fulfill its ability in finding of repeat patterns required in this analyses. Its program family, REPfind, REPselect and REPvis have also, one way or the other meet these given criteria.

REPfind (Kurtz et al., 2001) is a program family of the REPuter which uses an efficient and compact implementation of suffix trees to locate exact repeats in linear space and time. It computes all maximal repeats in the input sequence contained in the file and gives out the result as standard output. In REPfind, the size of the output can be limited by parameters for minimum length and maximum error and sorted by E-values calculated based on the distance model used. In cases where error occur during computation, the program quits and gives exit code 1, otherwise exit code 0.

To use REPfind, the input file should either be in fasta format or in plain format. The input sequences intended to be analysed are usually nucleotides (Adenine: A, Cytosine: C, Guanine: G and Thymine: T). There are REPfind options that could be used the input sequence file to get the required results, these options include

- -f : this reports maximal forward repeats

- -p : this reports maximal palindromic repeats
- -r : this reports maximal reversed repeats
- -c : this reports maximal complemented repeats
- -l L : this specifies length and repeats of length of at least L are reported, and this must be a positive integer
- -allmax : this reports all maximal repeats in the order in which they are found
- -s : this reports the substring of repeat sequences
- -l : this shows the distribution of the different repeats about the length

REPselect (Kurtz et al., 2001) is another program family of REPuter. It allows user to select interesting repeats for further use in subsequent analysis

REPvis (Kurtz et al., 2001) is also a program family of REPuter, it allows for the visualization of REPfind results. The display of the result is controlled by a scroll bar and zooming in and out is allowed for proper visualization of the results.

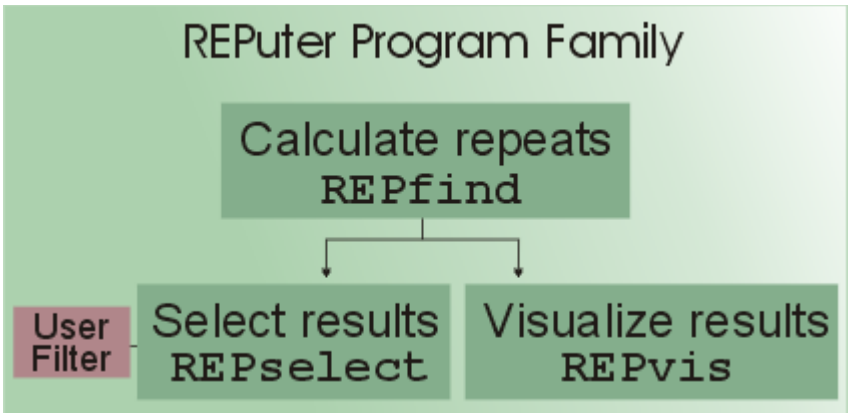


Figure 2.9. Reputer program family showing REPfind, REPselect and REPvis
Source: http://bibiserv.techfak.uni-bielefeld.de/library/reputer/manual/intro.html

ReAS (Li et al., 2005) tool works on transposable element (TEs) which have high copy number across the genome and are not too old, otherwise, still recognizable in comparison to their ancestral sequences. It first computes k-mer depth, which is the number of times

that a k-mer appears in the shotgun data, after which it retrieves all reads contained in the k-mer. The reads are then assembled into initial consensus sequence (ICS) by using ClustalW. Afterwards new k-mers at the end of the consensus are extended until no further extensions are possible (Li et al., 2005).

RepeatScout (Price et al., 2005) tool works by building a set of repeat families by using high frequency k-mer as seeds and then greedily extends it to repeat family definition (Price et al., 2005; Saha et al.,2008a,b)

RepeatGluer (Pevzner et al., 2004) tool finds all sub-repeats in a genomic sequence by using de Bruijn graphs. De Bruijn graph is a representation of every l-mer in a genomic sequence as vertex connected with edges especially if they are overlapping l-mer in the genome (Pevzner et al., 2004). Subsequent to this, the tool identifies the consensus sequence as well as the copy number in the sub-repeats, thus forms a copy of the genomic sequence with this and each sub-repeat is substituted by its consensus sequence (Pevzner et al., 2004)

RepeatFinder (Volfovsky et al., 2001) tool uses suffix tree data structure for identifying exact repeats. The set of exact repeats identified subsequently serves as a basis for building repeat classes (Volfovsky et al., 2001).

3 Aims and Objectives

The general aim of this research is to find nucleotide repeats that are present around variation sites of neutral and pathogenic SNP datasets from VariBench.

Objectives were to:

- Find suitable repeat analysis tool, which can detect four different patterns of repeats, which includes, direct/forward repeats, palindromic repeats, reverse repeats and compliment repeats, within a nucleotide sequence of specified length.
- Find the genomic sequences for pathogenic and non pathogenic SNP ids from the NCBI ftp site.
- Carry out repeat analysis with the analysis tool (REPuter) in order to detect the different types of repeats and their position within a 21 unit length of nucleotide sequence.
- Investigate the repeat types that are prevalent in the two data sets.
- Investigate the nucleotides present within different types of repeats and their abundance within the two dataset.
- Conduct statistical analysis to investigate if there are differences between the two dataset.

Research Problems.

Null Hypothesis

I) There is no significant difference between the repeat in neutral and pathogenic dataset.

II) There is no significant difference within the repeats of neutral dataset.

III) There is no significant difference within the repeats of the pathogenic dataset.

Alternative Hypothesis

I) There is significant difference between the repeats in neutral and pathogenic dataset.

II) There is significant difference within the repeats of neutral dataset.

III) There is significant difference within the repeats of the pathogenic dataset.

4 Materials and Methods

4.1 Materials

4.1.1 Dataset of neutral SNPs

This is a neutral dataset or non synonymous coding SNP dataset retrieved in April 2011 from a benchmark database for variations known as VariBench (Nair and Vihinen, 2012 http://bioinf.uta.fi/VariBench/) The dataset was from dbSNP database (Sherry et al., 2001) build 131 and it consists of 21,170 human neutral SNPs with allele frequency and chromosome sample count (Thusberg et al., 2011). The dataset has also been filtered for the disease-associated SNPs and variant position mapping has been extracted from dbSNP database.

4.1.2 Dataset of pathogenic SNPs

This is a pathogenic dataset, also downloaded in April 2011 from VariBench (Nair and Vihinen, (http://bioinf.uta.fi/VariBench/).The dataset consist of 19,335 missense variations, which were retrieved in June 2009 by Thusberg et al. (Thusberg et al., 2011) from PhenCode database(Giardine et al., 2007), IDbases (Piirilä et al., 2006) and 18 individual Locus Specific Database (LSDB). The dataset also consist of variant position mapped to RefSeq protein(>=99% match), RefSeq mRNA and RefSeq genomic sequences.

4.1.3 Database and tool

4.1.3.1 Reference Sequence (RefSeq) database at NCBI

The Reference sequence is a database of nucleotide and protein sequences with feature and bibliographic annotation. It is managed by National Center for Biotechnology Information (NCBI) a division of the National Library of Medicine located at US National institute of Health (NIH) (Pruitt et al., 2009). RefSeq resources can be accessed via FTP site on NCBI (Sayers et al., 2010) and several other methods which includes BLAST programs (Altschul et al., 1990; Altschul et al., 1997), scripted query by using E-Utilities and interactive query on the internet using text Entrez (Schuler et al., 1996). The NCBI builds RefSeqs from the sequence data which are available in the archival database GenBank (Benson et al., 2009). Although RefSeqs and GenBank records are both retrievable from NCBI interface (Pruitt et al., 2009; Benson et al., 2009). RefSeq records are not part of GenBank (McEntyre et al., 2002)

4.1.3.2 REPuter repeat analysis tool

REPuter is a repeat analysis tool. Its search engine REPfind uses an efficient and compact implementation of suffix trees to locate exact repeats in a linear space and time (Kurtz et al., 2001). REPuter takes in input data as a FASTA formatted file and gives the possibility of searching for repeats in four directions, which includes forward (direct) match, reverse match, complement match and palindromic match. It computes maximum repeats and shows repeats with small Expected value (E-value), which is a parameter describing the number of hits one can expect to see by chance (McEntyre et al., 2002) when searching the sequence of minimal repeat size of specified length. The output of the analysis has seven main parts indicating the repeat length of the first part of the repeat, starting position of the first part of the repeat, match direction of the repeat, repeat length of the second part, starting position of the second part of the repeat, distance of the repeat and the calculated E-value respectively. This output can be saved to a text file.

4.2 Methods

4.2.1 Download and processing of data

4.2.1.1 Download of Genomic sequence files

The whole human genomic sequences were downloaded from two file directories via the ftp site (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens). These are chromosomes directories and the Assembled_chromosomes directory. The files in the chromosome directories provide concatenated sequence data for scaffolds that have been assembled via individual GenBank records. These scaffolds are the same as those that are presented on the NCBI Map Viewer. The sequences thus include reference assembly and may also include alternate assemblies when available.

The files in the Assembled _chromosomes directory and its sub-directories ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/README., provide data for assembled chromosomes, unlocalized scaffolds (those scaffolds that are not associated with a specific chromosomes but which can be ordered on that chromosome), unplaced scaffolds (those scaffolds that are not associated with any chromosome), and in some cases scaffolds from alternate locus groups or genome patch. Patches are short sequences which are in form of contigs or scaffolds that have been released outside the normal cycle of major releases. To obtain the complete set of data for an assembly, seq subdirectory was visited and download of hs_ref_GRCh37.p5_chr(1-22,X and Y).fa.gz was done.

4.2.1.2 Preprocessing of Genomic sequence files

Command g unzip hs_ref_GRCh37.p5_chr* was executed on the command line interface so as to unzip the files for easy accessibility. The Fasta files contained the sequence identifiers followed by bare sequence lines. The sequence identifier includes:

- NCBI sequence identifier referred to as Geninfo integrated database id(gi|integer)
- Genomic RefSeq Accession.Version (ref|NT_* or NC_*). NT_ * is the Accession. Version for Scaffolds and NC_* is the Accession
- Version for chromosomes, while the last part of the sequence identifier indicates
- organism genomic sequence description.

The most important information for these analyses is the genomic RefSeq Accession. Version, which is useful for the retrieval of sequence length of 21 bp on every given data.

Python scripts were written in order to retrieve the sequence length of 21 bp. The 21 bp includes the variation sites located on the 10^{th} position (counting from 0 in python) in both neutral and pathogenic data sequences. The order in which the data was processed is given below.

4.2.1.3 Dictionary creation

For successful mapping of the variation position with the given genomic sequence id and adequate retrieval of nucleotide sequence length of 21 bases, a Python dictionary script was written. A dictionary is described as associative memories in languages. Python programming language holds an unordered set of key: value pair. The keys in this case are represented by the genomic sequence ids (Genomic RefSeq Accession.Version), while the values are represented by the corresponding variation positions on the sequences. A particular genomic sequence id might have more than one specified variation position, thus creating a Python dictionary script was the best option rather than using default dictionary creation python script, and this is given in Figure 4.1. The script written thus has the sequential steps, which are briefly described below and was applied for both pathogenic dataset and neutral dataset

- opening and reading of text file named “anoda.txt” which contains both the Genomic RefSeq Accession.Version and the corresponding variation positions

- preprocessing of each line in the file such that newline character ("\n") and tab ("\t") are striped off and processed under a new specified variable
- creation of empty dictionary, specified as d={}
- iterating over every line in the specified variable and keeping the first item (Genomic RefSeq Accession.Version) on the line as a key and the integer of the second item (variation positions) as value, given that the Genomic RefSeq Accession.Version is in the dictionary, variation positions are then assigned to it , otherwise the Genomic RefSeq Accession.Version is registered if it does not occur before. This continues until the code runs over all lines on the list.

```
f=open("anoda.txt","r")
mylines=f.readlines()
import string
my_lines=map(string.strip, mylines)
my_lines2=[]
for line in my_lines:
   line=line.split("\t")
   my_lines2.append(line)
        d={}
        for item in my_lines2:
        key=item[0]
        value=int(item[1])
       if key in d.keys():
                d[key].append(value)
     else:
      d[key]=[value]
```

Figure 4.1. Python dictionary script for analysis

4.2.1.4 Extraction of Genomic RefSeq Accession.Version

The downloaded Fasta files were preprocessed in order to retrieve the Genomic RefSeq Accession.Versions that is present in all downloaded files; this also applies to both dataset. This process helped in matching the given Genomic RefSeq Accession.Version in the dataset with downloaded ones so as to ascertain that they correspond to the given variation positions, thus, ease the subsequent retrieval of the nucleotide sequences.

The task required that python function "my_newid" be defined. In this function, split method was used on the seq_record.id at "|" positions and then input into a specified variable "myid" so as to get out the required information. Using assert method, the length of the split seq_record.id was true to be 5 and "myid" item 3 (myid[3]), holds the Genomic RefSeq Accession.Version . This function was later applied for sequence retrieval. Python scripts written for the extraction of the RefSeq Accession.version is given Figure 4.2

```
def my_newid(s):
    myid=seq_record.id.split("|")
    assert len(myid)==5 and myid[0]=="gi" and myid[2]=="ref"
    return myid[3]
```

Figure 4.2. Python script for RefSeq.accession.version retrieval

4.2.1.5 Sequence retrieval

Python script was written to retrieve 21 bp, which includes the variation position as the midpoint of the sequence. The first part of the script initiated the opening of a new file where the sequences were to be written as a text formatted file. The next part of the script used some Biopython, which is a set of libraries that provides ability to handle biological data (Cock et al., 2009). The Biopython script initiated the use of SeqIO, which is an interface for the input and output of sequences of diverse formats and it specially takes into use SeqRecord objects. After the initiation of the SeqIO, sequence input function was allowed to read and download genomic sequence in fasta format thus return a SeqRecord iterator.

On the return of SeqRecord iterator, genomic RefSeq Accession.version retrieval function which has been earlier defined as my_newid was invoked on the seq_record.id, to

retrieve the genomic RefSeq Accession.version present in the file into a variable "id". The genomic RefSeq Accession.version in the variable "id" was then checked to ensure that it is present in the already created dictionary. If the RefSeq Accession.version is present in the dictionary, then the variation positions corresponding to the RefSeq Accession.version as stored in the dictionary are further processed for sequence retrieval.

Given a variation position, the seq_record.seq item was iterated to retrieve 10 nucleotide bases upsteam and downstream, this makes 21 nucleotide bases, the requirement for the analysis. The results were then written to text files for further analyses. This process was repeated for all the downloaded fasta files (chromosomes 1-22,X and Y) for both datasets. Python code for this is given in the appendix.

4.2.2 Repeat Analysis

Usually REPuter takes in fasta formatted file, although, the written text file were merely lines of sequences, they were still usable, due to the fact that the tool usually ignores the description line of the fasta file and deals directly with sequences.

However, having many lines of sequence, each representing specific variation position, a need to handle one sequence line at a time becomes very important for accurate analysis. Thus preprocessing of the sequence generated files was done.

4.2.2.1 Preprocessing of sequence generated files

Python script was written to separate each line of the sequence file into a single file and it is described below.

- The python script opened each file in read mode to iterate over lines in the file
- Strip method was used on lines in the file to remove newline ("\n") characters.
- Unique variables were created using range method, which was applied to the length of lines of item in the file. This conferred uniqueness in the file naming system.
- Zip method was initiated on the generated variable and RefSeq Accession.Version of the sequences and written as text file.
- Parts of the hundreds of text file names created were distributed into different text files for further analysis.

4.2.2.2 Repeat generation with REPuter

Repeat analysis was carried out on the command line interface by invoking a Perl script on Repfind; a REPuter search engine. The script used repfind command on text files containing uniquely named sequence file and result was directed into another text file named according to the initial sequence file but differentiated by result_k*_seq tag. Below in figure 4.3 is a representative of the Perl script.

```
perl -p -e 'system("repfind -f -p -r -c -l 2 $_")' M_non_pathogenicset_k1.txt > M_non_pathogenicset_repeat_result_k1_seq.txt
```

Figure 4.3. Perl script for repeat generation.

4.2.2.3 Description of repfind command used in the analyses

The repfind command "repfind -f -p -r -c -l 2 -s" finds four pattern of repeats in the 21 bp, which are forward (-f), palindromic (-p), reverse (-r) and complement (-c) with minimum length of repeat (-l) specified as 2 bp alongside with the representing sequences (-s), in other words, it searches for repeat patterns of various kinds with length starting from 2 bp and above.

4.2.2.4 Preprocessing of repeat result files

Results obtained from the repeat analysis done with REPuter were being processed to get the necessary information required for further analysis. There are several lines in the result file, the first and second lines are the description lines while the following lines are the main result lines.

The first line shows the repfind script that was invoked on specified text file while the second line shows the length of the sequence in the file, the starting item number, counting from 0, the minimum length of repeat and file name containing the sequences respectively.

The main result lines following the description lines, for example the first result shown in the example below (8 4 R 8 4 0 1.89e-03 gccttccg), shows the length of the sequence (8 for this example), the starting position of the repeat (position 4 counting from 0), the repeat pattern which occurred at the position (R- reverse), again length of repeat (8), then second position at which the repeat occur (4 same as first position but may differ in most cases), the next 0 value indicates insersion or deletion was not specified, just a perfect match was sort for, then, the e-value obtained (indicating the significance of the repeat at

the position) and finally, the sequences involved (gccttccg for this example). Python scripts were used to retrieve the required information. A representative of the result file is given in Figure 4.4.

```
# repfind -f -p -r -c -l 2 -s NC_000024.9_2655246_21.txt
# 21 0 2 NC_000024.9_2655246_21.txt
8  4 R 8  4  0 1.89e-03 gccttccg
4  0 F 4  6  0 4.84e-01 cttc
4  0 R 4  6  0 4.84e-01 cttc
4  0 R 4  0  0 4.84e-01 cttc
4 10 P 4 17  0 4.84e-01 cgac
3  2 F 3 18  0 1.94e+00 tcg
3  2 P 3 10  0 1.94e+00 tcg
3  2 P 3 13  0 1.94e+00 tcg
3  3 R 3  3  0 1.94e+00 cgc
3  5 P 3 15  0 1.94e+00 cct
3  8 C 3 15  0 1.94e+00 tcc
3 10 F 3 13  0 1.94e+00 cga
3 13 P 3 18  0 1.94e+00 cga
3 14 R 3 14  0 1.94e+00 gag
2  0 R 2 18  0 7.75e+00 ct
2  0 C 2 14  0 7.75e+00 ct
```

Figure 4.4 Representative of the repeat results

Python scripts were also written prior to retrieving the following information.

- Retrieval of the total number of different pattern of repeats.
- Retrieval of repeat lengths whose first positions differs from the second positions and the first position is within the range of the variation sites (position 10)
- Retrieval of nucleotide counts at the variation sites.

4.2.3 Statistical analysis

4.2.3.1 Descriptive analysis

The descriptive analysis of the preprocessed data was done with Excel chart tool and some R scripts. Excel was used for the analysis, representation and presentation of the data using pie and radar charts. R scripts were used for calculations and measurements of central tendencies.

4.2.3.1.1 Measurement of central tendencies

Excel files could not be used for the measurement of central tendencies in this study, because numerous data files were involved. The data were read in using R scripts and command "summary" was invoked. The output were the minimum value, first quartile, median, mean, third quartile and maximum value.

4.2.3.1.2 Charts and graphs

Pie chart is used to compare the proportion of repeats in both datasets in this analysis, while the bar chart was used to show the distribution of the different pattern of repeats.
Box plot was also used to show the five number summary of the datasets, which includes minimum value, first quartile, median, mean, third quartile and maximum value.
Histogram and Q-Q plot was also used in normality check. The normality of a data is considered to be important in this analysis since large data sets are concerned. The results of this analysis usually give a clue on type analysis to be used in making the inference. The repeats of both pathogenic and neutral data sets were analysed using R statistics tools.

The normality of Forward, complemented, reversed and palindromic repeat was checked in both pathogenic and neutral dataset. Histogram was drawn to see if the result will model that of normal distribution. Similarly, the Q-Q plot (quantile-quantile plots) was also constructed to visualize the normality of the data**.** With the Q-Q plots, the normality of the distributions of the repeats in both pathogenic and neutral dataset can be easily compared. Since Q-Q plots compare distributions, there is no need for the values to be observed as pairs or even for the numbers of values in the two groups being compared to be equal.

The radar chart was also used to compare the differences between the distribution in both dataset, which was carried out with Microsoft excel. Radar charts were plotted for the nucleotides of the pathogenic dataset against the corresponding nucleotides of the neutral

dataset for forward, reversed, complemented and palindronmic at position 8, 9 and 10. A radar chart is a graphical method of representing multivariate data in the form of a two-dimensional chart of three or more qualitative variables represented on axes starting from the same point. As shown in figures, each radius or spoke represent the proportion of nucleotides count in the corresponding dataset.

4.2.3.2 Inferential statistics

4.2.3.2.1 Analysis of variance (ANOVA)

ANOVA literally means analysis of variance. As the name implies, it involves partitioning the variance in each variable to test if there is a significance difference between the means of the components of each variable**.**

A one-way ANOVA between the repeats in both datasets was conducted to test the null hypothesis that there is no significant difference between the means of both dataset with the following assumption that:-

- The distribution of genes in human genome is statistically independent in accordance with Mendelian law of independent assortments.
- The data is approximately normally distributed from the result of the QQ plots.

The ANOVA test is applied by calculating the estimates of variance which are:.

i) The variance between pathogenic and neutral dataset which is the mean square between samples and represented with MSB. It is also known as variance between samples.

ii) The variance within each data set i.e within forward, complemented, reversed and palindromic in each data set, which is the mean square between samples and represented with MSW. It is also known as variance within samples.

iii) The value of test statistics F for a test of hypothesis using ANOVA is given by the ratio of two variances, the variance between samples (MSB) and the variance within samples (MSW).

F = MSB/MSW

4.2.3.2.2 Tukey's HSD (Honest Significant Difference) test

Tukey's HSD is a single step multiple comparison procedure and statistical test generally used in conjunction with an ANOVA values that are significantly different. Tukey's HSD test simultaneously examines comparison between all pairs of groups. The adjusted P values of are compared in all the pathogenic and neutrals datasets and using 5% significant difference.

5 Results

5.1 Charts and tables

The data from the repeat analysis were preprocessed to obtain the total in numbers of different patterns of repeats that are present in pathogenic and neutral datasets with a minimum specified unit length of 2 bp. The total number of repeat patterns in each sequence of the dataset has been summarized by the repeat analysis tool, which helped in fast retrieval of results. The occurrences of the repeat pattern were observed to vary in the different data sets. The percentage of the repeats in pathogenic dataset is graphically represented as a pie chart and given in Figure 5.1, while Figure 5.2 shows percentage of repeats in neutral dataset. Table 5.1 is also a raw representation of the total number of repeats in both datasets The histograms shown in Figure 5.3 shows that both the pathogenic and neutral datasets are approximately normally distributed. The Q-Q plots also show that the datasets are normally distributed.

Table 5.1. Total number of repeats in pathogenic and neutral datasets

Pattern of Repeats	Total number of repeats of in the datasets	
	Pathogenic	Neutral
Forward_Repeat	186176	217654
Complemented_Repeat	166795	185598
Reversed_Repeat	310658	363351
Palindromic_Repeat	214279	236945
Total	877908	1003548

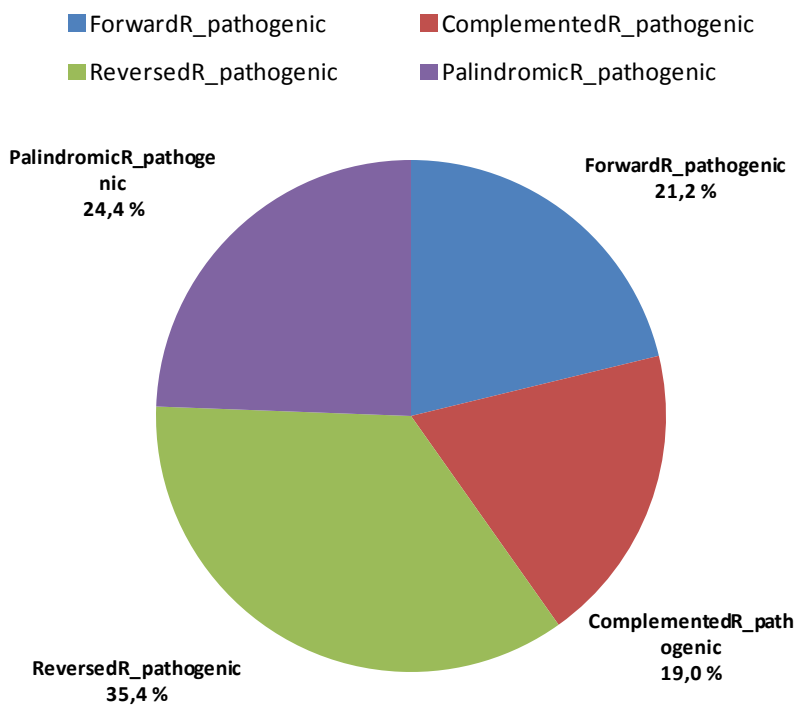


Figure 5.1. Pie chart showing the proportion of forward repeats (forwardR) complemented repeats (complementedR), reversed repeats (reverseR) and palindromic repeats (palindromicR) in pathogenic dataset.

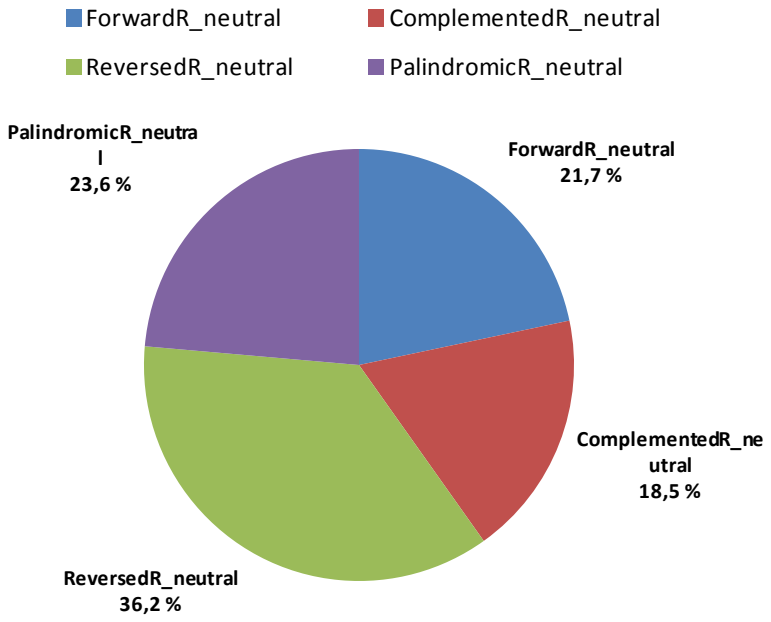


Figure 5.2. Pie chart showing the proportion of forward repeats (fowardR) complemented repeats (complementedR), reversed repeats (reverseR) and palindromic repeats (palindromicR) in neutral dataset.

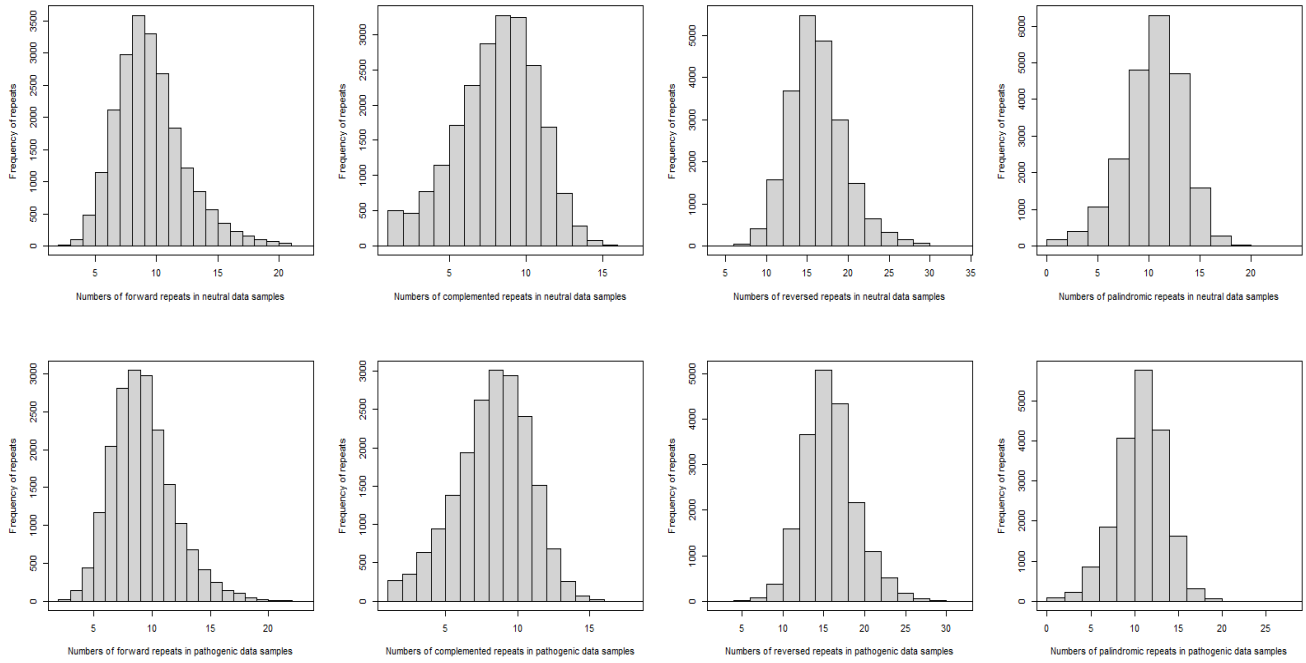


Figure 5.3. Histograms showing normality of the distribution of repeats in both pathogenic and neutral dataset.

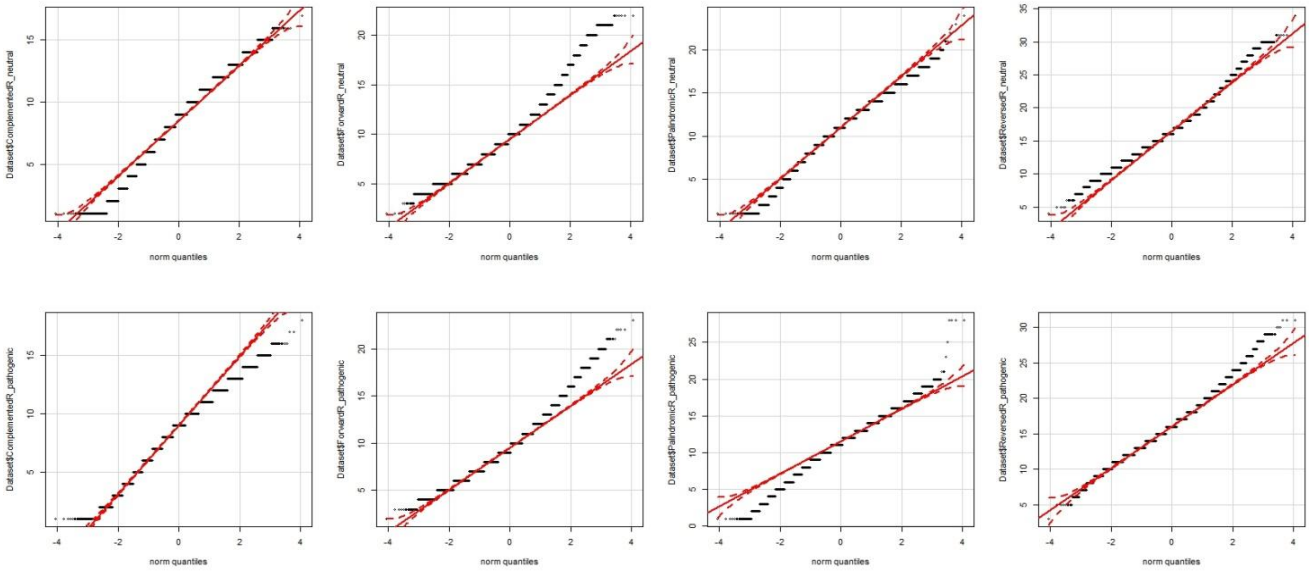


Figure 5.4. Q-Q plots of repeats in both pathogenic and neutral dataset.

5.2 Summary statistics of repeats in pathogenic and neutral dataset

Summary statistics of the repeat types were done using R statistical software. The major concern of this analysis is to observe and report the mean values of different repeats in

both datasets. The result of this analysis shows the mean value of repeats in pathogenic dataset as; complemented repeats (8.75), forward repeats (9.72), reversed repeats (16.22) and palindromic repeat (11.2), while those of the neutral dataset shows mean values as; complemented repeats (8.58), forward repeats (9.99), reversed repeats (16.67) and Palindromic repeat(10.9). Although the numerical comparison of the mean value shows that little differences occur, however this differences could either be significant or may not depending on subsequent analysis to be conducted. In both datasets, it can also be seen that reversed repeat has the highest value of mean and complemented repeats have the lowest values. Other summary results are given in Table 5.2 and 5.3.

The box plot in Figure 5.5 is a representation of the five number summaries (the smallest observation (sample minimum), lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation (sample maximum) of the datasets. The dots represent the outliers. The result of the box plot indicates that there are significant differences between the pathogenic and neutral datasets using their mean values.

The result of the box plot in figure 5.5 indicates that there are significant differences between the pathogenic and neutral datasets using their mean values.

Table 5.2. Summary statistics of repeats in pathogenic dataset

	Forward	Complemented	Reverse	Palindromic
Minimum	2.00	1.00	3.00	1.00
1st quarter	8.00	7.00	14.00	10.00
Median	9.00	9.00	16.00	11.00
Mean	9.72	8.75	16.22	11.20
3rd quarter	11.00	11.00	18.00	13.00
Maximum	23.00	18.00	31.00	28.00
NA'S	2646.00	2731.00	2651.00	2662.00

*NA's ; Not available was used to make all data equal

Table 5.3. Summary statistics of repeats in neutral datasets

	Forward	Complemented	Reverse	Palindromic
Minimum	2.00	1.00	4.00	1.00
1st quarter	8.00	7.00	14.00	9.00
Median	10.00	9.00	16.00	11.00
Mean	9.99	8.58	16.67	10.90
3rd quarter	11.00	10.00	19.00	13.00
Maximum	22.00	17.00	34.00	24.00
NA's	-	172.00	-	64.00

*NA's ; Not available was used to make all data equal

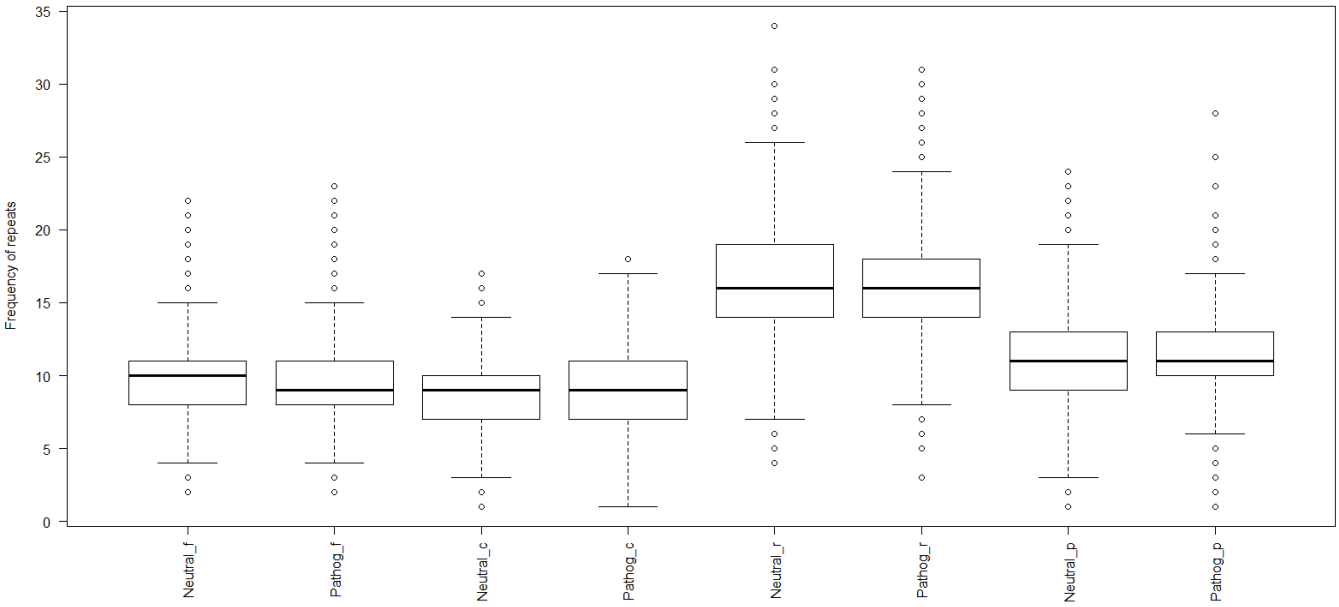


Figure 5.5. Box plot of forward repeats (neutral_f, pathog_f), complemented repeats (neutral_c, pathog_c), reversed repeats (neutral_r, pathog_r) and palindromic repeats (neutral_p, pathog_p) in neutral and pathogenic datasets.

5.3 Distribution of repeats with various lengths

Different repeats of varying lengths were observed to be present within the various pattern of repeats analyzed in both datasets. Just as specified during the analysis, the minimum repeat length sorted was two (2 bp) and a maximum length of ten (8 bp) nucleotide bases

along the 21bp. However, major focus was on the repeats of varying lengths which starts from the variation site. It was observed that the number of the repeats of varying length reduces as the length of the repeat increases. However, repeat lengths of 2 bp were the most abundant repeats found in both datasets. The Table 5.4, Table 5.5 and Table 5.6 and Table 5.7 below shows the raw and normalized distribution and of the repeats of various length from the variation site in pathogenic and neutral datasets respectively.

Table 5.4. Distribution of repeats of length 2-8 bp starting at variation site of the 21 bp nucleotide sequences in pathogenic dataset.

Pattern of Repeats	Distribution of repeats of 2-8 bp unit length from variation positions in pathogenic dataset						
	2 bp	3 bp	4 bp	5 bp	6 bp	7 bp	8 bp
Forward_Repeat	48088	11920	3336	888	256	32	32
Complemented_Repeat	44392	10920	2568	664	48	8	16
Reversed_Repeat	33432	7088	1416	0	0	0	0
Palindromic_Repeat	36920	8952	1568	232	0	0	0

Table 5.5. percentage of distribution of repeats of length 2-8 bp starting at variation site of the 21 bp nucleotide sequences in pathogenic dataset.

Pattern of Repeats	Distribution of repeats of 2-8 bp unit length from variation positions in pathogenic dataset						
	2 bp	3 bp	4 bp	5 bp	6 bp	7 bp	8 bp
Forward_Repeat	75.73	18.63	4.38	1.13	0.08	0.01	0.027
Complemented_Repeat	79.72	16.9	3.38	0	0	0	0
Reversed_Repeat	77.45	18.87	3.29	0.49	0	0	0
Palindromic_Repeat	74.49	18.47	5.17	1.37	0.39	0.05	0.05

Table 5.6. Distribution of repeats of length 2-8 bp starting at variation site of the 21 bp nucleotide sequences in neutral dataset

Pattern of Repeats	Distribution of repeats of 2-8 bp unit length from variation positions in neutral dataset						
	2 bp	3 bp	4 bp	5 bp	6 bp	7 bp	8 bp
Forward_Repeat	52272	15752	4120	1272	352	72	56
Complemented_Repeat	50056	12824	2144	642	152	0	32
Reversed_Repeat	36312	8152	1240	0	0	0	0
Palindromic_Repeat	40168	9512	1472	224	0	0	0

Table 5.7. Percentage of distribution of repeats of length 2-8 bp starting at variation site of the 21 bp nucleotide sequences in neutral dataset.

Pattern of Repeats	Distribution of repeats of 2-8 bp unit length from variation positions in neutral dataset						
	2 bp	3 bp	4 bp	5 bp	6 bp	7 bp	8 bp
Forward_Repeat	29.23	34.07	45.90	59.49	69.84	100.00	63.64
Complemented_Repeat	27.99	27.73	23.89	30.03	30.16	0	36.37
Reversed_Repeat	20.31	17.63	13.81	0	0	0	0
Palindromic_Repeat	22.46	20.57	16.40	10.48	0	0	0

5.4 Nucleotide base counts of repeats in pathogenic and neutral dataset.

Result of analysis done to carry out nucleotide base counts in both datasets, shows that neutral dataset has increased numbers of the four nucleotide bases (ACGT) when compared with pathogenic dataset. This count may not reflect true differences, because the sample size difference could have some effect on the total counts. However, below is

Table 5.8, Table 5.9 and Table 5.10, Table 5.11 showing the raw and normalized counts respectively of the nucleotides bases in pathogenic and neutral datasets.

Table 5.8. Nucleotide base count present at the variation sites in pathogenic dataset.

Pattern of Repeats	Count of nucleotide bases in repeats found at position 10 in both pathogenic dataset			
	Nucleotide bases			
	Adenine	Guanine	Cytosine	Thymine
Forward_Repeat	26472	45216	47488	32560
Complemented_Repeat	26256	43696	39784	25872
Reversed_Repeat	16200	28200	30296	19096
Palindromic_Repeat	21776	34760	30280	21312
Total	90704	151872	147848	98840

Table 5.9. Percentage of nucleotide base count present at the variation sites in pathogenic dataset

Pattern of Repeats	Count of nucleotide bases in repeats found at position 10 in both pathogenic dataset			
	Nucleotide bases			
	Adenine	Guanine	Cytosine	Thymine
Forward_Repeat	29.19	29.77	32.12	32.94
Complemented_Repeat	28.97	28.77	26.91	26.18
Reversed_Repeat	17.86	18.57	20.49	19.32
Palindromic_Repeat	24.01	22.89	20.48	21.56

Table 5.10. Nucleotide base count present at the variation sites in neutral dataset

Pattern of Repeats	Count of nucleotide bases in repeats found at position 10 in both neutral dataset			
	Nucleotide bases			
	Adenine	Guanine	Cytosine	Thymine
Forward_Repeat	38760	49408	53704	35832
Complemented_Repeat	30112	42288	45752	33296
Reversed_Repeat	23416	27976	30192	20456
Palindromic_Repeat	23128	31368	35576	25808
Total	115416	151040	165224	115392

Table 5.11. Percentage of nucleotide base present at the variation sites in neutral dataset

Pattern of Repeats	Count of nucleotide bases in repeats found at position 10 in both neutral dataset			
	Nucleotide bases			
	Adenine	Guanine	Cytosine	Thymine
Forward_Repeat	33.58	32.71	32.50	31.05
Complemented_Repeat	26.09	27.99	27.69	28.85
Reversed_Repeat	20.29	18.52	18.27	17.72
Palindromic_Repeat	20.04	20.77	21.53	22.36

Radar charts

The Figure 5.6 to Figures 5.21 below shows the radar charts which was plotted with the proportion of the nucleotide counts of (A,G,C,T) in each types of repeats in the pathogenic and neutral dataset. The differences in the size of spokes show differences in the proportions as shown below.

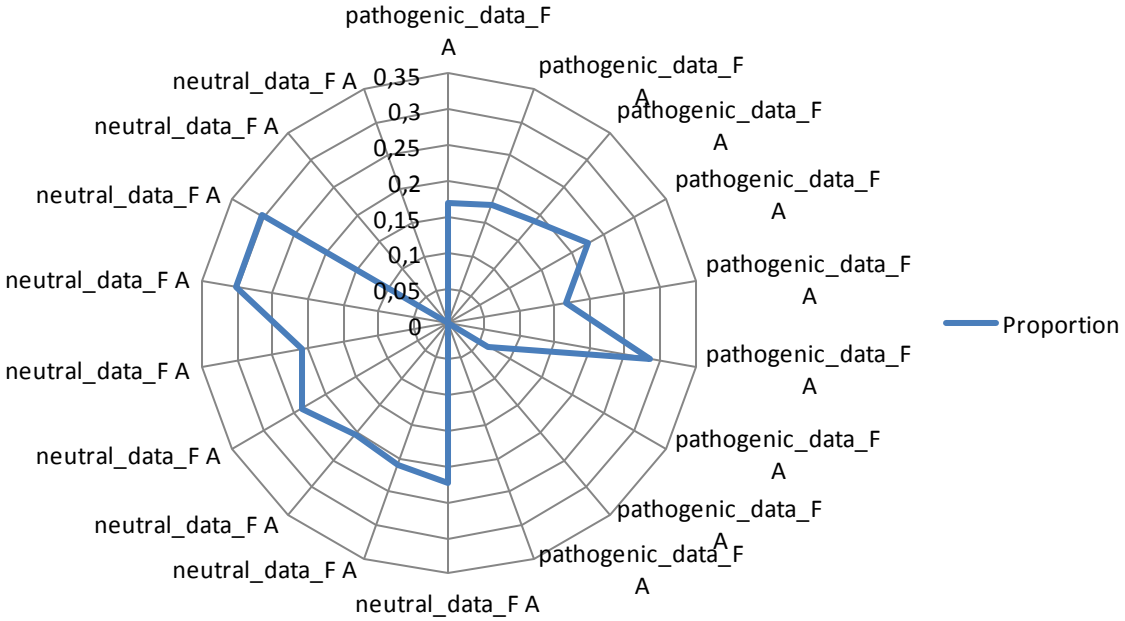


Figure 5.6. Adenine plot in pathogenic (pathogenic_data_FA) against neutral (neutral_data_FA) forward repeat from position 10 in both datasets

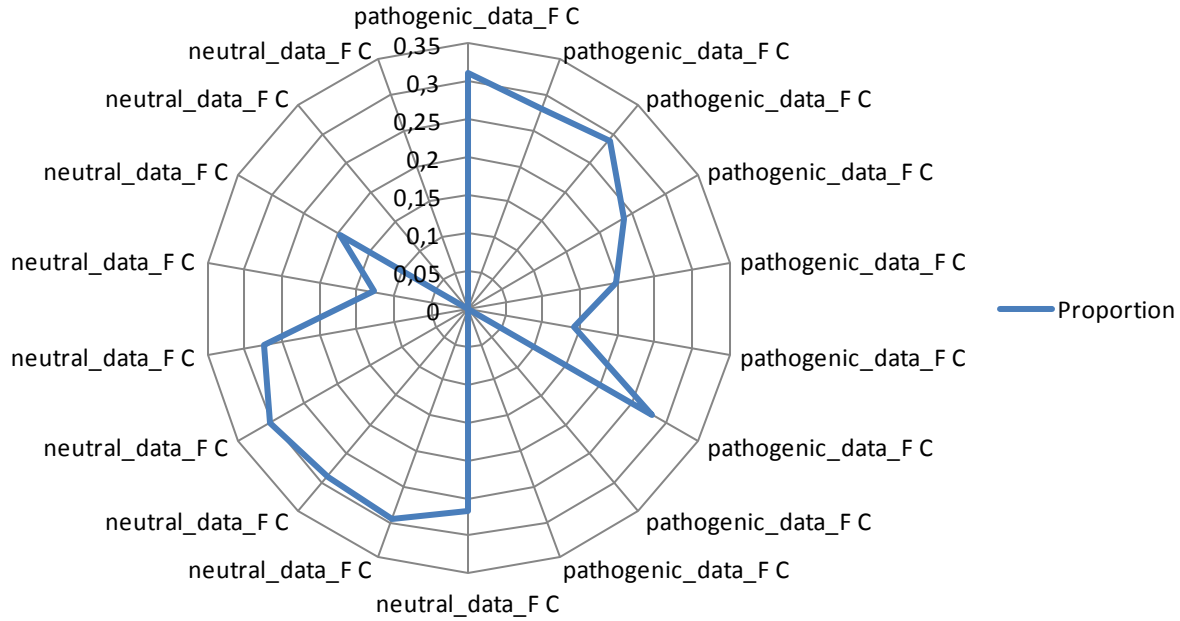


Figure 5.7. Cytosine plot in pathogenic (pathogenic_data_FC) against neutral (neutral_data_FC) forward repeat from position 10 in both datasets

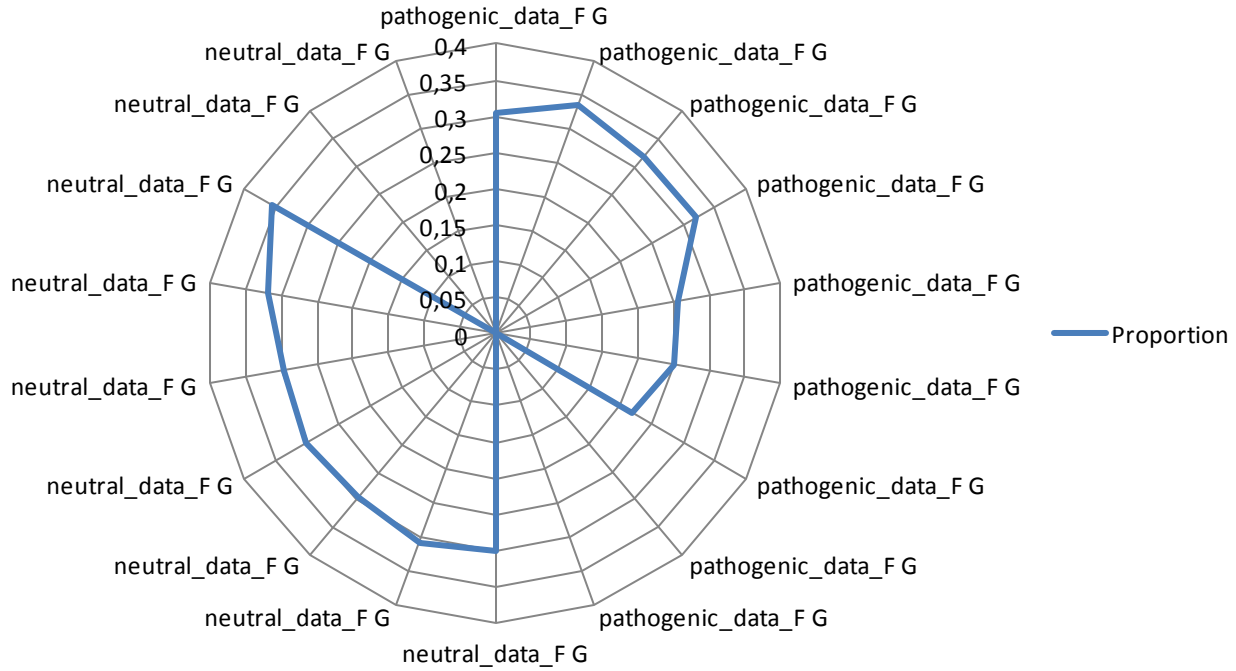


Figure 5.8. Guanine plot in pathogenic (pathogenic_data_FG) against neutral (neutral_data_FG) forward repeat from position 10 in both datasets

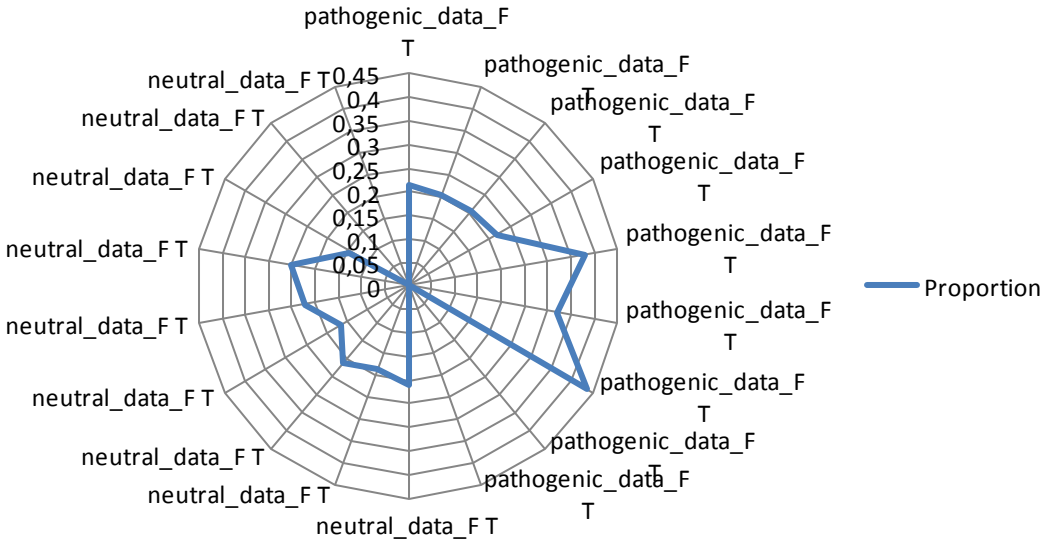


Figure 5.9. Thymine plot in pathogenic (pathogenic_data_FT) against neutral (neutral_data_FT) forward repeat from position 10 in both dataset

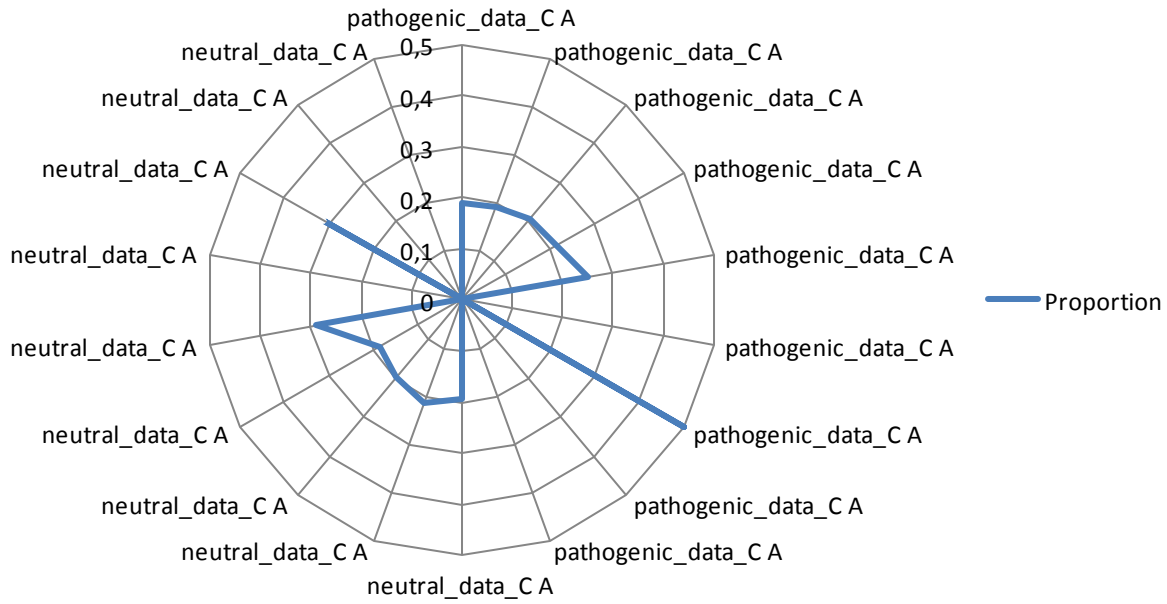


Figure 5.10. Adenine plot in pathogenic (pathogenic_data_CA) against neutral (neutral_data_CA) complemented repeat from position 10 in both datasets

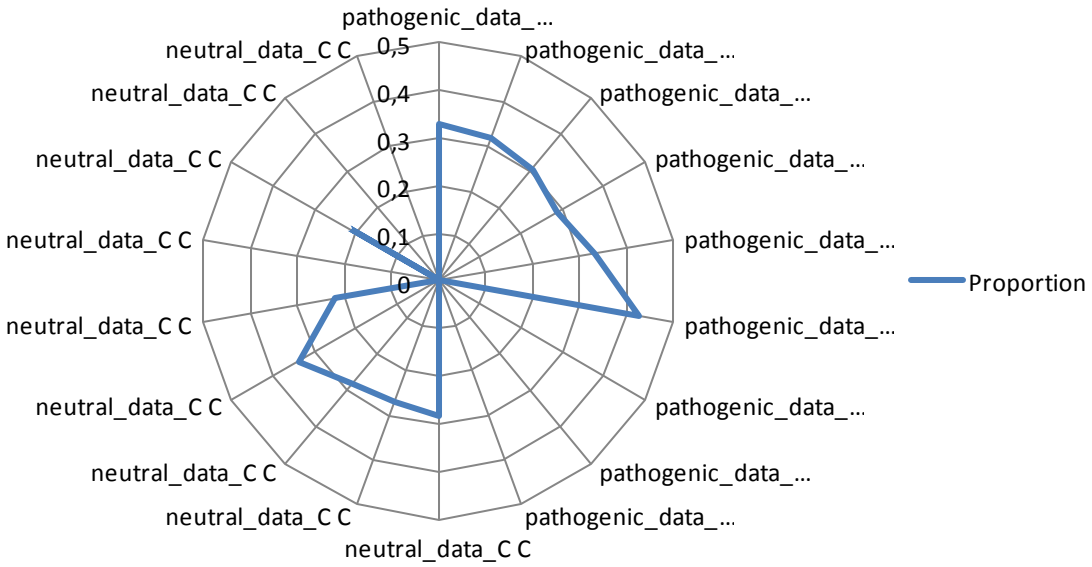


Figure 5.11. Cytosine plot in pathogenic (pathogenic_data_CC) against neutral (neutral_data_CC) complemented repeat from position 10 in both datasets

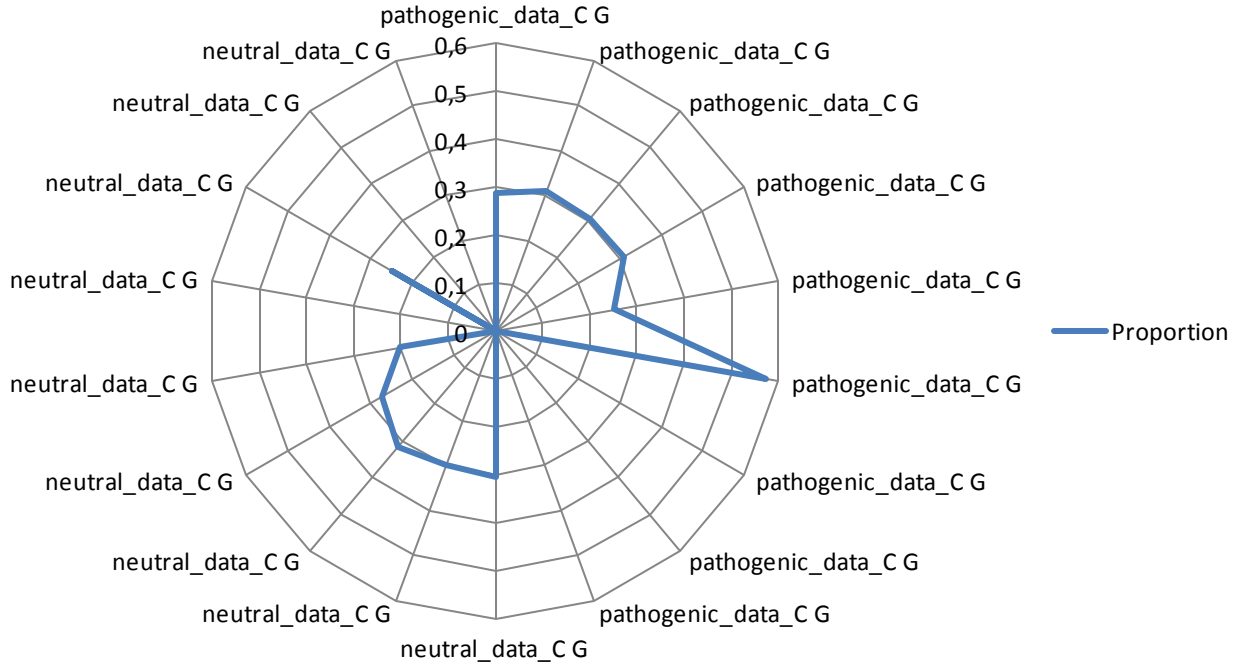


Figure 5.12. Guanine plot in pathogenic (pathogenic_data_CC) against neutral (neutral_data_CC) complemented repeat from position 10 in both datasets

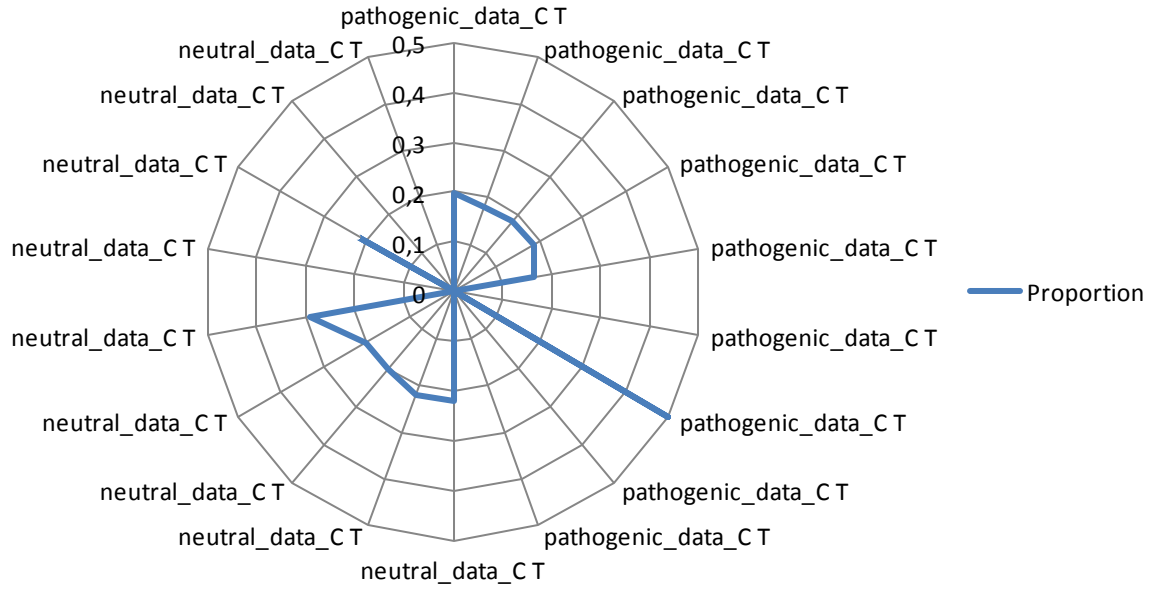


Figure 5.13. Thymine plot in pathogenic (pathogenic_data_CT) against neutral (neutral_data_CT) complemented repeat from position 10 in both datasets

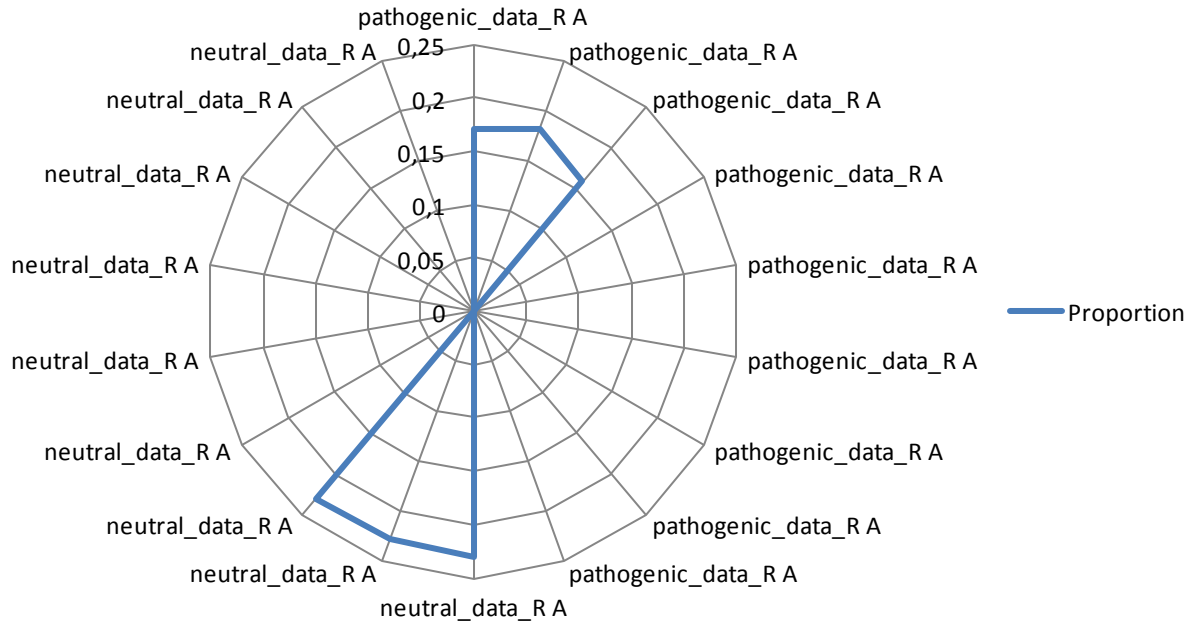


Figure 5.14. Adenine plot in pathogenic (pathogenic_data_RA) against neutral (neutral_data_RA) reversed repeat from position 10 in both datasets

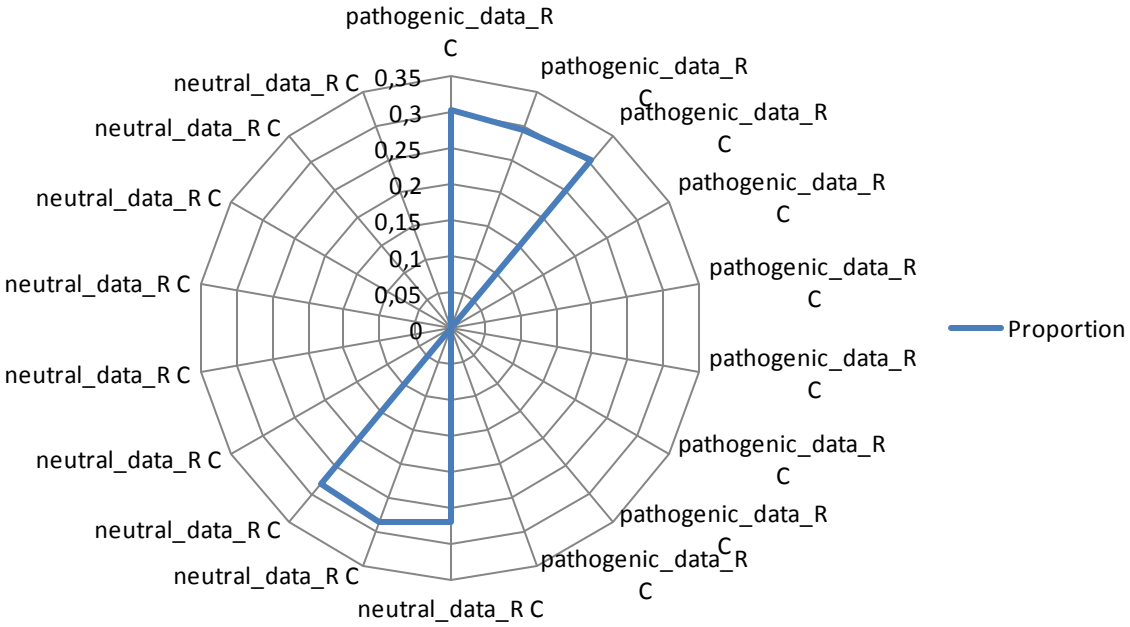


Figure 5.15. Cytosine plot in pathogenic (pathogenic_data_RC) against neutral (neutral_data_RC) reversed repeat from position 10 in both datasets

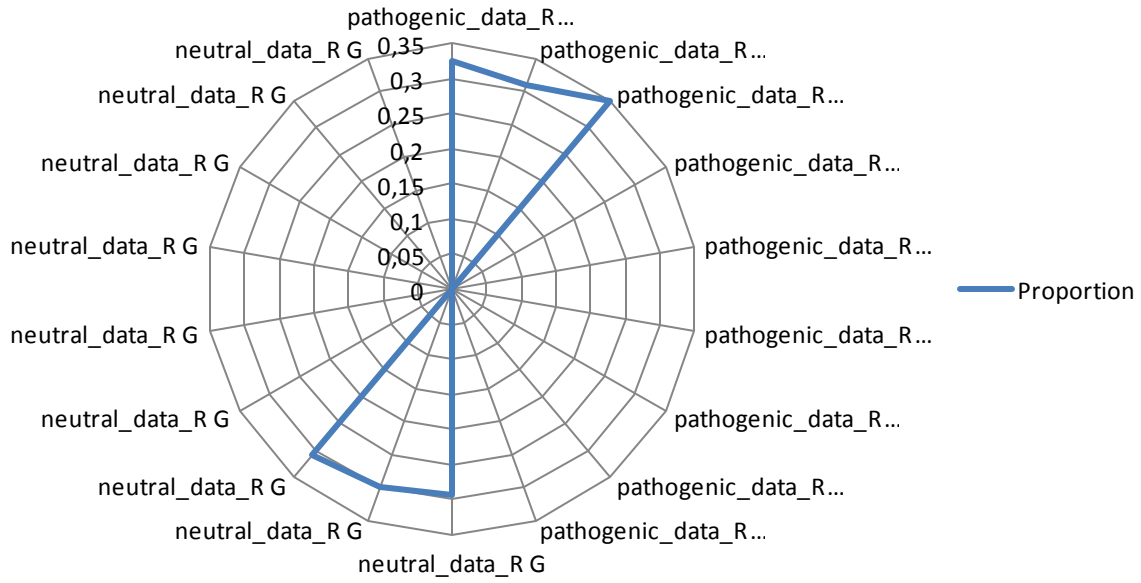


Figure 5.16. Guanine plot in pathogenic (pathogenic_data_RG) against neutral (neutral_data_RG) reversed repeat from position 10 in both datasets

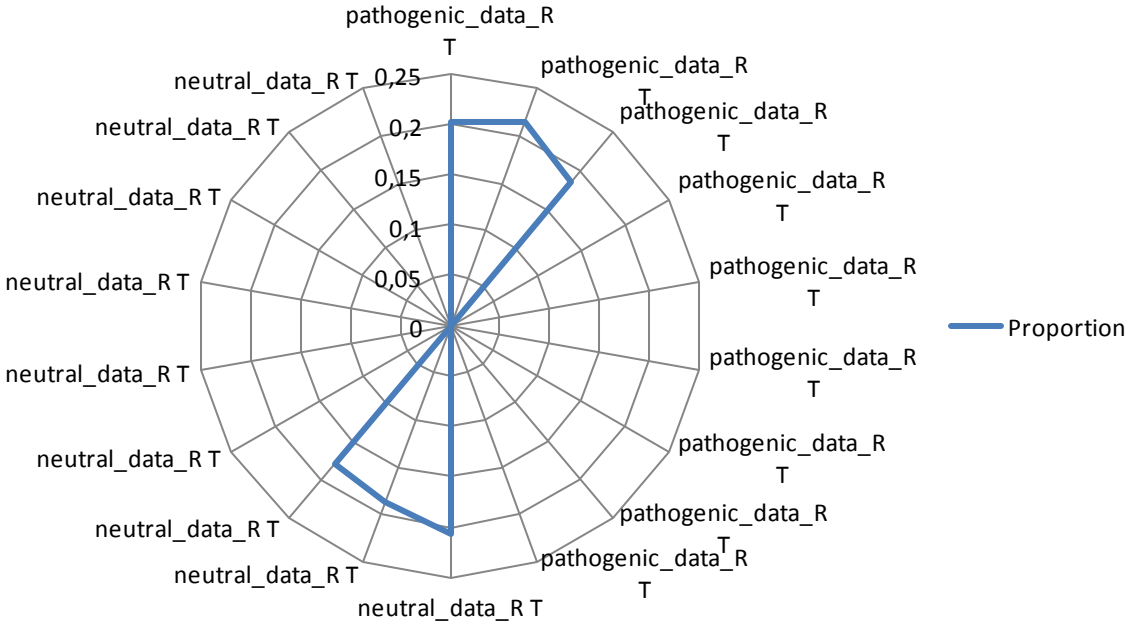


Figure 5.17. Thymine plot in pathogenic (pathogenic_data_RT) against neutral (neutral_data_RT) reversed repeat from position 10 in both datasets.

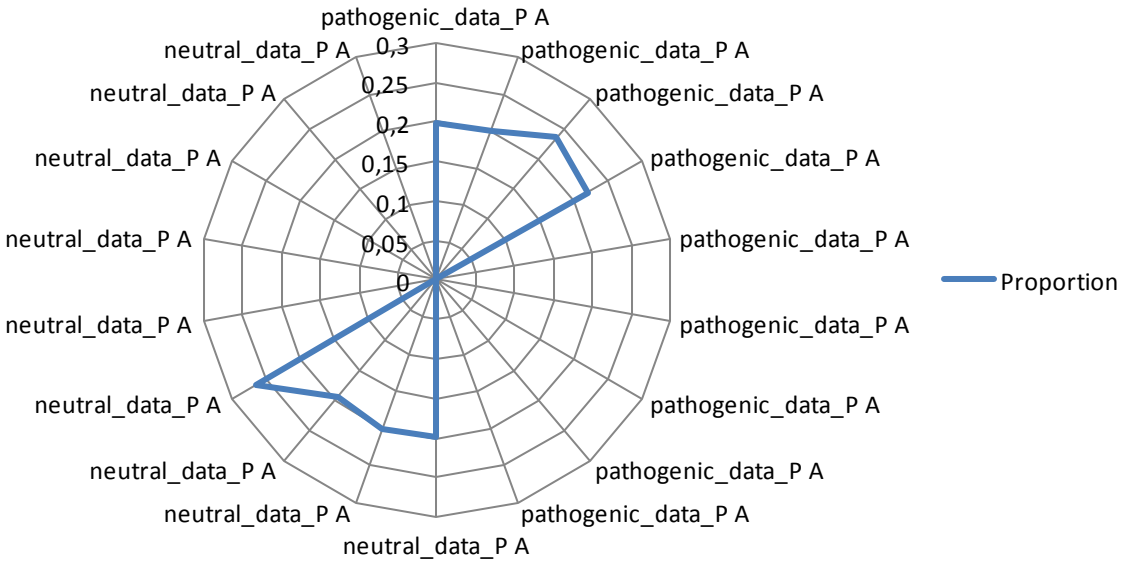


Figure 5.18. Adenine plot in pathogenic (pathogenic_data_PA) against neutral (neutral_data_PA) palindromic repeat from position 10 in both datasets

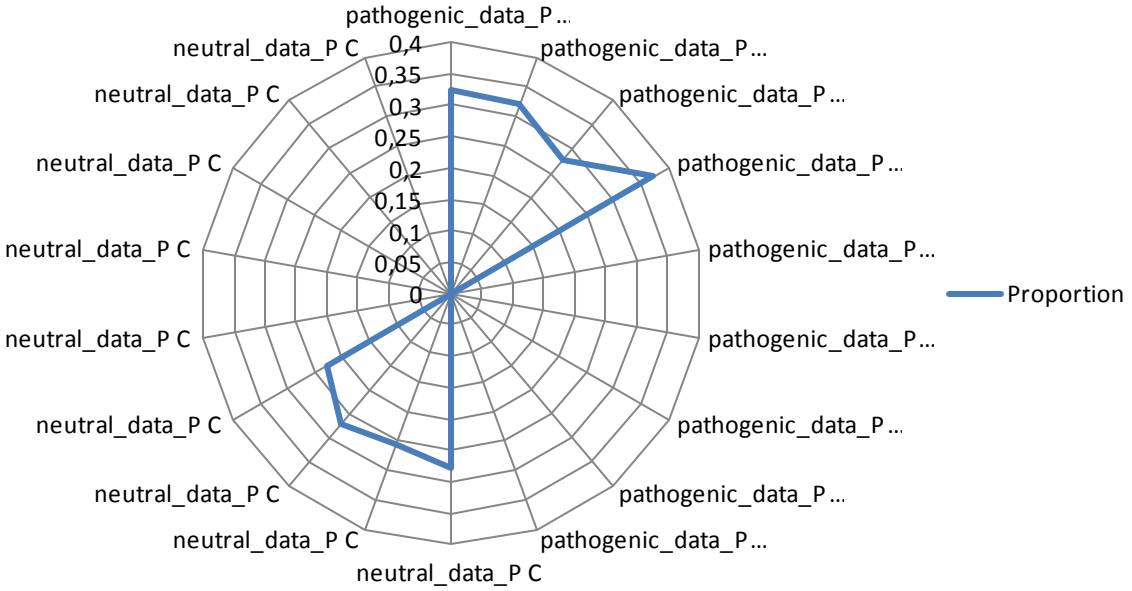


Figure 5.19. cytosine plot in pathogenic (pathogenic_data_PC) against neutral (neutral_data_PC) palindromic repeat from position 10 in both datasets

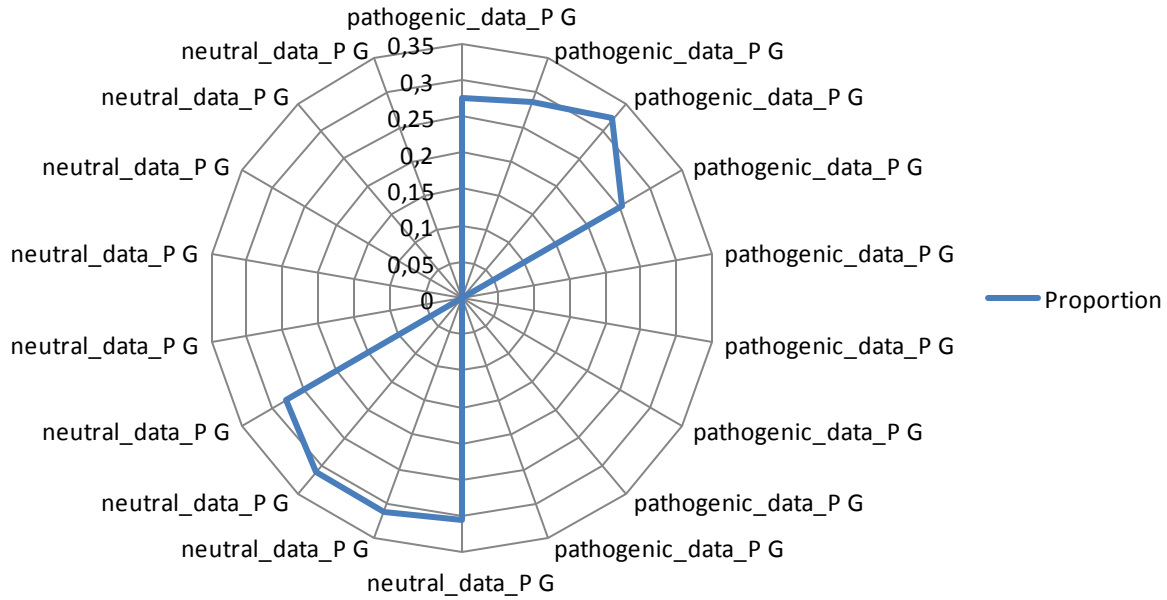


Figure 5.20. Guanine plot in pathogenic (pathogenic_data_PG) against neutral (neutral_data_PG) palindromic repeat from position 10 in both datasets

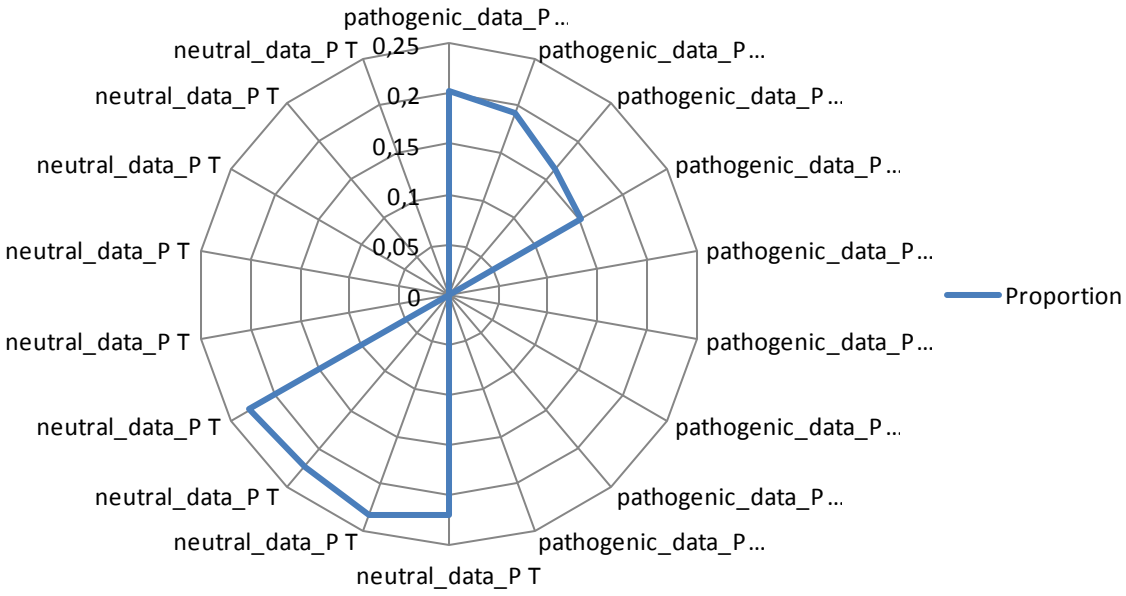


Figure 5.21. Thymine plot in pathogenic (pathogenic_data_PT) against neutral (neutral_data_PT) reversed repeat from position 10 in both datasets.

5.5 ANOVA Result

ANOVA analysis using 5% significance give the result in Table 5.12 below.

Table 5.12. ANOVA result

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Between Repeats	7	0.635	0.09077	5.265	$1.15e^{-05}$
Within Repeats	280	4.827	0.01724		

*Df represents degree of freedom, Sum Sq; represents sum of squares, Mean Sq; is mean of squares.

From the above result, it can be deduced that there is significant difference [F(7,280)= 5.265, $p=1.15e^{-05}$] . This gave room for further analysis in order to confirm where the differences actually occur.

5.6 Tukey's HSD (Honest Significant Difference) Test Results

The Tukey's HSD result shows that significant differences occur between the following pairs with p-values of 0.0092 which are also highlighted in Table 5.13.

- The repeat pattern reversed and forward type in neutral (neutral_R-neutral_F) dataset
- Reversed repeat in pathogenic dataset and forward repeats in neutral (pathogenic_R-neutral_F) dataset
- Complemented repeat in pathogenic dataset and reversed repeats in neutral (pathogenic_C-neutral_R) dataset
- Forward repeat in pathogenic dataset and reverse repeats in neutral (pathogenic_F-neutral_R) dataset
- Reversed repeat in pathogenic dataset and complemented repeats in neutral (pathogenic_R-pathogenic_C) dataset and

- Reversed repeat in pathogenic dataset and forward repeats in neutral (pathogenic_R-pathogenic_F) dataset.

However, the result does not indicate that similar pair of repeat pattern from pathogenic and neutral dataset has any significant differences. Other results of the Tukey's HSD analysis can be seen in Table 5.13 below.

Table 5.13. Tukey's HSD result

	diff	lwr	upr	p adj
neutral_F-neutral_C	$2.78e^{-02}$	-0.0667	0.1223	0.9861
neutral_P-neutral_C	$-5.56e^{-02}$	-0.1501	0.0390	0.6239
neutral_R-neutral_C	$-8.33e^{-02}$	-0.1778	0.0112	0.1291
pathogenic_C-neutral_C	$2.78e^{-02}$	-0.0667	0.1223	0.9861
pathogenic_F-neutral_C	$2.78e^{-02}$	0.0667	0.1223	0.9861
pathogenic_P-neutral_C	$-5.56e^{-02}$	-0.1501	0.0390	0.6239
pathogenic_R-neutral_C	$-8.33e^{-02}$	-0.1778	0.0112	0.1291
neutral_P-neutral_F	$-8.33e^{-02}$	-0.1778	0.0112	0.1291
neutral_R-neutral_F	$-1.11e^{-01}$	-0.2056	-0.0166	0.0092
pathogenic_C-neutral_F	$2.78e^{-09}$	-0.0945	0.0945	1.0000
pathogenic_F-neutral_F	$-8.33e^{-09}$	-0.0945	0.0945	1.0000
pathogenic_P-neutral_F	$-8.33e^{-02}$	-0.1778	0.0112	0.1291
pathogenic_R-neutral_F	$-1.11e^{-01}$	-0.2056	-0.0166	0.0092
neutral_R-neutral_P	$-2.78e^{-02}$	-0.1223	0.0667	0.9861
pathogenic_C-neutral_P	$8.33e^{-02}$	-0.0112	0.1778	0.1291
pathogenic_F-neutral_P	$8.33e^{-02}$	-0.0112	0.1778	0.1291
pathogenic_P-neutral_P	$-2.78e^{-09}$	-0.0945	0.0945	1.0000
pathogenic_R-neutral_P	$-2.78e^{-02}$	-0.1223	0.0667	0.9861

Table 5.13. (Continued).

	diff	lwr	upr	p adj
pathogenic_C-neutral_R	$1.11e^{-01}$	0.0166	0.2056	0.0092
pathogenic_F-neutral_R	$1.11e^{-01}$	0.0166	0.2056	0.0092
pathogenic_P-neutral_R	$2.78e^{-02}$	-0.0667	0.1223	0.9861
pathogenic_R-neutral_R	$5.56e^{-09}$	-0.0945	0.0945	1.0000
pathogenic_F-pathogenic_C	$-1.11e^{-08}$	-0.0945	0.0945	1.0000
pathogenic_P-pathogenic_C	$-8.33e^{-02}$	-0.1778	0.0112	0.1291
pathogenic_R-pathogenic_C	$-1.11e^{-01}$	-0.2056	-0.0166	0.0092
pathogenic_P-pathogenic_F	$-8.33e^{-02}$	-0.1778	0.0112	0.1291
pathogenic_R-pathogenic_F	$-1.11e^{-01}$	-0.2056	-0.0166	0.0092
pathogenic_R-pathogenic_P	$-2.78e^{-02}$	-0.1223	0.0667	0.9861
pathogenic_C-neutral_F	$2.78e^{-09}$	-0.0945	0.0945	1.0000

* diff : difference in observed means, lwr : the lower end point of interval, upr : upper end point of interval, p adj : gives the p-value after adjusting the multiple comparisons

6 Discussion

Several computational methods, tools and approaches have been devised towards the analysis of repetitive DNA sequences in both prokaryotic and eukaryotic genomic sequences. Some of these tools, methods and approaches have been briefly highlighted in this study. The fascinating thing about majority of the repetitive DNA sequence analysis tools or method is that they have several approaches which they use in conducting their

identification of dispersed repetitive DNA sequences in the genome, while some others use ab initio methods of repeat identification approaches.

The reference based identification approach has been widely studied and shown to work by making matching comparison between a given data set and a reference dataset from the database (Saha et al., 2008a; Lerat, 2009). However, a foreseen shortcoming of this kind of approach is that undiscovered repeat patterns or undocumented repetitive DNA sequences will not be available for such kind of comparison to be established.

Ab initio methods of repeat identification (Saha et al., 2008a; Lerat, 2009) are however different in it's repetitive DNA sequence identification. These have been studied and shown to build it repetitive DNA sequence from the scratch based on some specified parameters. A good example of such tool is REPuter.

REPuter was found to be a suitable tool for this study, due to its efficiency, flexibility and significance, interactive visualization and compositionality. There are three different analysis tools. Each tool has its own purpose in the identification and analysis of repeats. The three programs include REPfind, which was used in calculating the repeats, REPselect for selecting results and REPvis for visualization of the repeats. REPfind was the only program used for this analysis. Since all the results were meant to be used, REPselect was not used and the visualization of each sample in the data would generate a large number of data files that are not required for any further analysis.

REPfind uses some commands for selecting the repeats the largest default E-value of 50. These commands are important in order to output .

The results derived from the repeat analysis, using REPfind were further processed. These results were used for descriptive and inferential analysis from where suggestions and conclusions could be reached based on the given datasets.

From the pie chart in Figure 5.1 and Figure 5.2, it was observed that reversed repeats in both the pathogenic and neutral dataset were the highest followed by palindromic repeat and the least are the complementary repeat. The proportion of each component in both the pathogenic and neutral repeats is approximately the same. However, the numbers of repeats in each of the components of the neutral dataset are greater than the corresponding pathogenic dataset.

In the process of testing for the normality of the datasets, histogram gave approximately normal distribution. The QQ plot also gave a clearer picture of the normality and support that the data were approximately normally distributed.

The result of the analysis of the nucleotide base counts in both datasets shows that neutral dataset has increased numbers of the four nucleotide bases (ACGT) when compared with pathogenic dataset. This count may not reflect true differences, because the sample size difference could have some effect on the total counts. The boxplot gave a clearer picture between the means of repeats in neutral dataset and pathogenic dataset. The mean of complemented repeat in neutral dataset (neutral_C) is lower than the mean of complemented repeats in pathogenic dataset (pathogenic_C), the mean of forward repeats in neutral dataset (Neutral_F). is approximately equal to the mean of forward repeats in pathogenic dataset (Pathogenic_F). Likewise, the mean of palindromic repeats in neutral dataset (Neutral_p) is almost the same as the mean of palindromic repeats in pathogenic dataset (pathogenic_p) and similarly the mean of reversed repeats in neutral dataset (neutral _R) and the mean of reversed repeats in pathogenic dataset (Pathogenic_R) are approximately the same.

In the inferential analysis, ANOVA gave the following results [$F(7,280)= 5.265$, $p=1.15e-^{05}$] at 5% significance. At $p > 0.05$, the results show that there is a significant difference between the repeat patterns in neutral and pathogenic dataset. The results of the Tukey's HSD test gave an elaborate picture of exactly where the significant difference occurred in all the data set. This test is otherwise known as 'Honest significant difference' and it indicated that there is significant difference between:-

- the repeats pattern in reversed and forward type in neutral dataset;
- reversed repeat in pathogenic dataset and forward repeats in neutral dataset;

- complemented repeat in pathogenic dataset and reversed repeats in neutral dataset;
- forward repeat in pathogenic dataset and reverse repeats in neutral dataset,
- reversed repeat in pathogenic dataset and complemented repeats in neutral dataset and
- reversed repeat in pathogenic dataset and forward repeats in neutral dataset.

7 Conclusion

The result of the Anova [F(7,280)= 5.265, p=1.15e-05] indicates rejects the null hypothesis that there is no significant difference between the repeats in neutral and pathogenic dataset and therefore we accept the alternative hypothesis that, there is significant different between the repeats in neutral and pathogenic dataset.

The result of Tukey's HSD of p adjusted of 0.0092 for P> 0.05 of reverse against forward repeats in neutral dataset (neutral_R- neutral_F) rejects the null hypothesis that: there is no significant difference within the repeats of neutral dataset and therefore accept the alternative hypothesis that: there is significant difference within the repeats of neutral dataset.

Similarly, Tukey's HSD of p adjusted of 0.0092 for P> 0.05 of reverse against complemented repeats in pathogenic dataset (pathogenic_R- pathogenic_C) and reverse against forward repeats in pathogenic dataset (pathogenic_R- pathogenic_F) reject the null hypothesis that, there is no significant difference within the repeats of the pathogenic dataset and therefore accept the alternative hypothesis that, there is significant difference within the repeats of the pathogenic dataset.

Thus some differences occur between the two datasets, even though there are no exact matching pair of repeat pattern from both datasets.

With these findings, further studies could focus on specific unit length of nucleotide repeat patterns in close proximity to variation site, in pathogenic dataset to see if there is correlation between these repeating unit pattern and pathogenicity in the data.

8 References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol, 215(3): 403-10*

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res, 25: 3389–3402*

Amigo J, Salas A, Philips C. (2011). ENGINES: exploring single nucleotide variation in entire human genomes. *BMC Bioinformatics, 12: 105*

Andrieu O, Fiston AS, Anxolabehere D et al., (2004). Detection of transposable elements by their compositional bias. *BMC Bioinformatics, 5: 94*

Areshchenkova T, Ganal MW. (1999). Long tomato microsatellites are predominantly associated with centromeric regions. *Genome, 42: 536-544.*

Ashikawa I, Kurata N, Saji S, Umehara Y, Sasaki T. (1999). Application of restriction fragment fingerprinting with a rice microsatellite sequence to assembling rice YAC clones. *Genome, 42: 330-337*.

Bao Z, Eddy SR. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res, 12: 1269–1276*

Batzer MA, Deininger PL . (2002). ALU Repeats and Human genomic Diversity.
Nature Genetics, 3: 370-379

Bayliss CD, Dixon KM, Moxon ER. (2003). Simple sequence repeats (microsatellites) : mutational mechanisms and contributions to bacterial pathogenesis. A meeting review. FEMS, *Immunology & Medical Microbiology, 40: 11-19*

Belkum AV, Scherer S, Alphen LV, Verbrugh H. (1998). Short-Sequences DNA Repeats in Prokaryotic Genomes. *Microbio.Mol.Biol, 62(2): 275-293*

Benson G. (1998). Tandem repeats finder: a program to analyse DNA sequences. *Nucleic Acids Res, 27(2): 573-580*

Benson G, Waterman MS. (1998). A method for fast database search for all K- nucleotide repeats. *Nucleic Acids Res, 22(22): 4828-4836*

Britten RJ, Kohne DE. (1968). Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science, 161: 529–540.*

Brandes A, Thompson H, Dean C, Heslop-Harrison JS (1997). Multiple repetitive DNA sequences in the paracentromeric regions of *Arabidosis thaliana* L. *Chromosome Research, 5: 238-246.*

Brown TA. (2002). Genomes, 2nd edition. Oxford: Wiley-Liss; 2002. *The Human Genome.*

Cambareri EB, Aisner R, Carbon J. (1998). Structure of the chromosome VII centromere region in *Neurospora crassa*: degenerate transposons and simple repeats. *Molecular and Cellular Biology*, *18: 5465-5477.*

Campagna D, Romualdi C, Vitulo N et al., (2005). RAP: a new computer program for de novo identification of repeated sequences in whole genomes. *Bioinformatics, 21: 582–588*

Centola M, Carbon J. (1994). Cloning and characterization of centromereic DNA from *Neurospora crassa*. *Molecular and Cellular Biology*, *14: 1510 -1519.*

Charlesworth B, Sniegowski P, Stephan W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature, 371: 215–220*.

Chepelev I, Wei G, Tang Q, Zhao K. (2009).Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq *Nucleic Acids Res, 37 (16)e106*

DePristo MA, Banks E, Poplin R, Gerimella KV, Maguire JR, Hartl C, Philippakis AA, Angel GD Rivas MA, Hanna M, Mckenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabiel SB, Altshuler D , Daly MJ (2011)A framework for variation discovery and genotyping using next-generation DNA sequencing. *Nature Genetics, 43: 491–498*

Dereeper A, Nicolas S, Cunff LL, Bacilieri R, Doligez A, Peros J-P, Ruiz M, , This P. (2011). SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC Bioinformatics, 12: 134*

Diwan N, Cregan PB (1997). Automated sizing of fluorescent-labeled simple sequence repeat (SSR) markers to assay genetic variation in soybean. *Theor. Appl. Genet, 95: 723-733.*

Edgar RC, Myers EW (2005). PILER: identification and classification of genomic repeats. *Bioinformatics, 21: 152–158.*

Edgar RC. (2006). PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics, 8: 18*

Elder JF, Tuner JB. (1995). Concerted Evolution of Repetitive DNA Sequences in Eukaryotes The Quarterly Review of Biology, 70 (3): 297-320

Ellegren H. (2004). Microsatellites: Simple Sequences with Complex Evolution. *Nature Rev, 5: 437-445*

Fredman D, Siegfried M, Yuan YP, Bork P, Lehväslaiho H, , Brookes AJ (2002). HGVbase: a human sequence variation database emphasizing data quality and broad spectrum of data sources. *Nucleic Acids Res, 30(1): 387-91*

Fu YH, Kuhl DP, Pizzuti A, Pieretti M, Sutcliffe JS, Richards S, Verkerk AJ, Holden JJ, Fenwick RG, Warren ST, Oostra BA, Nelson DL, Caskey CT (1991). Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell, 67: 1047-058.*

Gangwal K, Lessnick SL. (2008). Microsatellites are EWS/FLI response elements: genomic "junk" is EWS/FLI's treasure. *Cell Cycle, 7(20): 3127-32*

Genetic Science Learning Center (1969, December 31) Making SNPs Make Sense. *Learn.Genetics*. Retrieved April 20, 2012, from http://learn.genetics.utah.edu/content/health/pharma/snips/

Gendrel CG, Boulet A, Dutreix M. (2000). (CA/GT) microsatellites affect homologous recombination during yeast meiosis. *Genes & Development, 14: 1261-1268*.

Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenski J, Sang Y, Elnitski L, Cutting G, Trumbower H, Kern A, Kuhn R, Patrinos GP, Hughes J, Higgs D, Chui D, Scriver C, Phommarinh M, Patnaik S, Blumenfeld O, Gottlieb B, Vihinen M, Väliaho J,Kent J, Miller W, , Hardison RC (2007). PhenCode: Connecting ENCODE Data With Mutation and Phenotype*. Human Mutation, 28(6): 554-562*

Gleicher N, Weghofer A, Oktay K, Barad DH. (2009). Correlation of triple repeats on the FMR1 (fragile X) gene to ovarian reserve: a new infertility test? *Acta Obstet Gynecol Scand., 88(9): 1024-30*

Gulcher J. (2012). Microsatellite markers for linkage and association studies. *Cold Spring Harb Protoc, (4): 425-32.*

Guo Y, Jamison DC. (2005). The distribution of SNPs in human regulatory regions. *BMC Genomics, 6: 140*

Havecker ER, Gao X, Voytas DF. (2004). The diversity of LTR retrotransposons. *Genome Biol, 5: 225*

Ho JW, Choi S-c, Lee Y-f, Hui TC, Cherny SS, Grcia-Barcelo M-M, Carvajal-Carmona L, Liu R, To S-h, Yau T-k, Chung CC, Yau CC, Hui SM, Lau PY, Yeun C-h, Wong Y-w, Ho S, Fung SS, Tomilinson IP, Houlston RS, Cheng KK, Sham PC. (2011). Replication study of

SNP associations for colorectal cancer in Hong Kong Chinese. *Br J Cancer, 104(2): 369-375*

Hofferbert S, Schanen NC, Chehab F , Francke U. (1997). Trinucleotide Repeats in the Human Genome: Size Distribution for All Possible Triplets and Detection of Expanded Disease Alleles in a Group of Huntington Disease Individuals b the Repeat Expansion Detection Method. *Human Molecular Genetics, 6(1): 77-83*

Huang W, Wang P, Liu Z, Zhang L. (2009). Identifying disease associations via genome-wide association studies. *BMC Bioinformatics, 30;10 Suppl 1: S68*

Hummerich H, Baxendale S, Mott R, Kirby SF, MacDonald ME, Gusella J, Lehrach H , Bates GP. (1994).Distribution of trinucleotide repeat sequences across a 2 Mbp region containing the Huntington's disease gene. *Hum. Mol. Genet, 3(1): 73-78*

Ijaz S, Khan IA. (2009). Molecular characterization of wheat germplasm using microsatellite markers. *Genet. Mol. Res, 8(3): 809-815*.

Ijaz S. (2010). Microsatellite markers: An important fingerprint tool for characterization of crop plants. African journal of biotechnology 10(40), 7723-7726. Mol. Res, 8(3): 809-815.

Jiang J, Nasuda S, Dong F, Scherrer CW, Woo S-S, Wing RA, Gill BS , Ward DC. (1996). A conserved repetitive DNA element located in centromeres of cereal chromosomes. *Proc Natl Acad Sci U S A, 93(24) 14210-14213*

Jiang J, Birchler JA, Parrott WA , Dawe RK. (2003). A molecular view of plant centromeres. *TrendsPlant Sci, 8: 570-575*

Jurka J, Klonowski P, Dagman V et al., (1996). CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Comput Chem, 20: 119–121*

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O , Walichiewicz J. (2005. Repbase Updated, a database of eukaryotic repetitive elements. *Cytogenet Genome Res, 110: 462-467*

Kalendar R, Vicient CM, Peleg O et al., (2004). Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics, 166: 1437–1450*

Katti MV, Ranjekar PK, Gupta VS. (2001). Differerntial Distribution of Simpe Sequence Repeats in Eukaryotic Genome Sequences. *Mol.Biol.Evol, 18(7): 1161-1167*.

Kiliszek A, Kierzek R, Krzyzosiak WJ, Rypniewski W. (2010). Atomic resolution structure of CAG RNA repeats: structural insights and implications for the trinucleotide repeat expansion diseases. *Nucleic Acids Res, 38(22): 8370-6*

Knowles JW, Assimes T, Boerwinkle E, Fortmann SP, Go A, Grove ML, Hlatky M, Iribarren C, Li J, Myers R, Risch N, Sidney S, Southwick A, Volcik K, Quertermous T. (2008). Failure to replicate an association of SNPs in the oxidized LDL receptor gene (OLR1) with CAD. *BMC Medical Genetics, 23*

Kolpavok R, Bana G, Kucherov G. (2003). mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res, 31(13): 3672-3678*

Komissarov AS, Gavrilova EV, Demin SJ, Ishov AM, Podgornaya OI. (2011). Tandemly repeated DNA families in the mouse genome. *BMC Genomics, 12: 531*

Kremer EJ, Pritchard M, Lynch M, Yu S, Holman K, Baker E, Warren ST, Schlessinger D, Sutherland GR, Richards RI. (1991). mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science, 252(5013): 1711-1714*

Kulikova O, Geurts R, Lamine M, Kim DJ, Cook DR, Leunissen J, de Jong H, Roe BA, Bisseling T. (2004). Satellite repeats in the functional centromere and pericentromeric heterochromatin of Medicago truncatula. *Chromosoma,113(6): 276-83*.

Kurtz S, Ohlebusch E, Schleiermacher C. (2000). Computational and Visualization of degenerate repeats in Complete Genomes. *Molecular Biol, (ISMB2000), AAAI-Press, pp 228-238.*

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res, 29(22): 4633-4642*

Kwok P-Y, Chen X. (2003). Detection of Single Nucleotide Polymorphism. *Molecular Biol, 5, 43-60*

Lerat E, (2009). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity, 104: 520-533*

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod et al., (2007). The diploid genome sequence of an individual human. *PLoS Biol., 5: 2113–2144.*

Li R, Ye J, Li S et al., (2005). ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput Biol, 1: e43*

Li YC, Korol AB, Fahima T, Beiles A, Nevo E. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol, 11(12): 2453-65*

McCarthy EM, McDonald JF. (2003). LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics, 19: 362–367*

McClintock B. (1950). The origin and Behavior of Mutable Loci in Maize. *Proc Natl Acad Sci USA, 36(6): 344-355*

McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Glossary.

Michael Allaby. "replication slippage." A Dictionary of Ecology. 2004. Retrieved May 29, 2012 from Encyclopedia.com: http://www.encyclopedia.com/doc/1O14-replicationslippage.html

Murphy DT, Karpen GH**.** (1998). Centromeres take flight: alpha satellite and the quest for the human centromere. *Cell, 93: 317-320*

Myles S, Davison D, Barrett J, Stoneking M, Timpson N. (2008). Worldwide population differentiation at disease-associated SNPs. *BMC Medical Genomics 1: 22*

Nair P, Vihinen M. (2012). A benchmark database for variations. Retrieved April 2011 from http://bioinf.uta.fi/VariBench/.

Nei M, Kumar S. (2000). Molecular Evolution and Phylogenetics. Oxford University Press

Oberlé I, Rousseau F, Heitz D, Kretz C, Devys D, Hanauer A, Boué J, Bertheas MF, Mandel JL. (1991). Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science, 252: 1097-102.*

Ohno S. (1972). So much 'junk' in our genomes. *Brookhaven Symp Biol, 23: 366–370*

Ohshima K, Okada N. (2005). SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. *Cytogenet Genome Res, 110: 475–490*

Ouyang S, Buell CR. (2004). The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res, 32: D360–D363*

Paulson HL , Fischbeck KH. (1996). Trinucleotide Repeats in Neurogenetic Disorders. *Neuroscience, 19: 79-107*

Pevzner P.A, Tang H , Tesler G. (2004). De Novo Repeat Classification and Fragment Assembly. *Genome Res, 14: 1786-1796*

Piirilä H, Väliaho J, Vihinen M. (2006). Immunodeficiency mutation databases (IDbases) *Hum. Mutat, 27(12): 1200-1208*

Plohl M, Luchetti A, Mestrović N, Mantovani B. (2008). Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. *Gene, 15: 409(1-2): 72-82.*

Pray L. (2008). Transposons, or Jumping Genes: Not junk DNA. *Nature Education, 1(1)*

Price AL, Jones NC, Pevzner PA. (2005). De novo identification of repeat families in large genomes. *Bioinformatics, 21(Suppl 1): i351–i358*

Ramensky V, Bork P ,Sunyaev S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res, 30(17): 3894-3900*

Rao SR, Trivedi S, Emmanuel D, Merita K ,Hynniewta M. (2010). DNA repetitive sequences-types, distribution and function: A review. *Journal of Cell and Molecular Biology, 7(2) & 8(1): 1-11*

Richard F.G., Kerrest A. Dujon B. (2008). Comparative genomics and Molecular Dynamics of DNA repeats in Eukaryotes. *Microbiol. Mol. Biol, 72(4): 686-727*

Rudd MK, Willard HF. (2004). Analysis of the centromeric regions of the human genome assembly. *Trends in genetics, 20: 529–33*

Rudd MK, Wray GA, Willard HF. (2006). The evolutionary dynamics of alpha-satellite. *Genome Res, 16(1): 88-96*

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, WatersonRH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumensteil B, Baldwin J, Stange- Thomann N, Zody MC, Linton L, Lander ES, Altshuler D. (2001). International SNP Map Working Group A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature, 409: 298-993*

Saha S, Bridges S, Magbanua ZV,Peterson DG. (2008a). Empirical comparism of ab initio repeat finding programs. *Nucleic Acids Res, 36(7): 2284-2294*

Saha S, Bridges S, Magbanua ZV, Peterson DG (2008b). Computational Approaches and Tools Used in Identification of Dispersed Repetitive DNA Sequences. *Tropical Plant Biol, 1: 85-96*

Salat U, Bardoni B, Wöhrle D,Steinbach P.(2000). Increase of FMRP expression, raised levels of FMR1 mRNA, and clonal selection in proliferating cells with unmethylated fragile X repeat expansions: a clue to the sex bias in transmission of full mutations? *J Med Genet, 37: 842-850*

Schmidt T, Heslop-Harrison JS. (1996). The physical and genomic organization of microsatellites in sugar beet. *Proceedings of the National Academy of Sciences USA*, *93: 8761-8765.*

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. (2001). Genomic and genetic definition of a functional human centromere. *Science, 294: 109–15.*

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29: 308-311.*

Shuqing Xu, Terry Clark, Hongkun Zheng, Søren Vang, Ruiqiang Li, Gane Ka-Shu Wong, Jun Wang, Xiaoguang Zheng (2008) Gene conversion in the rice genome
BMC Genomics, 9: 93.

Sirota M, Schaub MA, Batzoglou S, Robinson WH, Butte AJ (2009). Autoimmune Disease Classification by Inverse Association with SNP Alleles. *PLoS Genet, 5(12): e1000792. doi: 10.1371/journal.pgen.1000792*

Siwach P, Ganesh S. (2008). Tandem repeats in human disorders: mechanisms and evolution. *Front Biosci, 1(13): 4467-84*

Smit AFA, Hubley R, Green P (1996–2004) RepeatMasker Open-
3.0. http://www.repeatmasker.org

Stenger JE, Lobachev KS, Gordenin D, Darden TA, Jurka J, Resnick MA. (2001). Biased Distribution of Inverted and Direct Alus in the human Genome: Implications for Insertion, Exclusion, and Genome Stability. *Genome Res, 11: 12-27*

Subramanian S, Mishra RK, Singh L. (2002). Genome-wide analysis of microsatellite repeats in humans: their distribution and density in specific genomic regions. *Genome Biol, 4(2) R13*

Subramanian S, Madgula VM, George R, Mishra RK, Pandit WM, Kumar CS, Singh L. (2003). Triplet repeats in human genome: distribution and their association with genes and other genomic regions. *Bioinformatics, 19(5): 549-552*

Sutherland GR, Richards RI. (1995). Simple tandem DNA repeats and human genetic disease. *PNAS, 92(9): 3636-3641*

Tang J, Leunissen J AM, Voorrips R E, Linden CGV, Vosman B. (2008) HaploSNPer: a web- based allele and SNP detection tool. *BMC Genetic, 9: 23*

Teshima KM, Innan H .(2004). The effect of gene conversion on the divergence between duplicated genes. Genetics, *166(3): 1553-1560.*

Toth G, Gaspari Z, Jurka J. (2000). Microsatellites in Different Eukaryotic Genomes: Survey and Analysis. *Genome Res, 10: 967-981*

Ugarkovic D, Plohl M. (2002). Variation in satellite DNA profiles- causes and effect. *The EMBO journal 21: 5955-5959*

Usdin K. (2008). The biological effects of simple tandem repeats: Lesson from the repeat expansion diseases. *Genome Res, 18: 1011-1019*

Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH,Kuhl DP, Pizzuti A, Reiner O, Richards S, Victoria MF, Zhang FP, Eussen BE, van Ommen GJB, Blonden LAJ, Riggins GJ, Chastein JL, Kunst CB, Galjaard H, Caskey CT, Nelson DL, Oostra BA,Warren ST. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell, 95: 905-14.*

Vincent JB, Paterson AD, Strong E, Petronis A, Kennedy JL. (2000). The unstable trinucleotide repeat story of major psychosis. *Am J med Genet, 97(1): 77-97*

Volfovsky N, Haaa JB, Salzberg SL. (2001). A clustering method for repeat analysis in DNA sequences.*Genome Biol, 2*

Vukich M, Giordani T, Natali L, Cavallini A. (2009). Copia and Gypsy retrotransposons activity in sunflower (*Helianthus annuus L.) BMC Plant Biology, 9: 50*

Warburton PE, Giordano J, Cheung F et al., (2004). Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res, 14: 1861–1869*

Wicker T, Sabot F, Hua-Van A et al., (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet, 8: 973–982*

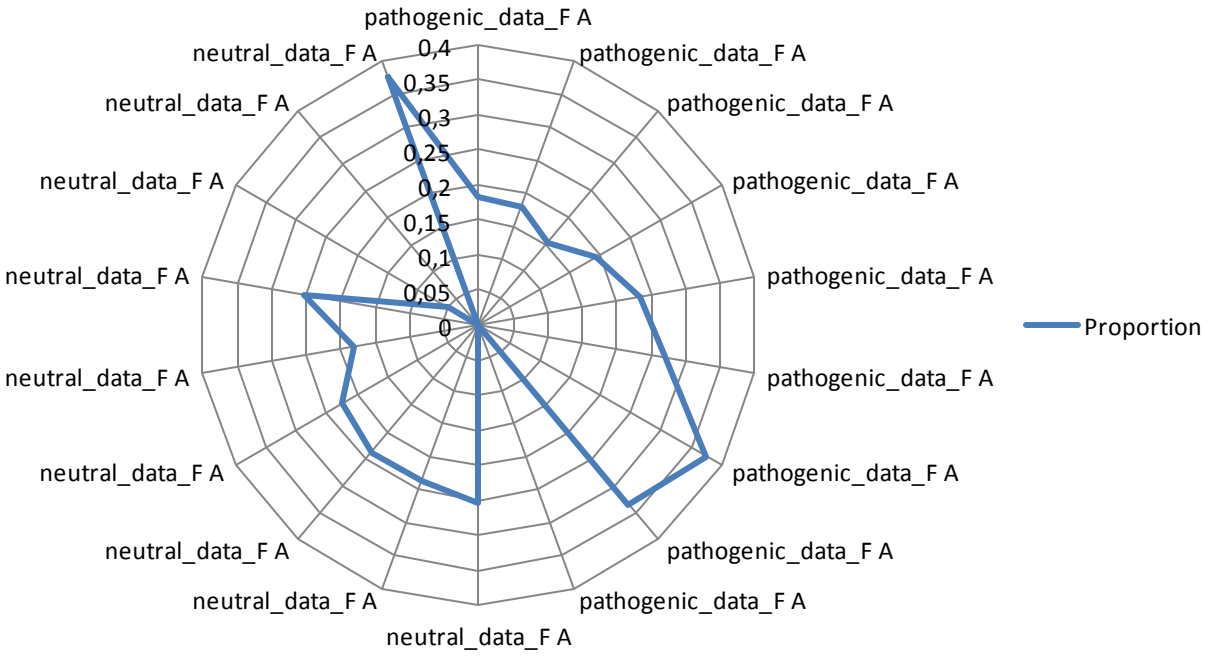
Yan HM, Dong C, Zhang EL, Tang CF, A XX, Yang WY, Yang YY, Zhang FF, Xu FR. (2012). Analysis of genetic variation in rice paddy landraces across 30 years as revealed by microsatellite DNA marker. *Yi Chuan 34(1): 87-94*

9 Appendices

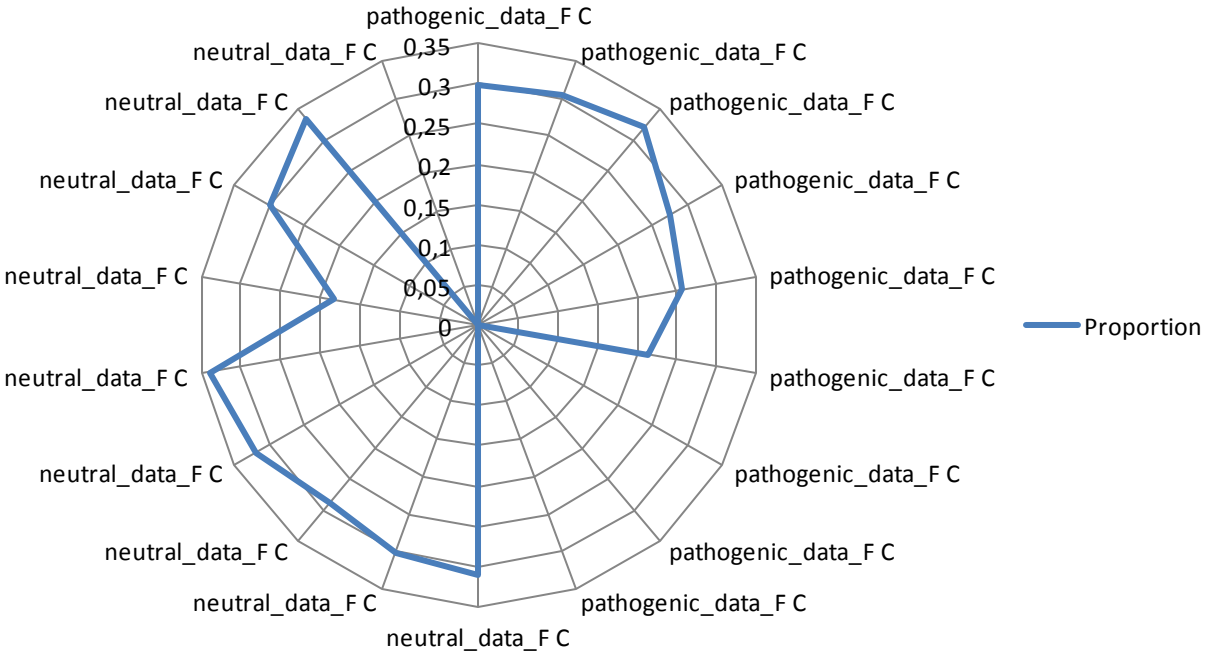
```
myfile=open("M_nonpathogenicset.txt","w")
from Bio import SeqIO
handle=open("NM_Seqfasta.fasta")
for seq_record in SeqIO.parse(handle, "fasta"):
   id=my_newid(seq_record.id)
   #print id
   if id in d.keys():
      pos=d[id]
      for p in pos:
         p= p+1
         seq1=seq_record.seq[p-10:p+11]
         seq2=seq_record.seq[p]
         lines=id,seq2
         lines2= id+"\t" +str(seq1)+"\n"
         myfile2.writelines(lines2)
         lines2=id+"_"+str(p)+"\t"+str(seq1)+"\t"+"\n" str(p)+"\t"+str(seq1)+"\n"
         lines2=id+"|"+seq2+"|"+str(p)+"\n"+str(seq1) +"\n"
         myfile.writelines(lines2)
handle.close()
myfile.close()
```

Figure A- 1. Sequence retrieval script

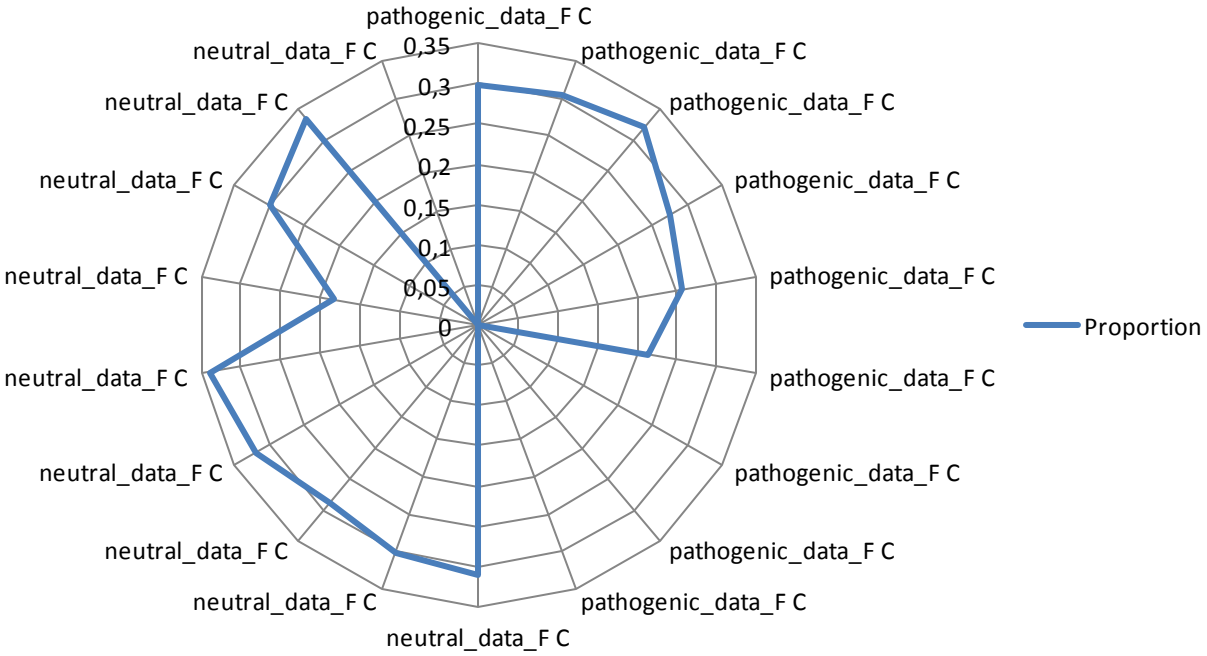
Adenine plot in Pathogenic vs Neutral datasets Forward Repeat from position 8



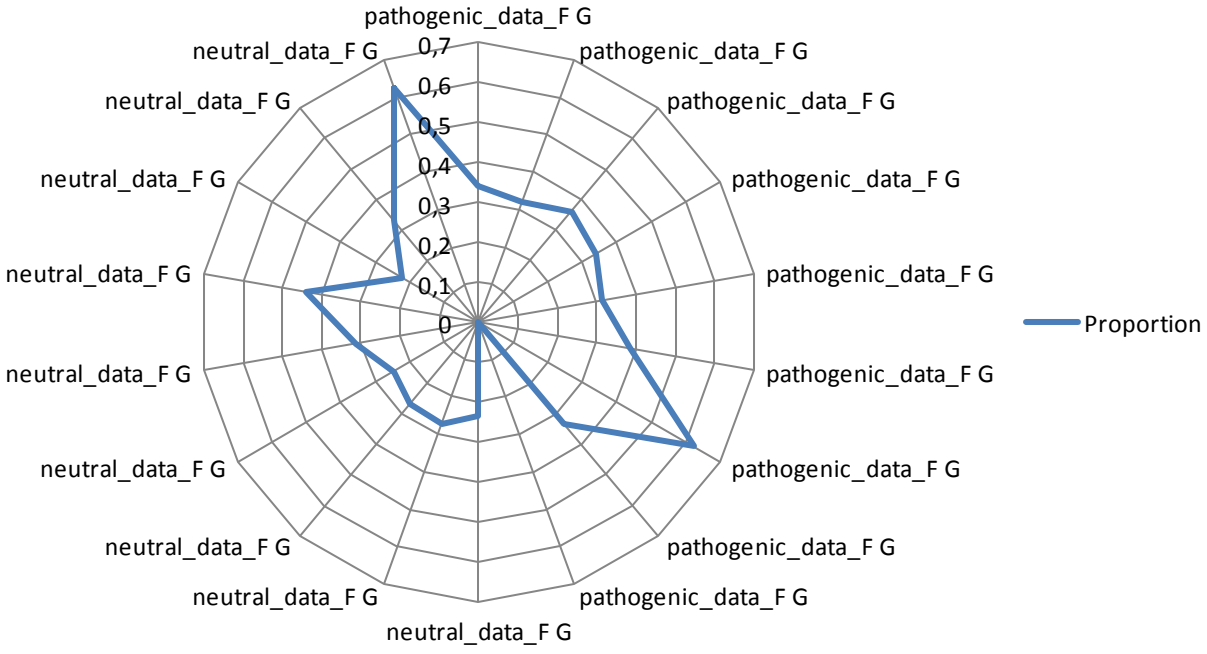
Cytosine plot in Pathogenic vs Neutral datasets Forward Repeat from position 8



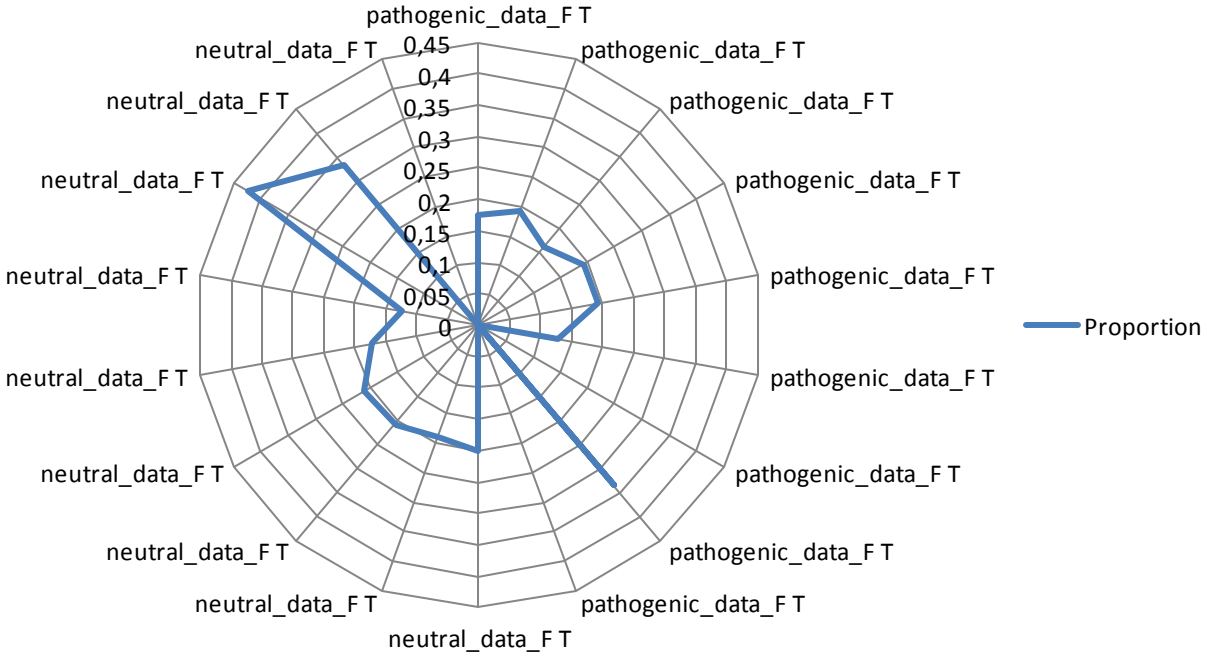
Cytosine plot in Pathogenic vs Neutral datasets Forward Repeat from position 8



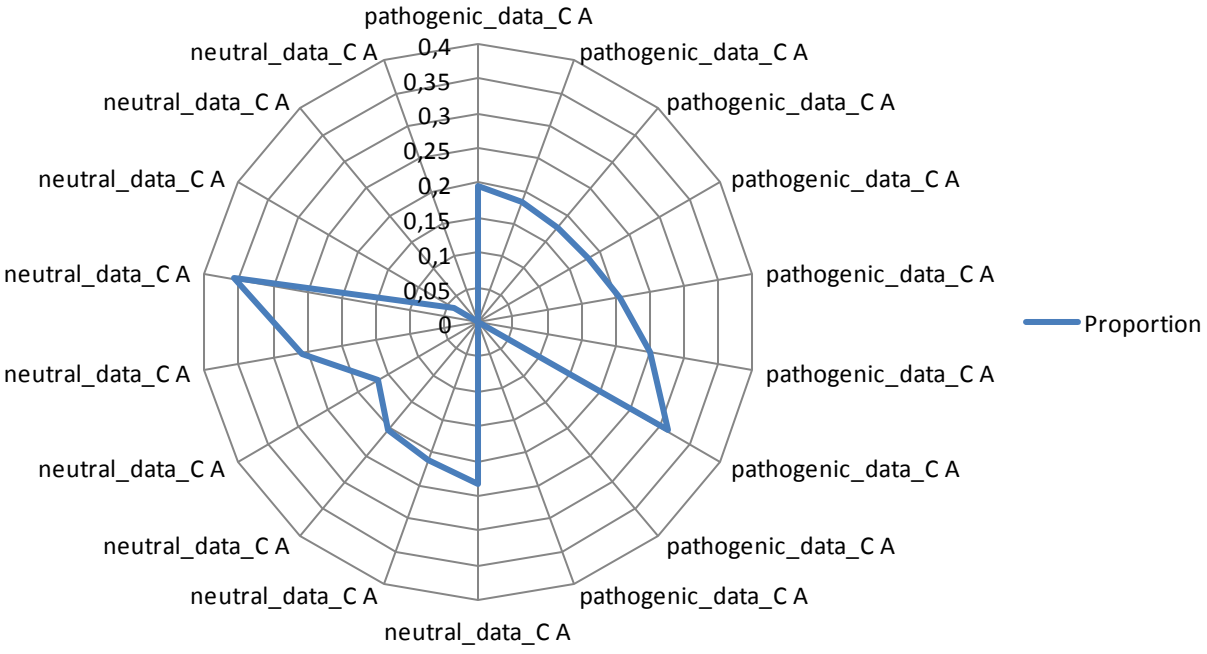
Guanine plot in Pathogenic vs Neutral datasets Forward Repeat from position 8



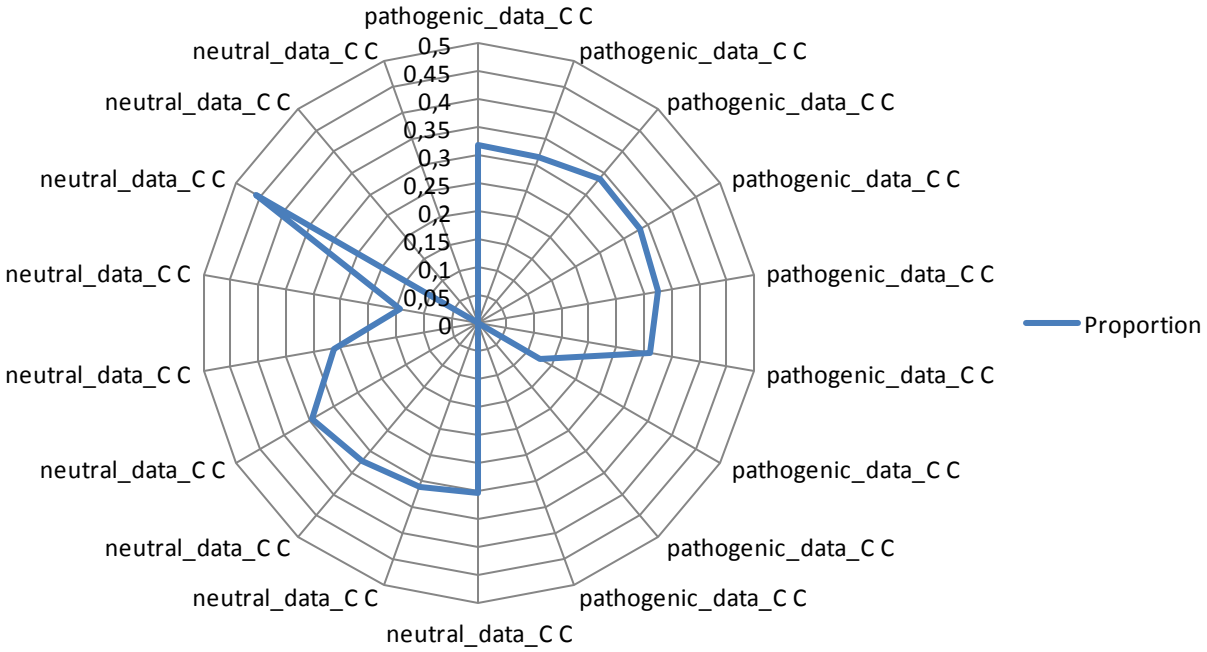
Thymine plot in Pathogenic vs Neutral datasets Forward Repeat from position 8



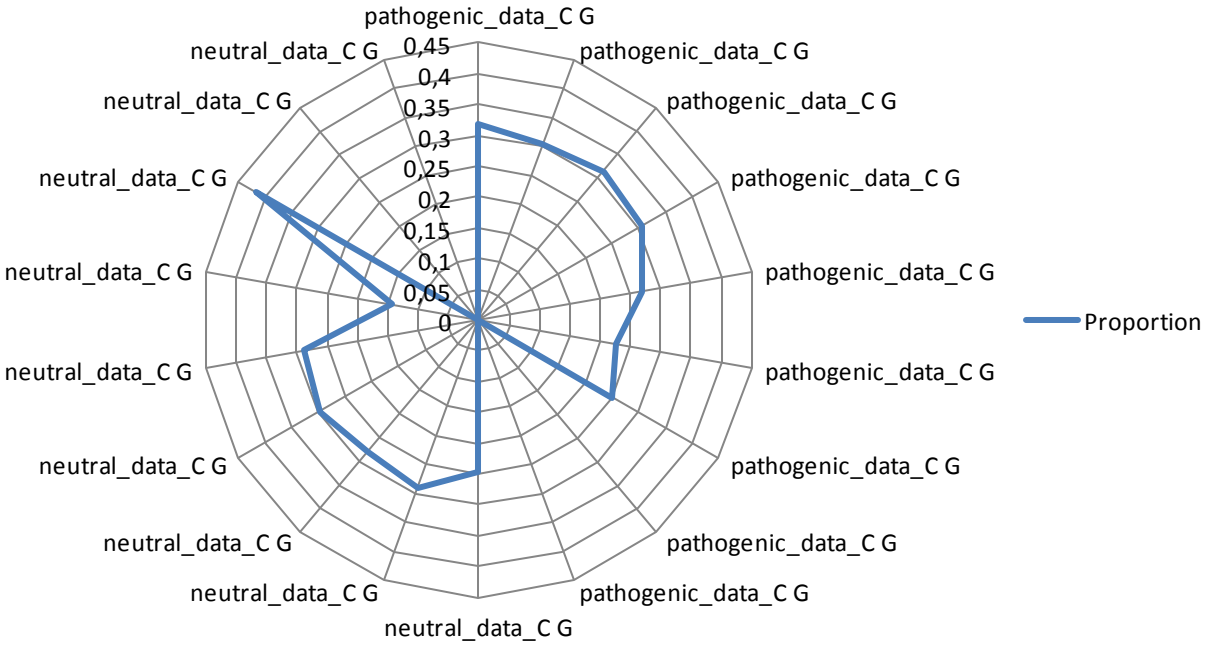
Adenine plot in Pathogenic vs Neutral datasets Complement Repeat from position 8



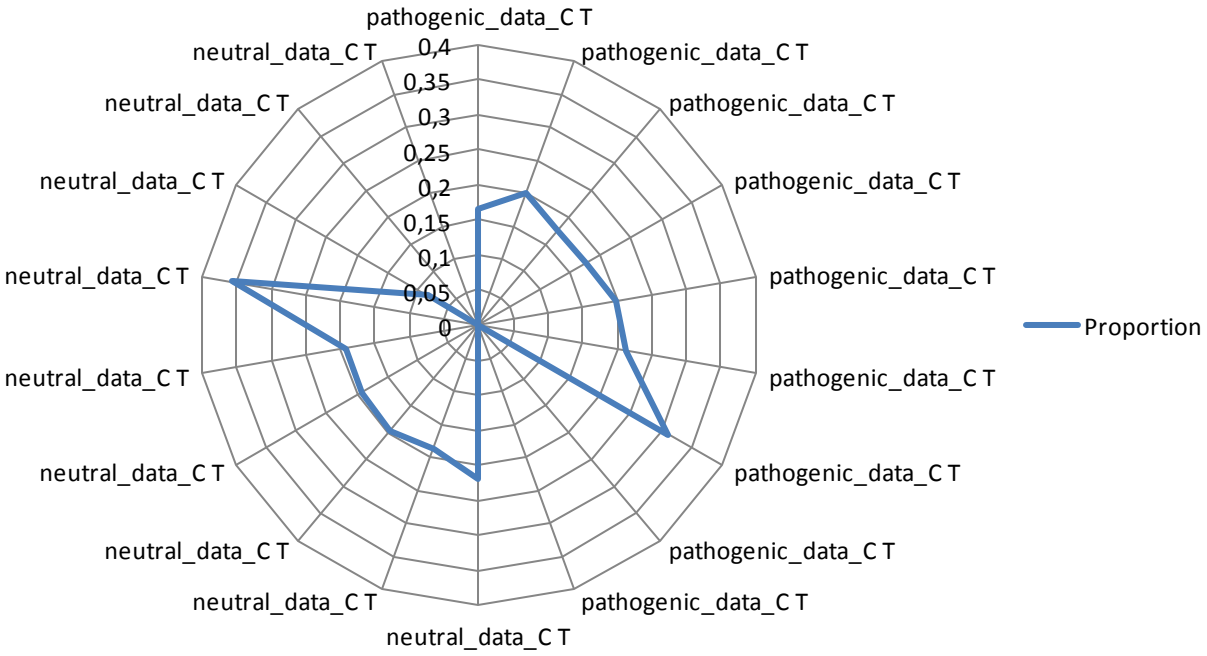
Cytosine plot in Pathogenic vs Neutral datasets Complement Repeat from position 8



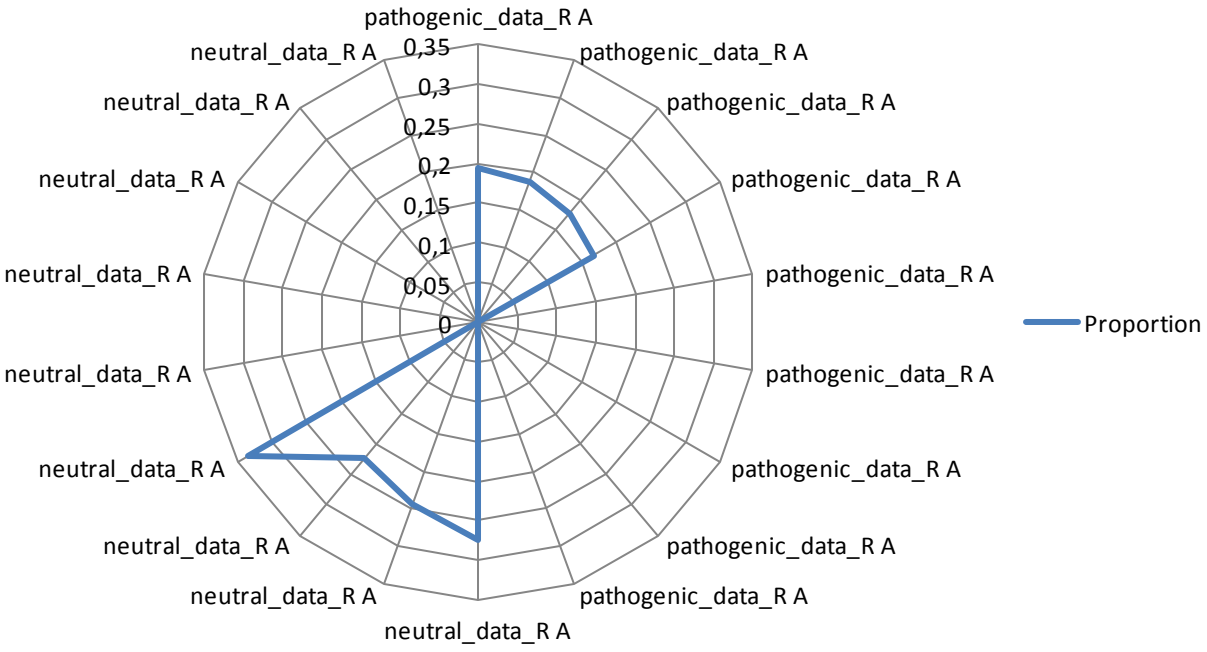
Guanine plot in Pathogenic vs Neutral datasets Complement Repeat from position 8



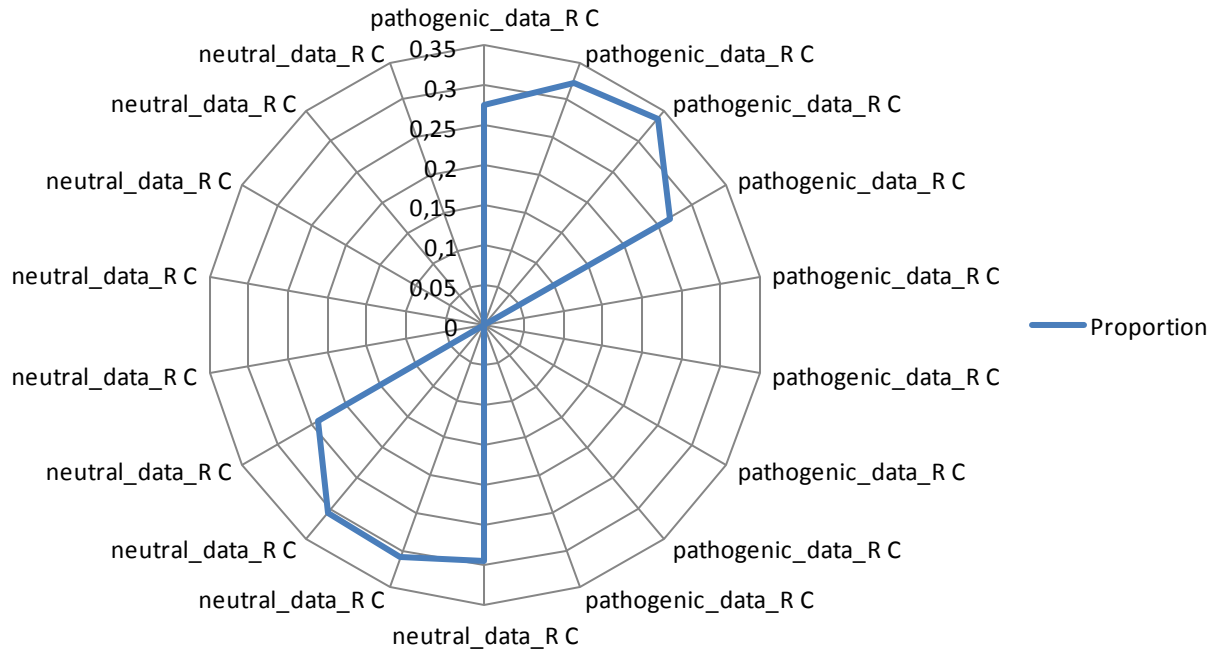
Thymine plot in Pathogenic vs Neutral datasets Complement Repeat from position 8



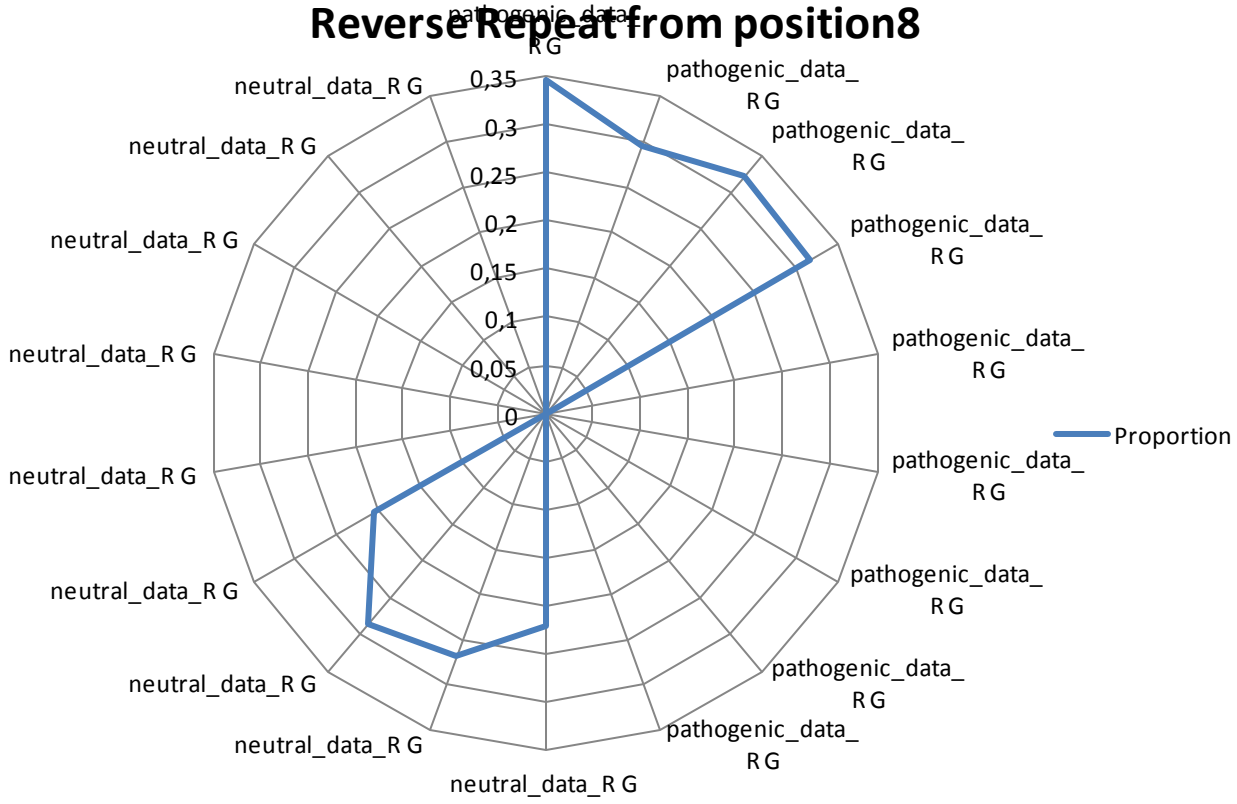
Adenine plot in Pathogenic vs Neutral datasets Reverse Repeat from position8



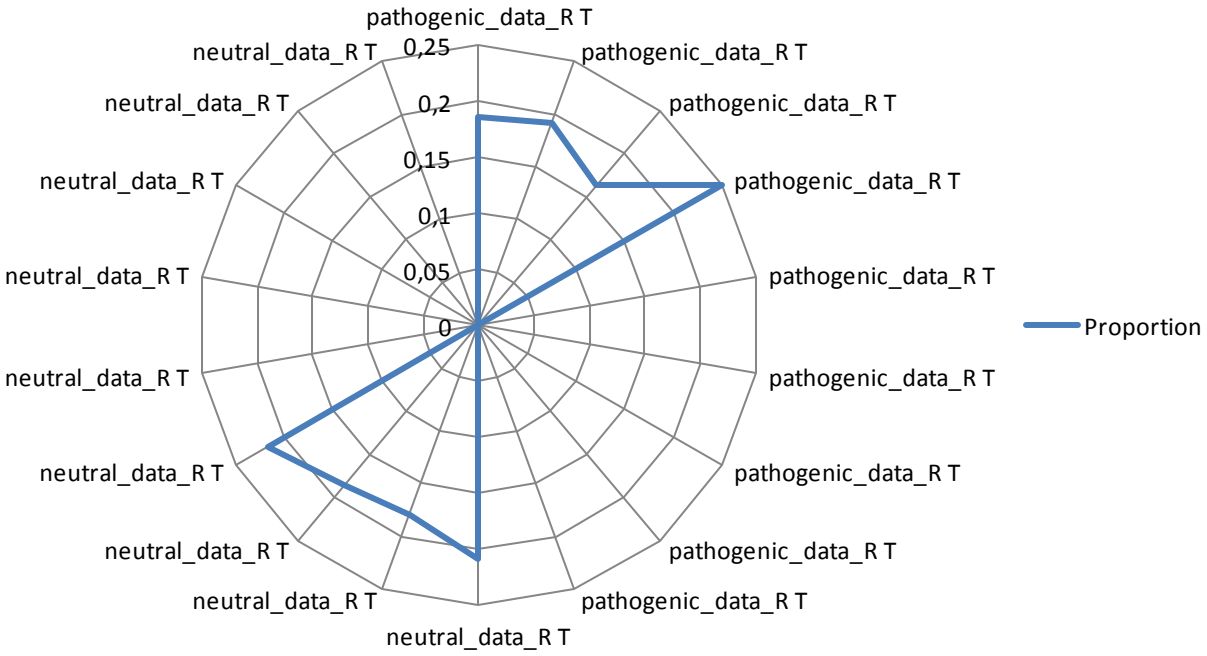
Cytosine plot in Pathogenic vs Neutral datasets Reverse Repeat from position8



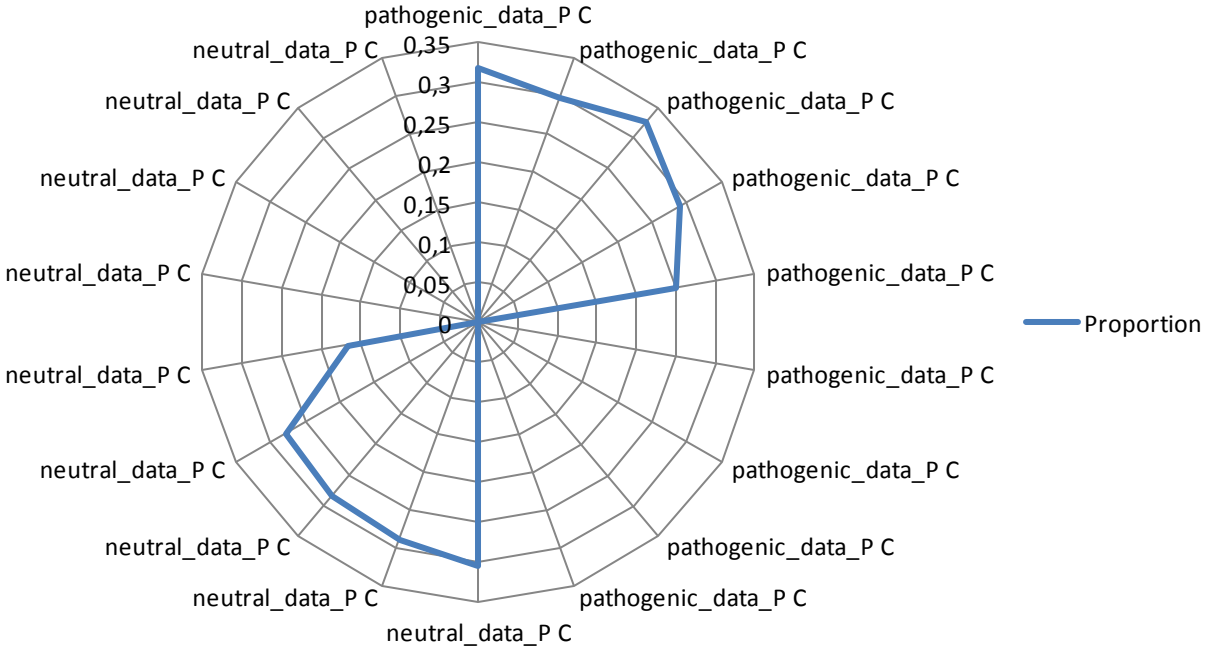
Guanine plot in Pathogenic vs Neutral datasets Reverse Repeat from position8



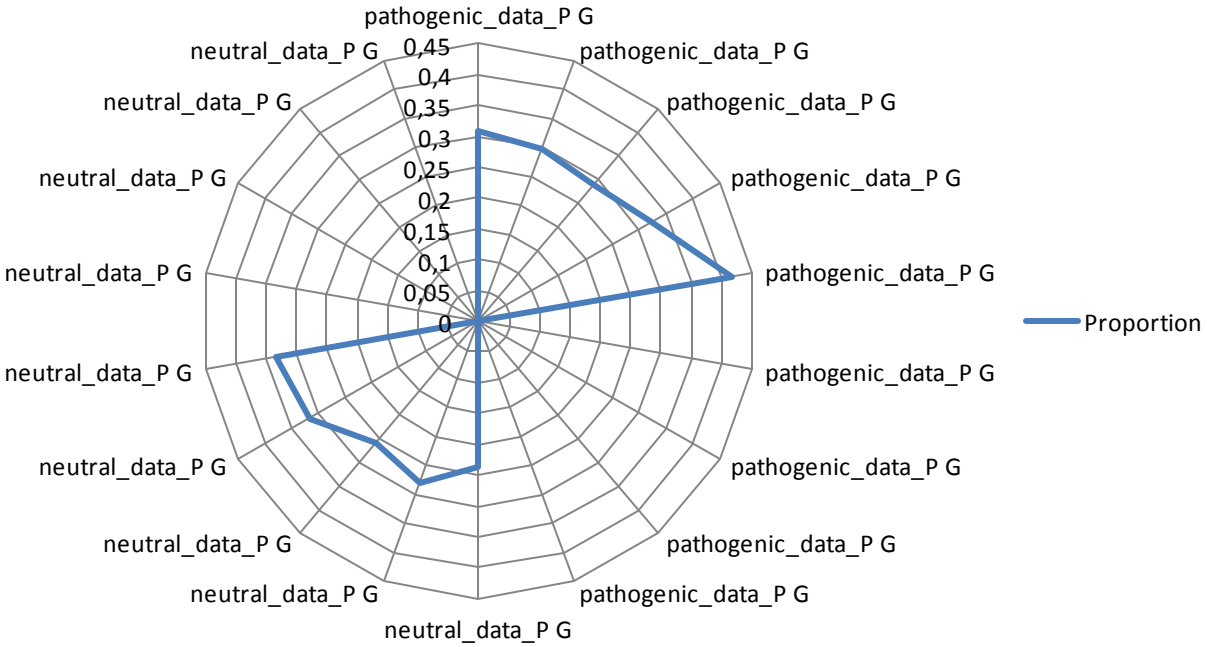
Thymine bplot in Pathogenic vs Neutral datasets Reverse Repeat from position 8



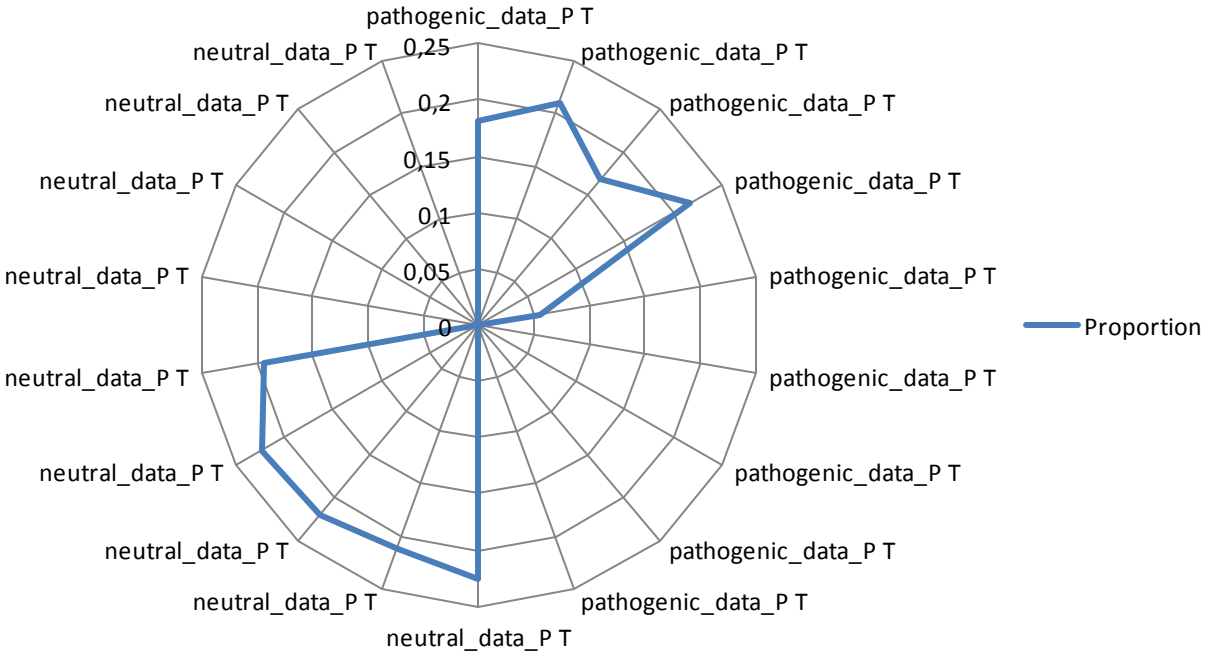
Cytosine plot in Pathogenic vs Neutral datasets Palindromic Repeat from position 8



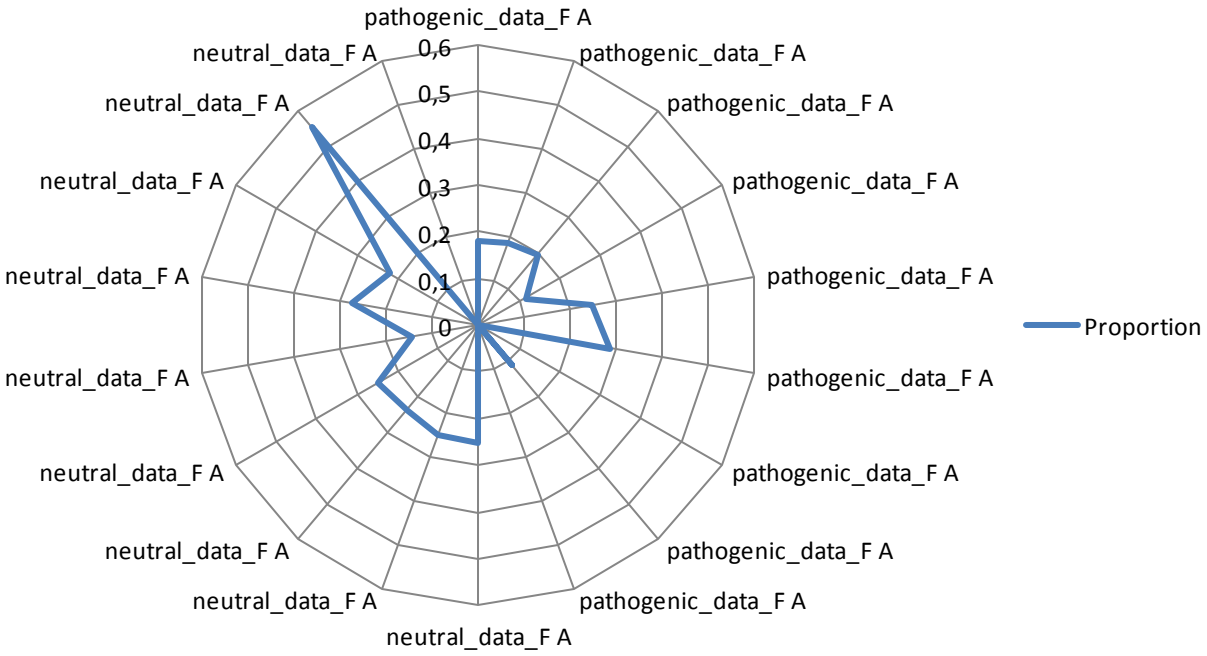
Guanine plot in Pathogenic vs Neutral datasets Palindromic Repeat from position 8



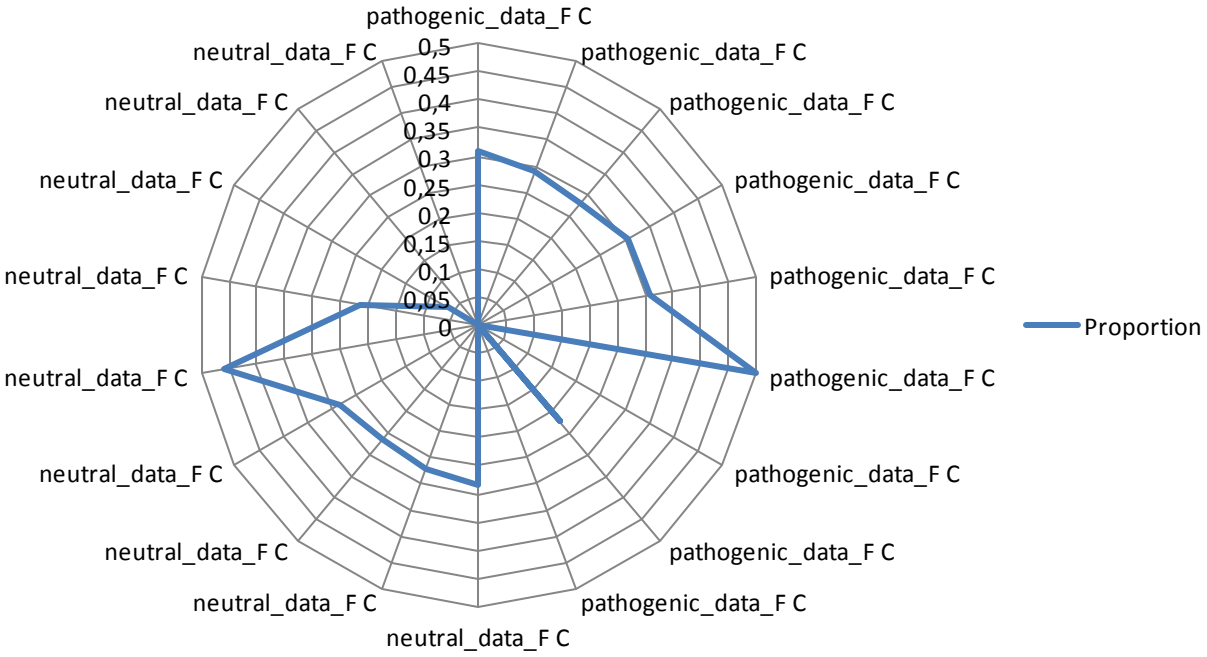
Thymine plot in Pathogenic vs Neutral datasets Palindromic Repeat from position 8



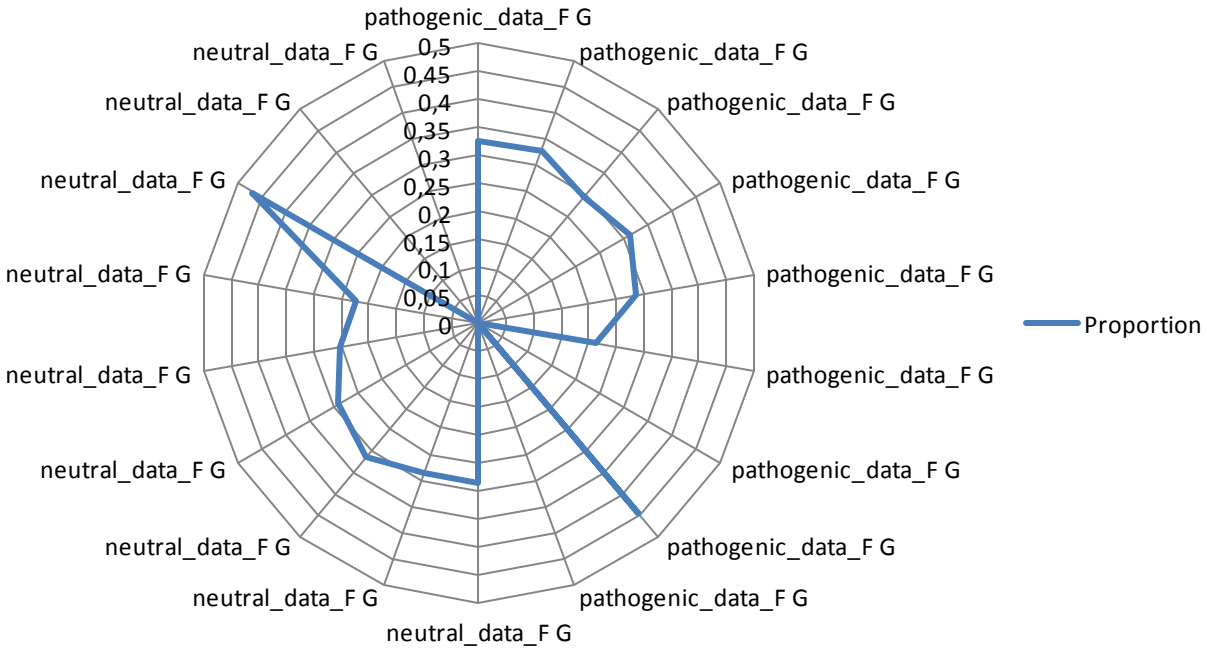
Adenine plot in Pathogenic vs Neutral datasets Forward Repeat from position 9



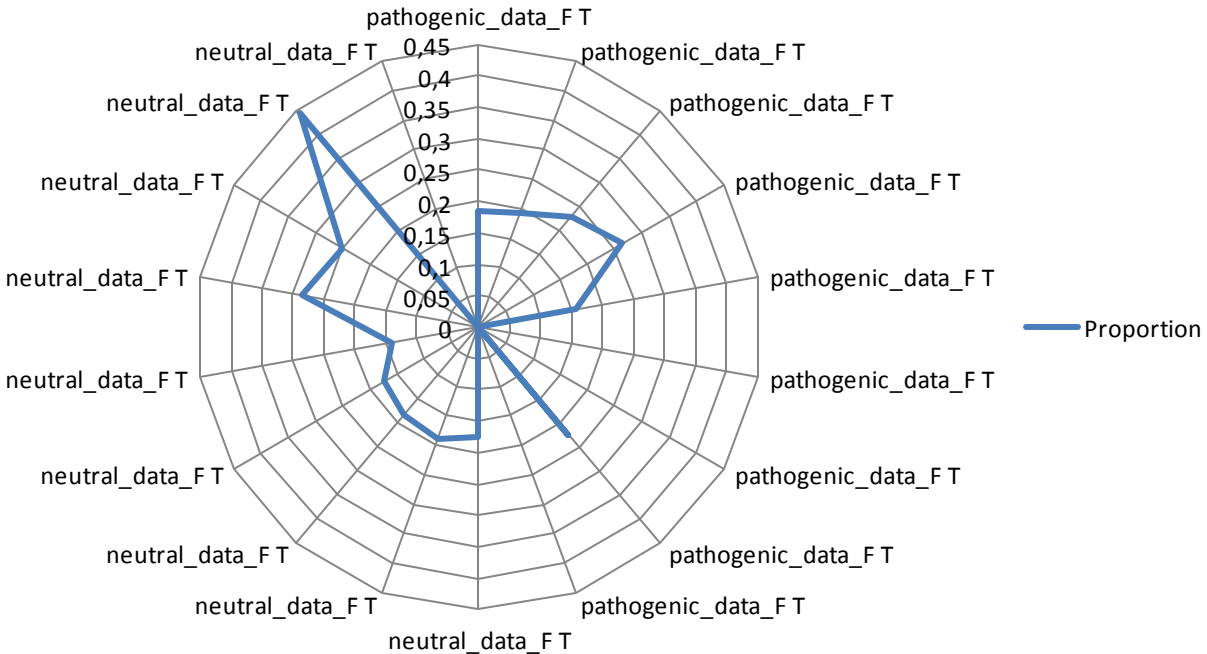
Cytosine plot in Pathogenic vs Neutral datasets Forward Repeat from position 9



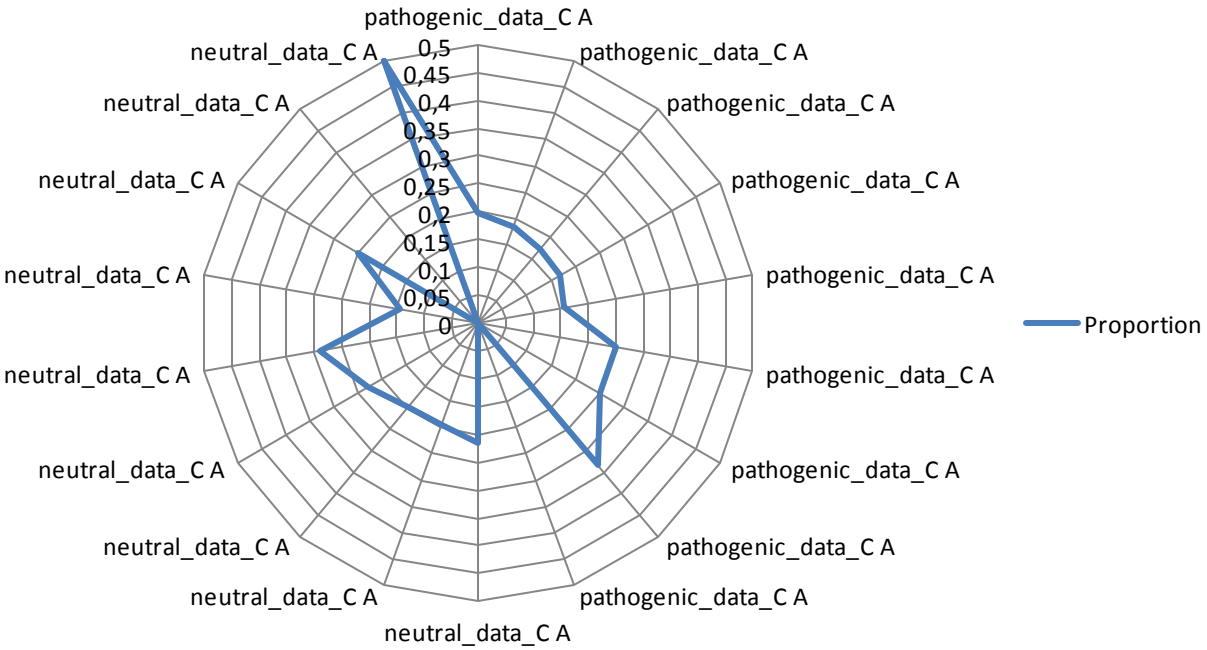
Guanine plot in Pathogenic vs Neutral datasets Forward Repeat from position 9



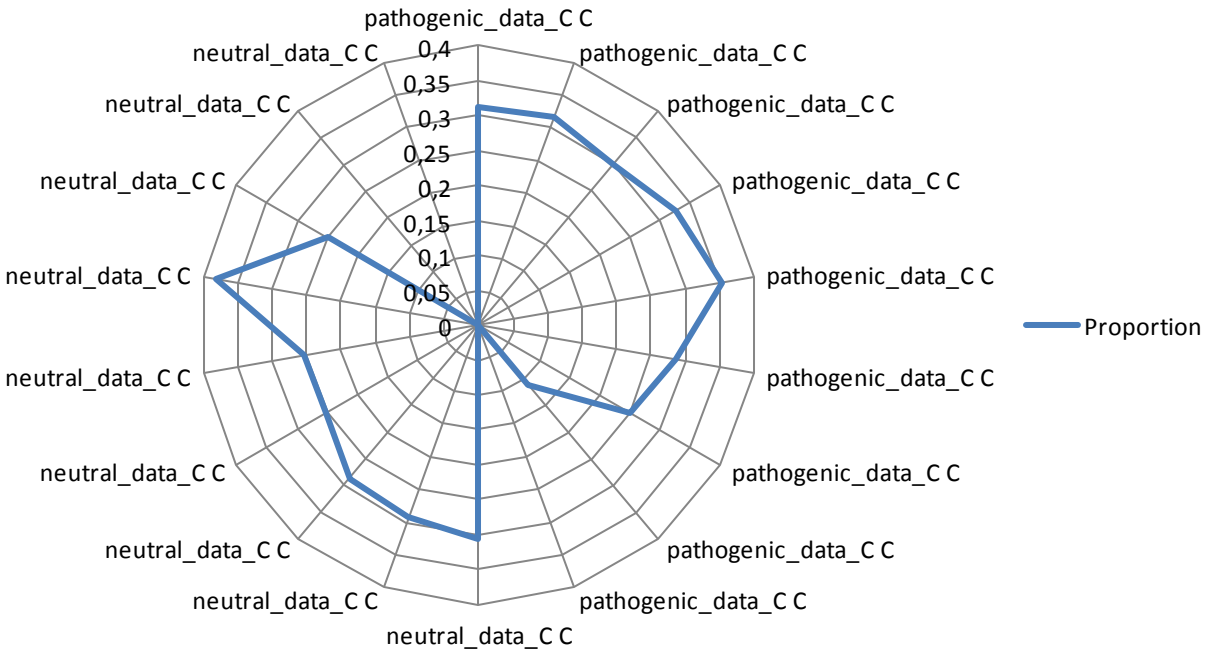
Thymine plot in Pathogenic vs Neutral datasets Forward Repeat from position 9



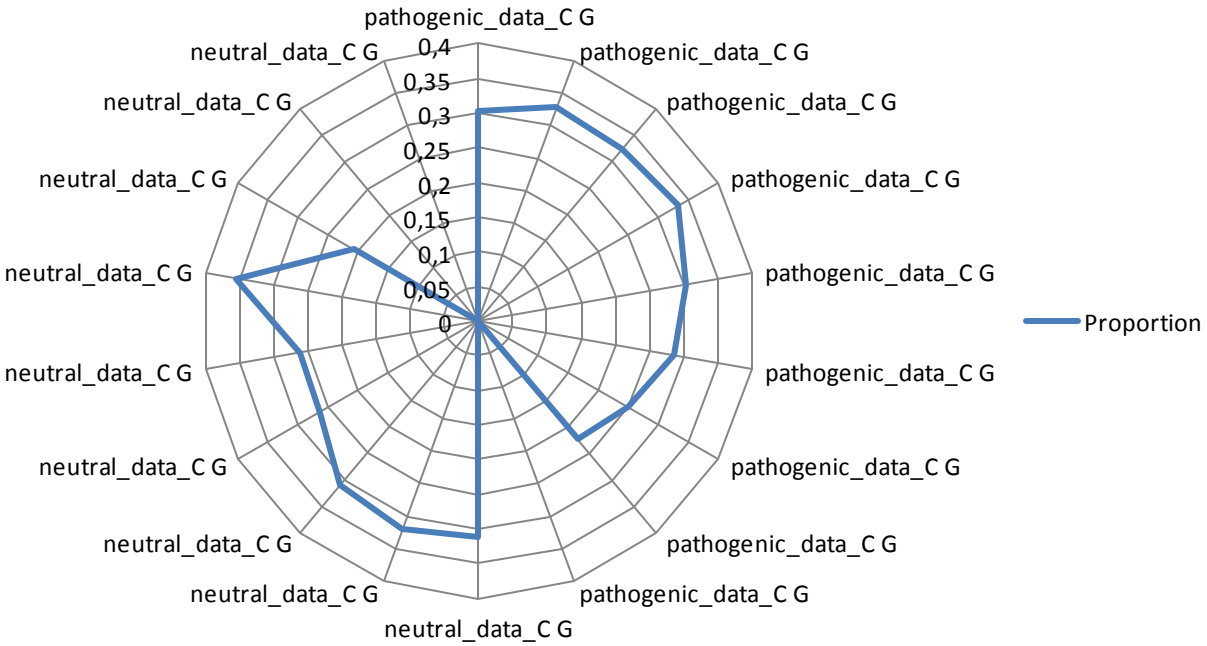
Adenine plot in Pathogenic vs Neutral datasets Complement Repeat from position 9



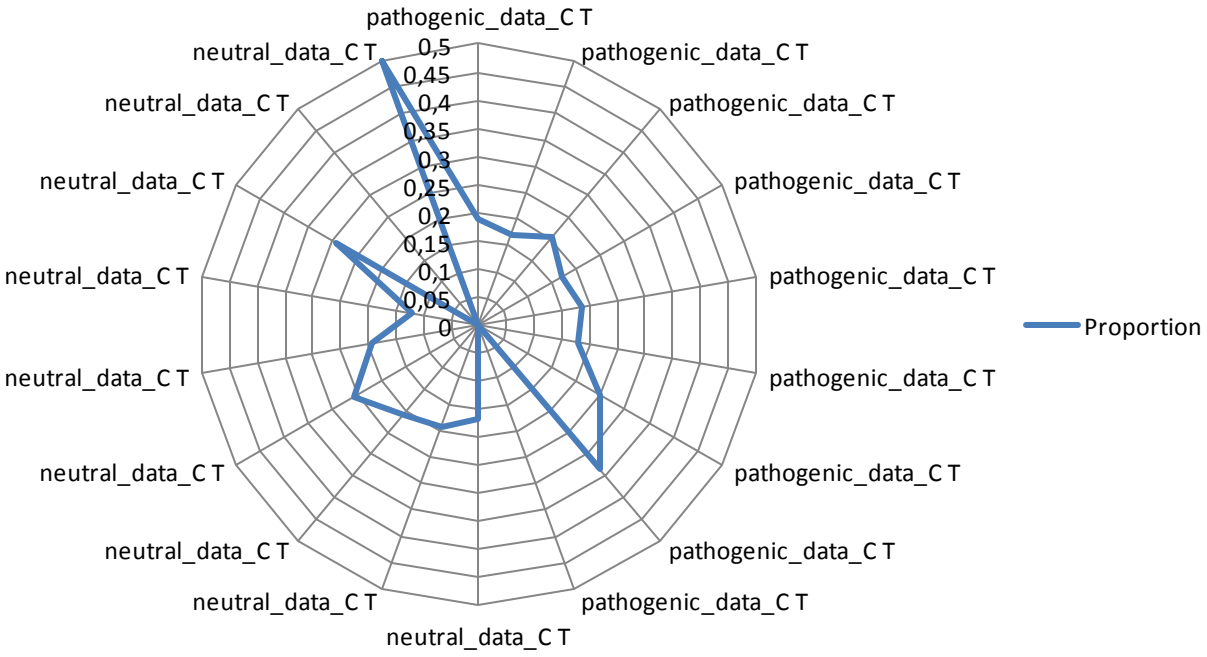
Cytosine plot in Pathogenic vs Neutral datasets Complement Repeat from position 9



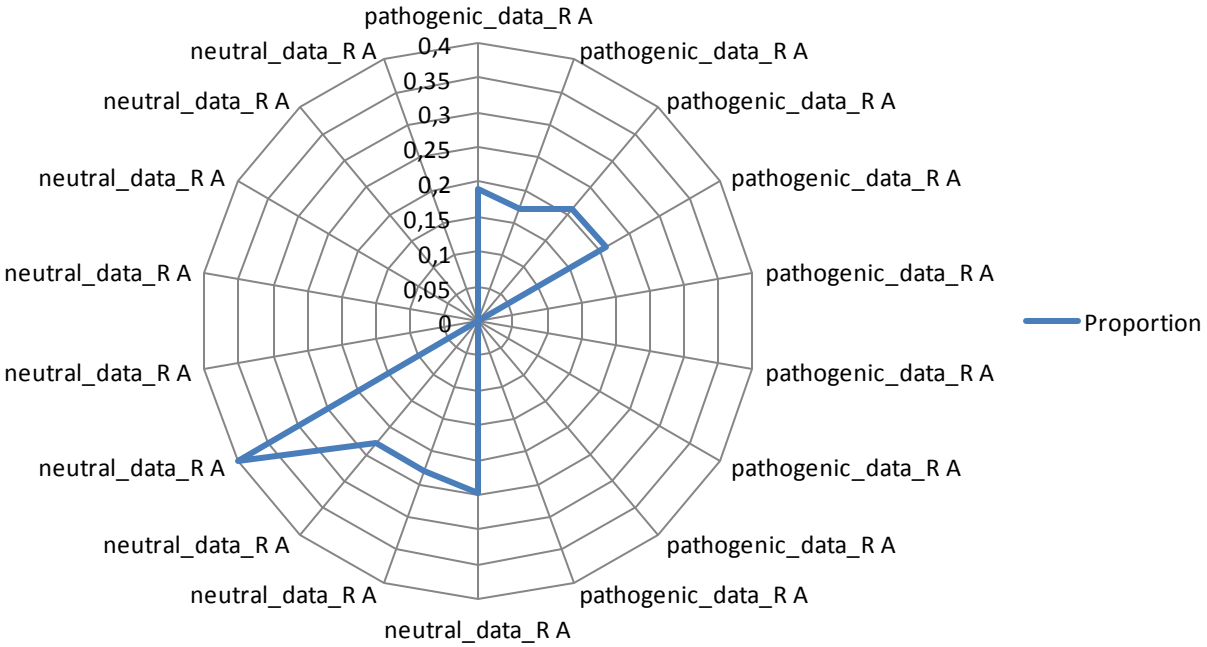
Guanine plot in Pathogenic vs Neutral datasets Complement Repeat from position 9



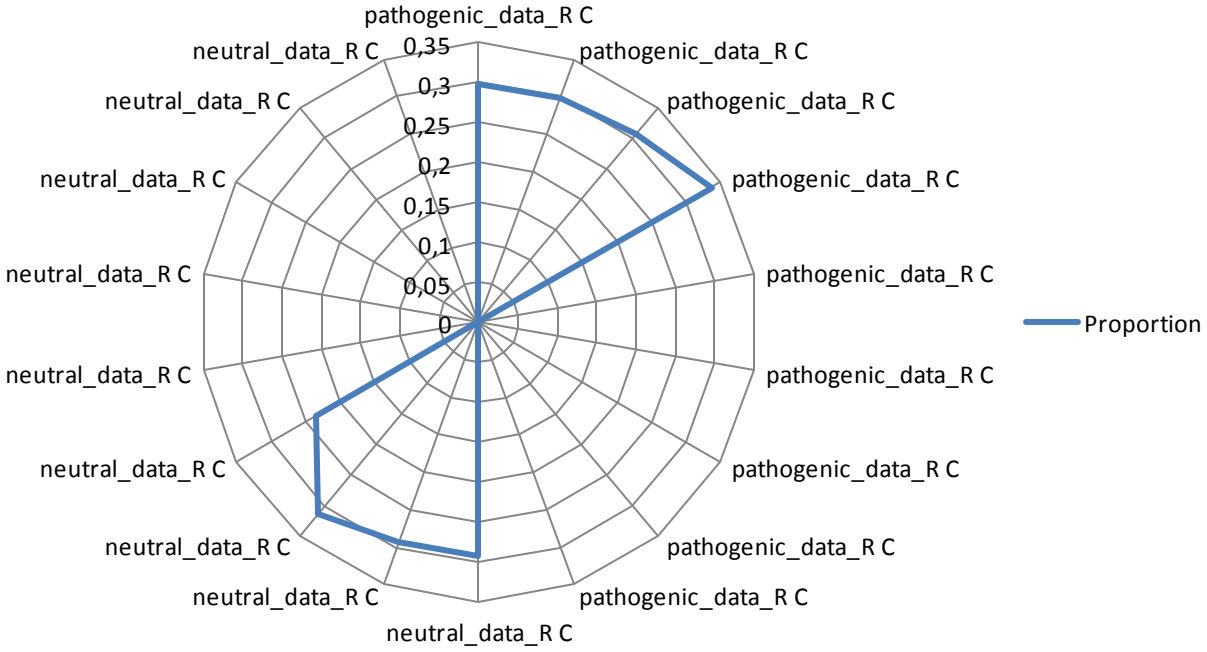
Thymine plot in Pathogenic vs Neutral datasets Complement Repeat from position 9



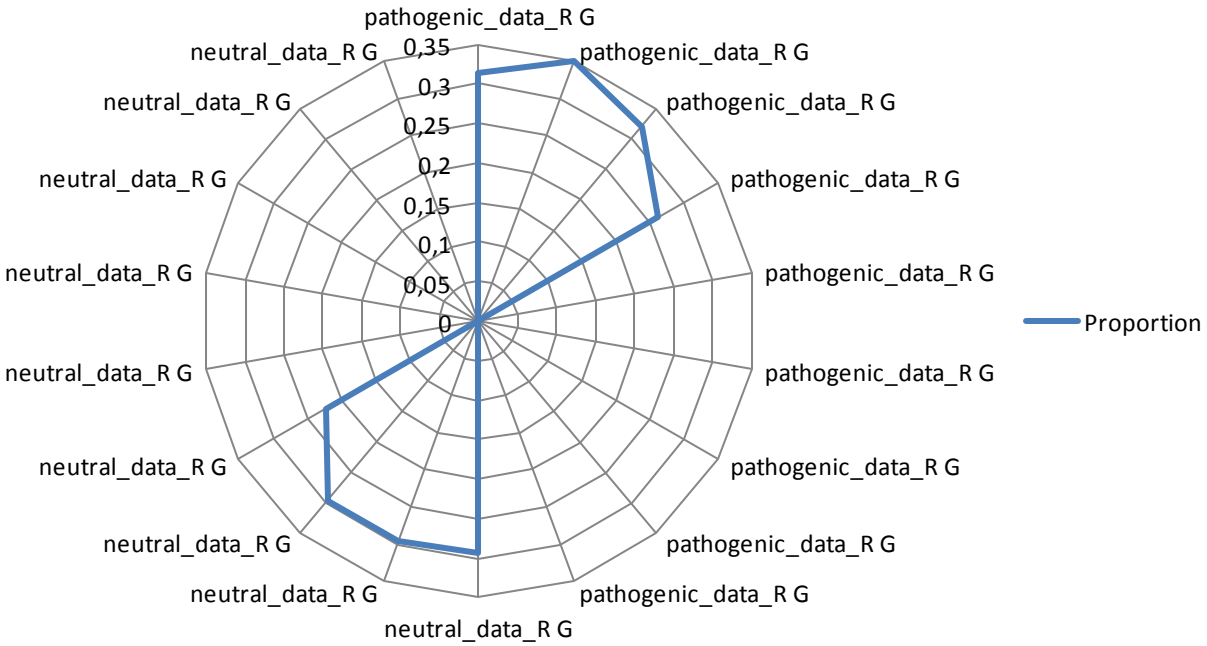
Adenine plot in Pathogenic vs Neutral datasets Reverse Repeat from position9



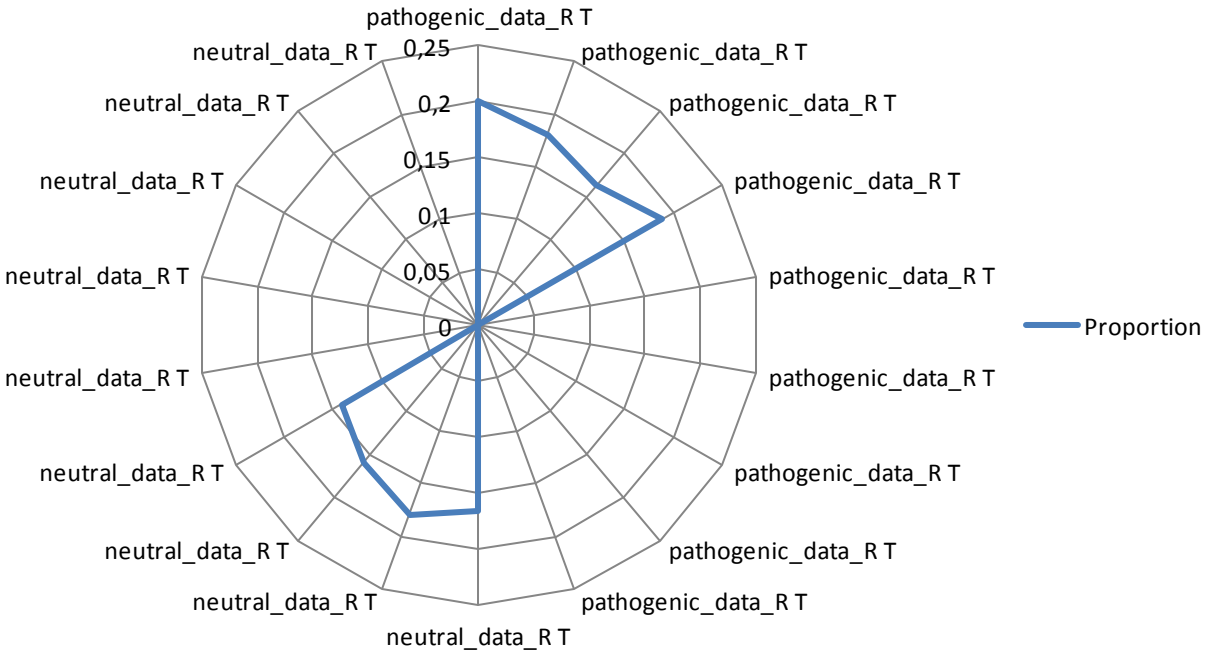
Cytosine plot in Pathogenic vs Neutral datasets Reverse Repeat from position9



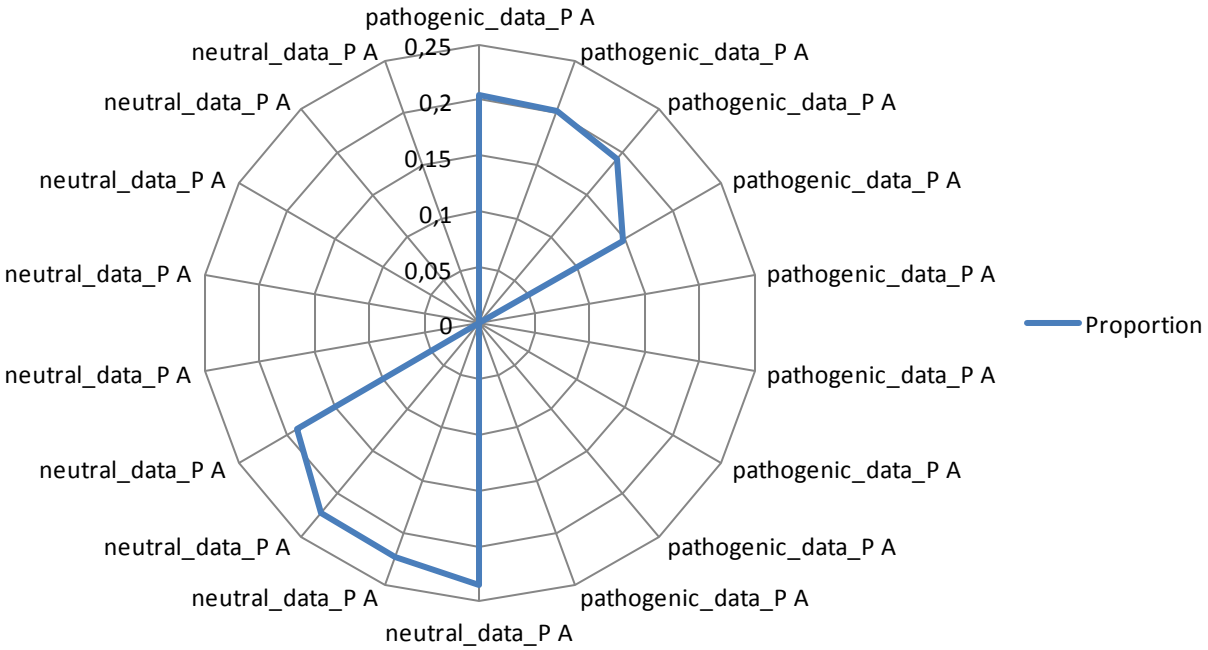
Guanine plot in Pathogenic vs Neutral datasets Reverse Repeat from position9



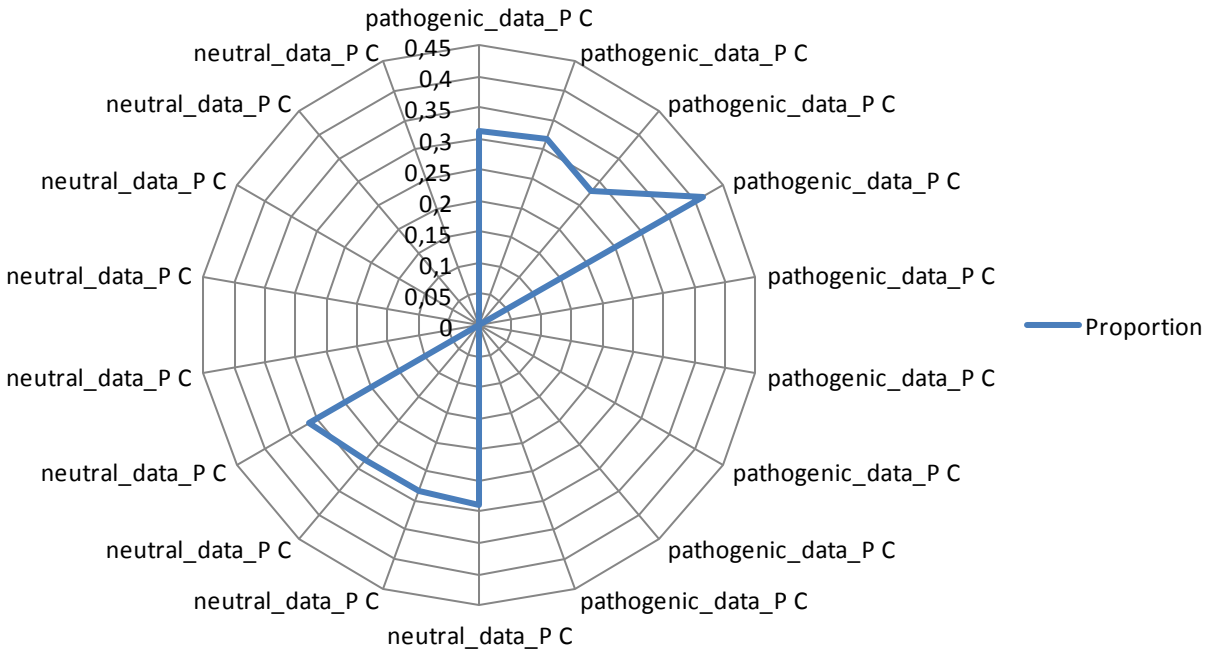
Thymine plot in Pathogenic vs Neutral datasets Reverse Repeat from position 9



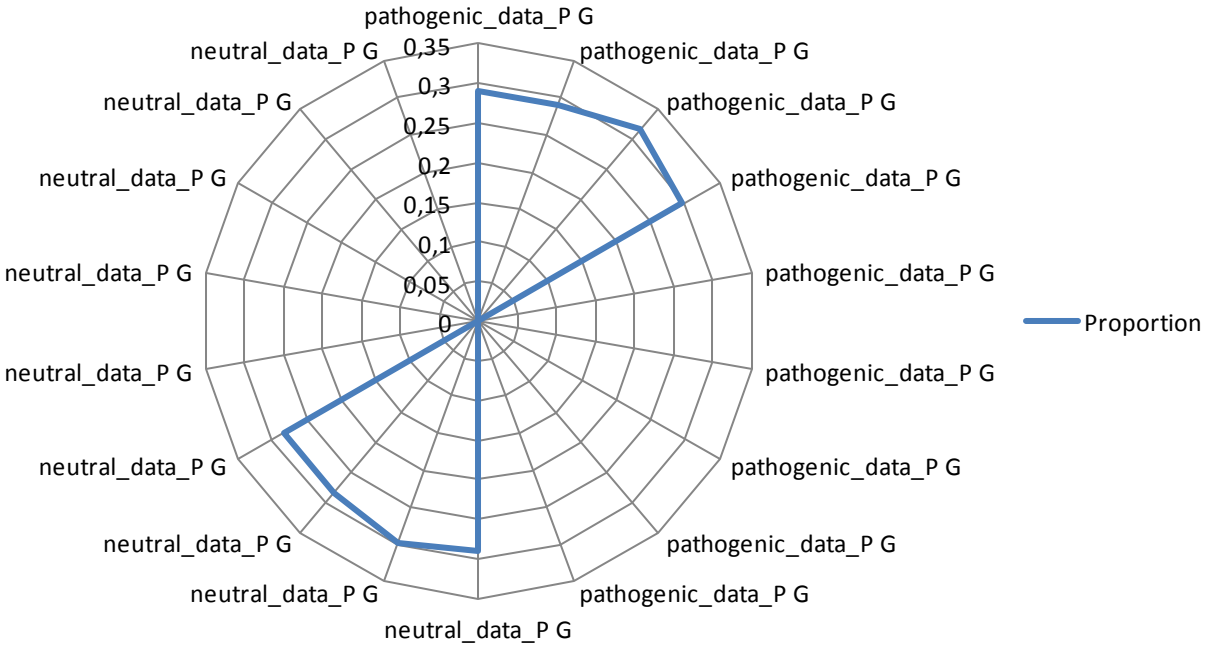
Adenine plot in Pathogenic vs Neutral datasets Palindromic Repeat from position 9



Cytosine plot in Pathogenic vs Neutral datasets Palindromic Repeat from position 9



Guanine plot in Pathogenic vs Neutral datasets Palindromic Repeat from position 9



Thymine plot in Pathogenic vs Neutral datasets Palindromic Repeat from position 9