5-2023

Clustering and analysis of g quadruplex sequences.

Aryan Neupane
University of Louisville

CLUSTERING AND ANALYSIS OF G QUADRUPLEX SEQUENCES

By

Aryan Neupane
Btech, Kathmandu University, 2016

A Dissertation
Submitted to the Faculty of the
Graduate School of
University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy in
Interdisciplinary Studies:
Specialization in Bioinformatics

Graduate School
University of Louisville
Louisville, Kentucky

May 2023

CLUSTERING AND ANALYSIS OF G QUADRUPLEX SEQUENCES

By

Aryan Neupane
Biotechnology, Kathmandu University, 2016

A Dissertation Approved on

March 24, 2023

By the following Dissertation Committee

_____

Dr Eric Rouchka, Dissertation Chair

_____

Dr Riten Mitra

_____

Dr Jeffrey Petruska

_____

Dr Juw Won Park

DEDICATION

This dissertation is dedicated to the resilience of so many families that are torn apart and to those who struggle to find hope in the midst of despair.

This dissertation is dedicated to everyone who have left their home in search of learning.

This dissertation is dedicated to my parents Mr. Tilak Neupane and Mrs. Jamuna Chapagain Neupane for providing me everything.

ACKNOWLEDGMENTS

I would like to thank my professor, Eric Rouchka, DSc, for his guidance and patience. I would also like to thank, Dr. Juw Won Park, Dr Julia Chaliker and Dr Jae Hwang for their comments and assistance over the past five years. I would like to thank my dissertation committee members Dr Jeffrey Petruska and Dr Riten Mitra for their guidance over the years. I would like to thank my lab members over the years, Dr Ernur Saka, Dr Muhammed Sayed, Dr Kalpani De Silva, Muhammed Chabane, Tae Lim Kook, Swati Saha, Aachal Malhotra and Uddalok Jana.

I would also like to express my thanks to the special Muna Fuyal for your understanding and patience during those times when there was no light at the end of anything. You encouraged me and without you, I wouldn't dare dream this dream. Also, many thanks to the members of my family in Nepal, Dr. Amrit Neupane and all my friends here in US and Nepal for their support.

ABSTRACT
CLUSTERING AND ANALYSIS OF G QUADRUPLEX SEQUENCES

Aryan Neupane

March 21, 2022

G quadruplex structures are secondary structures located throughout the genome of various organisms with involvement in regulatory functions in different transcription, translation, genome stability, epigenetic regulation as well as cell division. Even with the diverse acknowledgement of G4 structure in vivo, there are no current search tools for G quadruplexes based on already identified G quadruplexes and identified families across different genomes based on sequence diversity. Construction of families of G4 sequences and identifying their polymorphisms within disease and disorders will lead to a better understanding of their functional roles and will further research into the biophysical modeling of interactions with oligonucleotide treatments of disease.

The first project aims to develop a framework for clustering G quadruplex (G4) sequences into families based on sequence, structure, and thermodynamic properties. No current search tools exist to filter G4s based on their properties, and the diversity of G4 sequences across the genome is not fully understood. To address this gap, we utilized a combination of clustering and annotation methods to identify 95 families of G4 sequences within the human genome. Profiles for each family were created using hidden Markov models, and their thermodynamic properties, functional annotations, and transcription factor binding motifs were analyzed.

The second project aims to investigate the effect of single nucleotide variations (SNVs) on G4 structures in disease contexts. Although the role of G4s in cancer and metabolic disorders are well-established, the effect of SNVs on G4s has not been extensively studied. Using the COSMIC and CLINVAR databases, we identified over 37,000 G4 SNVs and analyzed their effects on G4 secondary structures. We found that a significant proportion of SNVs result in G4 loss or gain, and we identified genes enriched for destabilizing SNVs in G4-forming regions. We also analyzed mutational patterns in the G4 structure and found a higher selective pressure on the coding region of the template strand. Our findings provide insights into the effects of SNVs on G4 structures and highlight potential targets for therapeutic intervention in diseases associated with G4 dysregulation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

G-quadruplexes are alternative nucleic acid structures of DNA or RNA with multiple stacked guanine bases held together by hydrogen bonds to form a structure that is stabilized by the presence of a cation (1, 2). G-quadruplexes can adopt a variety of conformations, and the specific arrangement and length of the guanine bases can affect the stability and function of the structure.

Identified across variety of biological roles including transcription regulation (3), DNA replication (4), telomere maintenance (5, 6). Despite extensive work through computational tools have been used to predict the formation of stable G quadruplex across different genomes, the full spectrum of diversity and role of G quadruplex in regulation is yet to be uncovered.

1.1 Motivation

Typically, prediction of G quadruplex sequence is carried out to address the question of if the query is a G quadruplex or not through a specific pattern, or the presence of specific G and C repeats from the primary sequence structure which allow for the prediction if a sequence.

Although many tools have been developed to carry out identification, the methods and results have some limitations. G quadruplex sequences carry a diverse arrangement of sequences despite majority of sequences being a guanine. These tools only identify

Neupane, A.; Chariker, J.H.; Rouchka, E.C. Structural and Functional Classification of G-Quadruplex Families within the Human Genome. Genes 2023, 14, 645. https://doi.org/10.3390/genes14030645

sequence with proximity to form a sequence but do not identify patterns through these sequences to identify similar sequences in different genomes.

The aim of this dissertation is to make significant contributions to the field of G-quadruplex biology through the development of computational methods for the identification of G-quadruplex forming sequences and their associated families. This study explores the role of single nucleotide variants in G-quadruplex regions in disease samples, using the COSMIC and CLINVAR databases as a resource. Furthermore, this dissertation presents a new tool for the prediction of G-quadruplex forming sequences and their classification into families.

In Part I of this dissertation, we address the need for a method to identify the patterns present in G-quadruplex forming sequences. We discuss the challenges associated with the analysis of these complex structures and propose a novel computational approach for the identification of G-quadruplex families.

In Part II, we highlight on the role of single nucleotide variants in G-quadruplex regions in disease samples. The purpose of the study is to identify disease specific region and genes based on the specificity of G quadruplex.

In Part III, we present a new tool for the prediction of G-quadruplex forming sequences in genomic data. This tool classifies G-quadruplex sequences into families, based on their sequence features, and provide a valuable resource for researchers studying the biology of G-quadruplexes.

Neupane, A.; Chariker, J.H.; Rouchka, E.C. Structural and Functional Classification of G-Quadruplex Families within the Human Genome. Genes 2023, 14, 645. https://doi.org/10.3390/genes14030645

1.1.1 Part I: Identification of families of G Quadruplex sequences

The overall goal of part I is to build specific families of G quadruplex sequences based on homologous sequence data, conservation patterns, functions, and expression patterns in model organisms. Analyzing the G4 patterns across human (Hg38) genome, we provide a rationale for identification of sequence characteristics and functional annotation of G quadruplex sequences.

These structures have been shown to play a role in various biological processes, including transcription regulation and DNA replication. In addition, G4 DNA has been implicated in several diseases, including cancer, and is therefore of significant interest for the development of novel therapeutic strategies. More recently, the potential role of G4 DNA in cancer has been the subject of much research, with studies suggesting that the formation of G4 DNA in oncogenes and tumor suppressor genes could potentially contribute to the development of cancer.

With more than hundred thousand pG4 identified in the human genome(7), majority of these sequences are of length 15-35bp, and formation of G4 DNA structures is influenced by several factors, including the sequence and length of the guanine-rich region, as well as the presence of cations such as potassium and sodium. The diversity of the sequence is aided by short linker nucleotide bases of length 1-10bp. The ability to selectively stabilize G4 DNA structures has been the subject of much research, as it could potentially be used to modulate their function in vivo. Identification of these patterns and clustering of G4 sequences can provide invaluable support needed to understand the molecular mechanisms and biological functions carried out by this structure across different genomes.

1.1.1 Part II: Identification and Effects of SNVs in G quadruplex in Disease samples

Despite efforts on experimental identification of structure on individual sequences, the diversity of G quadruplex sequences is still an enigma. Studies have linked point mutations within G quadruplex to destabilize the G4 structure. We analyze large diversity in G4 sequences, through the lens of single nucleotide variants identified experimentally in cancer and disease samples. Destabilization of G4s is likely to disrupt functional association with several proteins and other transcription factors causing instability in their functions. Previous studies have highlighted the changes in loops of G quadruplexes and stability led to a significant alteration in gene expression among individuals further fueling the structural role of G4s in regulation and binding of transcription factors. Relying on experimental data, we aim to provide insight into relating newly identified G quadruplexes or G quadruplexes for a certain purpose as aptamers. These G quadruplexes are being used for a molecular carrier for the delivery of various ligands at the target site. Experiments indicate various G quadruplexes show great binding strength and lower the ligand's cytotoxicity towards non-malignant cells. Disease based association of G quadruplex can help identify the target site for binding and therapeutic uses.

1.2 Dissertation Contributions

To achieve the aims, we developed the following framework.

1.2.1 A framework for identification of G-Quadruplex families

G4 sequences are used as a novel target for various molecules with the ability to modulate gene expression. Any sequence variation in the G4 disrupting G-tracts or change in the loop composition or length affect G4 formation, topology and subsequently

their functional roles. We identified putative G quadruplex in the human genome and identify clusters of G4 utilizing existing DNA clustering tools. Because of the short and diverse nature of sequences, additional methods are required to filter redundant clusters. Based on the identified clusters we annotate the analyzed sequence characteristics including electrostatic potential, groove width, TF binding sites, presence in proximity of a gene and region specific to coding regions, among others. We develop clusters demonstrating a functional relationship between G quadruplex sequences having similar properties including sequence homology which was used to construct the Pu27 family.

1.2.2 Identification of SNVs in G-Quadruplex sequences

Point mutations within a G quadruplex can potentially destabilize the G4 structure. SNPs within or near G quadruplexes that have been associated with disease phenotypes has been studied. We aim the identification of SNPs across G4 regions will provide an insight into differences in folding properties based on sequence and external factors. Utilizing COSMIC and CLINVAR database for the identification of putative G4 regions, we investigate the biological, cellular, and molecular functions based on the folding energies of each G4 by different SNV to predict the role of G4 structural integrity in relation with diseases.

1.3 Dissertation Outline

The dissertation is organized as follows. Chapter 2 introduces G quadruplex and clustering techniques. Chapter 3 introduces the proposed method for identification of clusters of G quadruplexes. This chapter provides the analysis of annotated clusters as families and provide a prediction tool using hidden Markov models. Chapter 4 presents the description of identified single nucleotide variants in G quadruplex utilizing available

resources. Chapter 5 provides the description of tool “G4-samuha” based on the identified models of G quadruplex families. This tool allows to search for putative G quadruplex sequences based on similarity. Chapter 6 provides the conclusions and potential future work.

CHAPTER 2 BACKGROUND AND LITERATURE REVIEW

2.1 Introduction to Nucleic acids

Nucleic acids are complex organic molecules that play a crucial role in the biology of all living organisms. They are made up of long chains of nucleotides joined together by covalent bonds.

2.1.1 DNA and RNA

There are two main types of nucleic acids: DNA and RNA. DNA, or deoxyribonucleic acid, is the genetic material of living cells, and it carries the instructions needed for the cell to function and reproduce. RNA, or ribonucleic acid, plays a variety of important roles in the cell, including serving as a template for the synthesis of proteins. Each nucleotide contains a nitrogenous base with sugar molecule (ribose or deoxyribose) attached to a phosphate group as a backbone (Figure 2-1). Possible nitrogenous bases include pyrimidines cytosine, thymine (in DNA), uracil (in RNA) and purines, adenine and

guanine.

2.1.2 Base pairs

The Watson-Crick or canonical rule for base pairing in nucleic acids states that adenine (A) always pairs with thymine (T) or Uracil (U) in RNA and cytosine (C) pairs with guanine (G). This binding pattern allows for the stable, double stranded helical structure of DNA (9), and hairpin loops(10), stem-loop structures, pseudoknots (11), and more

Neupane, A.; Chariker, J.H.; Rouchka, E.C. Structural and Functional Classification of G-Quadruplex Families within the Human Genome. Genes 2023, 14, 645. https://doi.org/10.3390/genes14030645

complex tertiary structures of RNA (12).

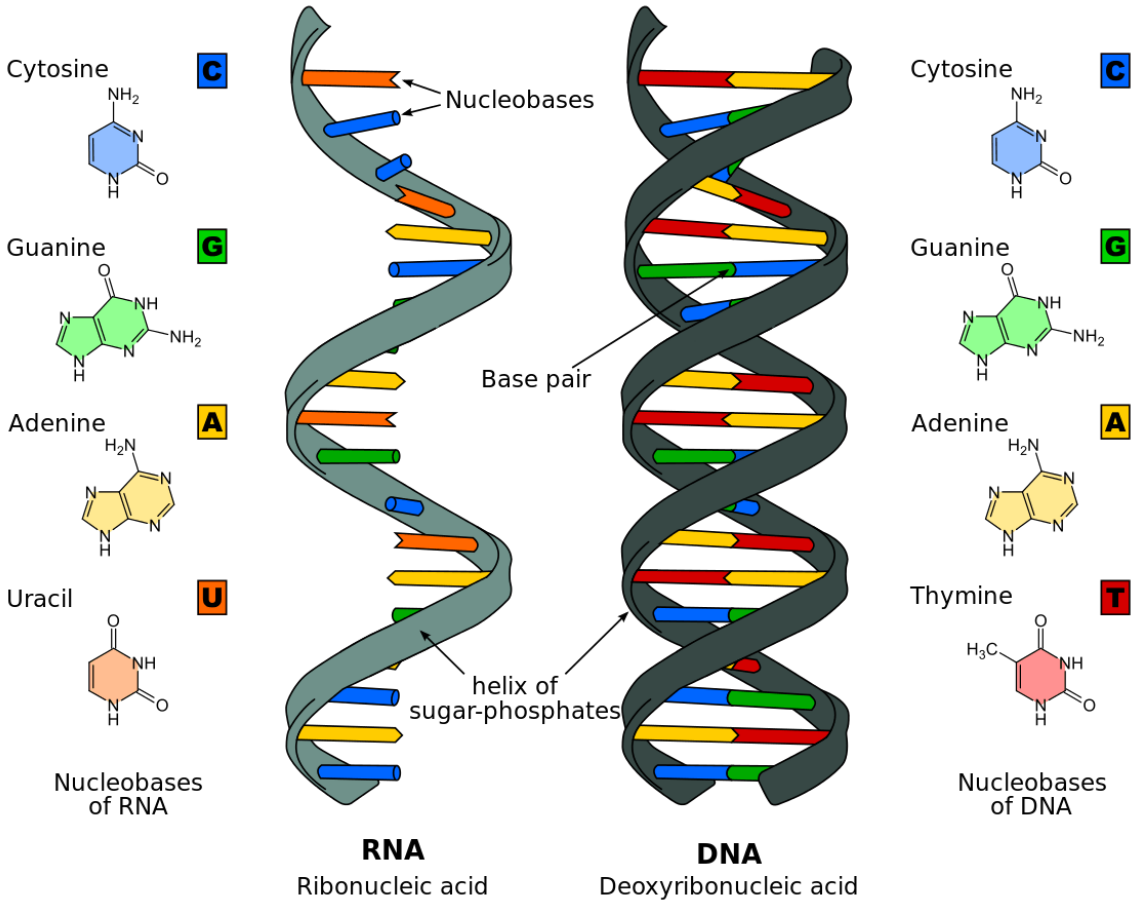


Figure 2-1 Structure and chemical composition of DNA and RNA (Image from:8) (CC-BY-4.0)

The double helix structure of DNA is stabilized through the Watson-Crick base pairing between the complementary nucleotides on each strand wound around each other. The two strands are held together by hydrogen bonds between the bases (

Figure 2-2).

2.1.3 Base stacking

Base stacking refers to the way in which the bases (nucleotides) in a molecule of DNA or RNA are arranged on top of one another. The bases are relatively hydrophobic and since

Neupane, A.; Chariker, J.H.; Rouchka, E.C. Structural and Functional Classification of G-Quadruplex Families within the Human Genome. Genes 2023, 14, 645. https://doi.org/10.3390/genes14030645

they are flat structures, they stack on top of each other to maximize the hydrophobic surface causing the twist leading to a helical structure (9) Sequence or base pairing interaction in the secondary helix structure determines the stacking energy of a double stranded DNA. The energy required to melt the double strand DNA can be calculated using the sequence and base pair interactions..

Table 2-1 lists the stacking energies in a B DNA helix for all the dinucleotide combinations. It has been established that pyrimidine-purine dinucleotide has the least energy (requiring least energy to melt) and the GC dinucleotide require the highest energy to melt.

Another way to measure the stacking of bases is the propeller twist. When the two bases do not line up perfectly in a pair, the angles created by the planes of the two bases is measured as the propeller twist. Higher propeller twist generally indicates rigid helix structure (13).There are several different types of helical structures that have been identified in DNA molecules, including the A form, B form, and Z form.

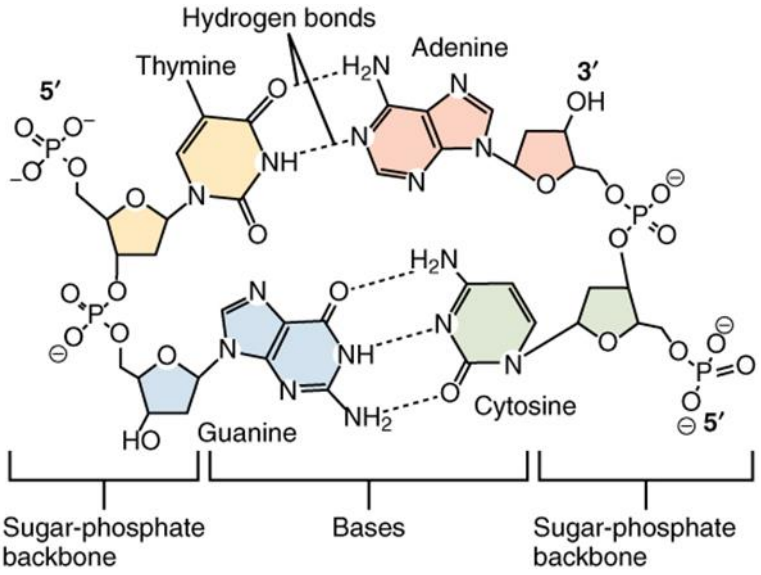


Figure 2-2 Structure of DNA nucleotides in DNA structure Image from:(1) (CC-BY-4.0)

2.1.4 Different forms of DNA (A-DNA, B-DNA, Z-DNA)

A-DNA, discovered by Rosalind Franklin is a right-handed helix structure formed under dehydrating conditions (2). Smaller rise per turn causes A-DNA to be shorter than the predominant double helix B-DNA (14). The A form helix is a type of helical structure found in DNA molecules, characterized by a wide, rigid, flat shape, with a pitch (the distance between successive turns of the helix) of about 28° and a rise (the distance along the helix axis between successive base pairs) of 0.26 nanometers. The rigid wide shape due to the off-center stacking in A DNA makes them less flexible as compared to B DNA. The B form helix is another common type of DNA helical structure (15). The B helix has a slightly narrower and more compact shape, with a pitch of about 34° and a rise of 0.34 nanometers. The B form helix is thought to be the most common form of DNA found in cells and is thought to play a key role in the process of DNA replication. The Z form helix is a less common type of helical structure found in DNA molecules (16). It is characterized by irregular, zigzag, highly twisted, left-handed shape, with a

pitch of about 45° and a rise of 0.45 nanometers and can be formed mostly in alternating purine-pyrimidine tracts.

2.1.5 Non-Canonical Structures

In addition to the Watson-Crick base pairing rules, there are also non-canonical or non-standard base pairing interactions that can occur in nucleic acids. These non-canonical base pairs are not as common as the Watson-Crick base pairs, but they can still play important roles in the structure and function of nucleic acids.

Examples of non-canonical base pairs include 1) Hoogsteen base pairing (17), in which the edges of the base are involved in hydrogen bonding, instead of the usual Watson-Crick face-to-face binding, 2) Wobble base pairing (18–20), in which a single nucleotide base can pair with more than one type of complementary base. This can occur, for example, when a guanine (G) base pairs with either a cytosine (C) or a uracil (U) in RNA, 3) triplex-forming oligonucleotides (TFOs) (21) , which can form stable, three-stranded structures with DNA or RNA through non-canonical base pairing interactions, 4) Reverse Hoogsteen base pairing(22), in which the bases are flipped over 180° and the edges of the base are involved in hydrogen bonding, instead of the usual Watson-Crick face-to-face binding.

Non-canonical base pairing can play important roles in regulating gene expression and in the formation of specialized structures within nucleic acids, such as G-quadruplexes and i-motifs. We discuss G-quadruplex structure later in detail.

A summary of differentiation based on structural parameters for duplex and quadruplex DNA is presented in Table 2-2 A summary of structural parameters for duplex and Quadruplex DNA (23, 24).

Table 2-1 Dinucleotide values for helical parameters for B-DNA

Dinucleotide step	Stacking energy (kcal mol–1)	Twist angle (°)	Propeller twist (°)
AA	-5.37	35.6	-18.66
AC	-10.51	34.4	-13.1
AG	-6.78	27.9	-14
AT	-6.57	32.1	-15.01
CA	-6.57	34.5	-9.45
CC	-8.26	33.7	-8.11
CG	-9.61	29.8	-10.03
CT	-6.78	27.9	-14
GA	-9.81	36.9	-13.48
GC	-14.59	40	-11.08
GG	-8.26	33.7	-8.11
GT	-10.51	34.4	-13.1
TA	-3.82	36	-11.85
TC	-9.81	36.9	-13.48
TG	-6.57	34.5	-9.45
TT	-5.37	35.6	-18.66
Average	-7.92±2.57	35.7±8.0	-12.60±3.2

Table 2-2 A summary of structural parameters for duplex and Quadruplex DNA (23, 24)

Structural type	B-DNA	A-DNA	Z-DNA	Quad-parallel	Quad-anti-parallel
Rise(A˚)	3.4	2.9	3.7	3.13	3.3
Twist	36.7	32.7	-30	30	30
Groove width (A˚)	11.7/5.7	2.7/11	8.5	10.2	12 8.9/12.2
Strand polarity	+-	+-	+-	++++	+-+-,+---,++--
Helix	RH	RH	LH	RH	RH
No. of bases per turn	10.5	11	12	12	12
Base pair tilt (˚)	-6	20	7	NA	NA
Width(A˚)	18	26	23	21–23	
C10-C10	10			16	16
Sugar pucker	C20	C30	C20/C30	C20	C20

Groove width: Backbone phosphate i and the iþ3 phosphate on the opposing strand.

While helical structures are the most common type of nucleic acid structure, G-quadruplexes have been found to play important roles in various biological processes, such as gene regulation and telomere which is discussed further. Despite their potential

importance, G-quadruplexes are less well understood than helical structures, and there is still much to be learned about the differences between these two types of structures and their roles in biology. Understanding the structure, stability, and function of both helical and G-quadruplex structures is therefore an important area of study in bioinformatics and molecular biology.

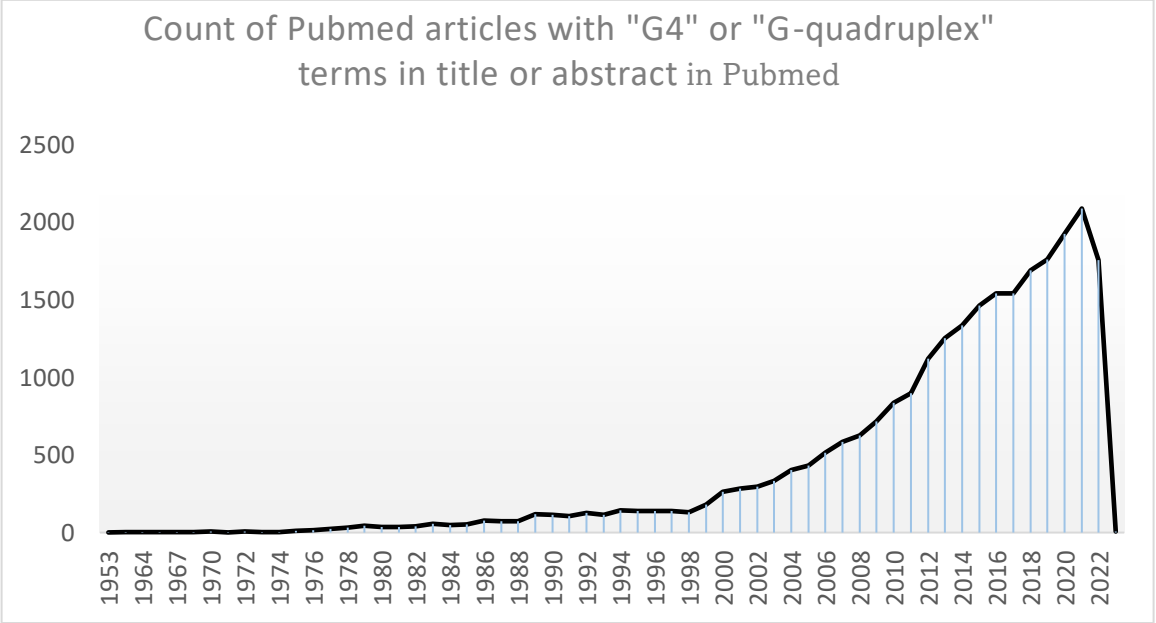


Figure 2-3 : Count of pubmed articles published with “G4 quadruplex” or “G4” and “miRNA” in title or abstract

2.2 G quadruplex sequences

G-quadruplexes are alternative nucleic acid structures that are formed by guanine-rich sequences. In a G-quadruplex, four strands of DNA or RNA are held together by hydrogen bonds to form a structure that is stabilized by the presence of multiple guanine bases. G-quadruplexes can adopt a variety of conformations, and the specific arrangement of the guanine bases can affect the stability and function of the structure. The base pair

angle and groove size in a G-quadruplex can also vary depending on the specific conformation of the molecule and environmental (25).

G-quadruplexes have four runs of at least two to three guanines separated by short stretches of other bases. These sequences can fold into a four-stranded structure with the guanine tetrads stacked on top of each other, held together by mixed loops of DNA that have the nucleobases on the inside and the sugar-phosphate backbone on the outside. The binding energy of G-quadruplexes is due to hydrogen bonding between the guanines, called Hoogsteen base pairing, which is stabilized by π- π interactions and charge interactions between the sixth position of oxygen (O6) and cations (such as K+ and Na+) between the stacks(26). G-quadruplexes can form different topologies based on various factors, including the size and orientation of the looping nucleotide bases and the direction of the sequence. pG4s have been found in the proximity of oncogenes that have a role in regulation. This implies that the location of PG4s do not occur randomly but have a functional role. G-quadruplex formation requires the destabilization of the B helix structure of DNA. A Transcription bubble is formed due to which regions of positive and negative supercoiling moves in either of the direction and G-quadruplex forming putative sequences change their confirmation (27, 28).

Based on the location of the formation of G-quadruplexes, the activity may differ. When present upstream of the TSS, the G4s may have a positive or negative effect based on the ability to interfere with RNA polymerase, Transcription factors and other binding proteins. If formed downstream of the TSS on the coding strand, transcription re-initiation may be aided by opening the confirmation of DNA or affinity to specific TFs may be increased but it may impede DNA polymerase movement which can cause transcriptional

repression by the quadruplex in the template strand (29) (Figure 2-4). G-quadruplexes can adopt a number of different topologies, including diagonal and lateral structures, depending on factors such as the size and orientation of the looping nucleotide bases and the direction of the sequence (30). The sequence can be classified as parallel, anti-parallel, or a mixed "3+1" hybrid based on its orientation. The structural architecture of G-quadruplexes is highly diverse and can vary in terms of the number of strands, the number of loops, and the orientation of the strands. Parallel structures have strands that run in the same direction, while anti-parallel structures have strands that run in opposite directions. Hybrid structures, also known as "3+1" structures, have three parallel strands and one anti-parallel strand. (31). It is also worth noting that G-quadruplexes can adopt different conformations depending on the specific sequence and the presence of specific ions or ligands. The conformation of a G-quadruplex can influence its stability and function and may be an important factor in its ability to interact with other molecules or participate in various biological processes.

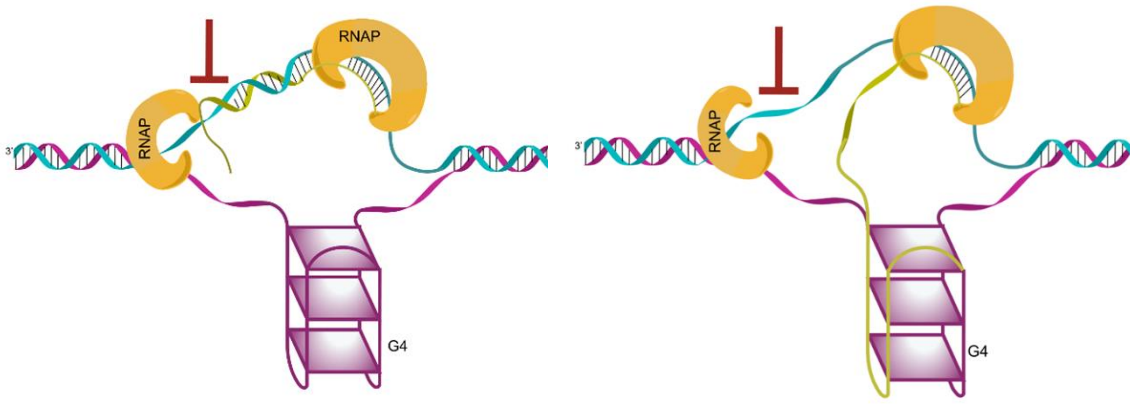


Figure 2-4 Formation of G Quadruplex Structures: Intramolecular and Intermolecular Interactions of G4 Motifs During Transcription. The presence of G4 motifs on either strand within the transcribed region can result in the co-transcriptional formation of G quadruplex structures that can physically interfere with RNA polymerase movement. Additionally, the synthesized RNA transcript can pair with the non-template DNA strand

to form a heteroduplex structure known as an RNA loop, which can then fold into an intermolecular DNA:RNA hybrid G quadruplex.

G-quadruplexes, or four-stranded secondary structures in nucleic acids with multiple repeats of guanine, were first identified in 1962 by Gellert et al. (32). These structures have now been found to play important roles in various biological processes, including gene expression regulation, through the involvement of various G-quadruplex binding proteins (33)(34). A number of techniques, such as circular dichroism, spectroscopy, nuclear magnetic resonance spectroscopy (35), x-ray diffraction (36) and fluorescence spectroscopy (37) have been used to study the folding and formation of G-quadruplexes in vitro. CD spectroscopy is useful to differentiate all parallel structures from anti parallel structures. However, these techniques do not provide a complete understanding of G-quadruplex formation and stability in vivo. Biffi et al. were able to visualize DNA G-quadruplex structures in the genomic DNA of mammalian cells through nuclear staining with a single chain phage display antibody, BG4 (38). More recently, advances in techniques such as Ribo-seq (39) and next-generation sequencing, as well as the identification of specific fluorescence probes, have allowed for the identification of G-quadruplexes in various sequences in vivo (40).

716,310 G4-forming sequences stabilized by G4 ligand pyridostatin (PDS) and 525,890 G4-forming sequences stabilized by K+ were identified in the human genome by combining polymerase stop assay with Illumina next-generation sequencing (G4-seq) (41). Prokaryotic(42) and eukaryotic(43) genomes have large number of putative G-quadruplexes located all over the genome in various locations. Many sequences have been identified near the Transcriptional start site (TSS) in their promoters implying regulatory

function of the formation of G4s in gene expression. Also, it is now known that about 43% of the genes contain at least a putative G-quadruplex within 1 kb upstream of the TSS (44). It has been found that oncogenes are more likely to contain G rich sequences whereas tumor suppressors have less amount of G rich sequences (45). Oncogenes with G-quadruplexes in their proximity could be affected by the destabilization of the sequences.

2.2.1 Pu27 family

The Pu27 family is a group of 17 putative G-quadruplex-forming DNA sequences that are homologous to the Pu27 genomic (Pu27ge) sequence associated with the promoter region of the human c-Myc gene. The c-Myc gene is regulated by a region known as the nuclease hypersensitive element (NHEIII1), which is located -115 bp upstream of the P1 promoter and has the ability to form i-motif structures. The complementary G-rich non-coding strand of this region is able to form a G-quadruplex structure. Members of the Pu27 family are able to bind specifically to the parent Pu27 target sequence and are found throughout the human genome on different chromosomes. It has been shown that each sequence of the Pu27 family is able to form a stable G-quadruplex structure and is able to bind in a sequence-specific manner to the NHEIII1 region of the c-MYC promoter, repressing the expression of the gene and inhibiting cell growth. G-quadruplex sequences have also been found in the mRNA of certain genes and may bind to these sequences to further stabilize them, acting as an "off switch" for transcription. In some cases, the expression of G-quadruplexes in the untranslated regions (UTR) of transcribed genes such as SOX2, NAV2, and SPTLC2 may be relative to their presence in different cell or tissue types. It has also been suggested that the c-MYC transcribed mRNA may regulate further transcription in a DNA-RNA "back-loop" mechanism by binding to the complementary NHEIII1 sequence

in the c-MYC promoter. Enrichment of certain transcription factors has also been observed at putative G-quadruplex locations (47)

c-MYC transcription is regulated by a region -115 bp upstream of the P1 promoter known as nuclease hypersensitive element (NHEIII1). The region is known to form i-motif structures and the complementary G rich noncoding strand forms the G quadruplex structure.

2.3 Computational methods associated with G-quadruplex

Computational methods for identifying and conserving G-quadruplexes are diverse and can be categorized into several groups. Sequence-based methods predict G-quadruplex formation based on sequence parameters solely, mostly analysing repeats of guanine interspersed by loop region with variable length and bases in between. Similarly, Structure-based methods predict G-quadruplex formation based on the hydrogen bond for Guanine bonding (base pairing) required fold the DNA sequence into a quadruplex structure. One example is the ViennaRNA Suite (RNAfold) (3) that uses RNA secondary structure prediction and analysis to predict G-quadruplex secondary structure and thermodynamic profile. Additionally, Hybrid methods combine the sequence and structure-based method with machine learning methods

2.3.1 Tools used to identify G-quadruplex

Several computational approaches for detecting putative G-quadruplex structures within sequence data have been constructed, including Quadparser (4), G4Hunter (5), G4HMM (6), QGRS mapper (7), G4P (8), Quadpredict (9) and QuadBase (10). Quadparser considers sequence-based approach with repeats of three G's with a loop length of 1 to 7 in between with various similar approaches being taken in other tools as well varying the number of guanine repeats and loop length. QGRS mapper takes a variable number of G and the loop length providing a G score as the likelihood score for the formation of G4s. These provide an overview of the possibility of a sequence being able to form a G quadruplex or not.

Various sequences identified by the tool is enriched in the human promoter regions. G4P Calculator evaluates the G4 DNA potential percentage which depends upon the runs of guanines in a sliding window. The final score is the percentage of the 'hit' in all the windows searched (8). G4Hunter is based on calculating the skewness of G and C by associating runs of G with positive score and presence of C penalized. A probability score to form G quadruplex is associated with the final score. For the prediction of G quadruplexes, G4Hunter is applicable for RNA or DNA sequences with no reference to its complementary strand. The nucleotides in the loop region are also not taken into any account.

Another tool, QuadBase offers the analysis of G quadruplexes across various species across prokaryotes and eukaryotes (10). Pqsfinder, based on flexible folding rule on G quadruplex is trained using 392 in vitro experimentally validation sequences (11). Another tool for identification of G4 RNA, G4RNA screener is based on combining different scores given by different tools, cGcC score, G4 Hunter and G4NN into a single tool (12). G4NN, a part of G4screener, uses a k-mer of 3 as input through a feed-forward single-layer neural network that learns from sequences available on G4RNA database (149 G4 and 179 non-G4). Diving into predicting and characterization of different types of G quadruplex is necessary as the pattern of working of these sequences in different genome seems to be different. Trinucleotide composition of nucleotides is used as input, but accurate training of G quadruplex model would require much more data and more layers to the network. The Quadron algorithm is based on a tree-based gradient boosting machines using more than 200 sequence and structure-based features trained from over 700,000 sequences in vitro G4-formation dataset which was obtained for the human genome using the G4-seq methodology, and specifically for DNA G4s in this case (13). Another R based tool, G4-iM Grinder is designed to identify and analyze G-quadruplex

sequences in DNA and RNA using a flexible folding rule to predict the formation of G-quadruplexes based on the presence of G-rich stretches and the potential for Hoogsteen base pairing (14). The tool also incorporates several qualification functions, including scoring systems like G4hunter, cGcC, and PQSfinder, which help limit the sequences after a search. These functions evaluate the likelihood of quadruplex formation and can be used to calculate a quantitative interest score for each sequence. G4-iM Grinder also includes functions for quantifying predefined patterns and localizing known-to-form and known-NOT-to-form quadruplex sequences. The resulting sequences can then be prioritized for in vitro evaluation based on their scores, frequency, or other filters. G4boost, another tool utilizes decision tree-based models utilizing sequence composition, structural identify G4 motifs and predict their secondary structure folding probability and thermodynamic stability (15)

2.3.2 G quadruplex conservation

Todd and Neidle used single linkage hierarchical agglomerative clustering to create clusters based on 87,697 intronic regions on the non-template strand of the human genome, based on the pattern G3-5 L1-7 G3-5 L1-7 G3-5 L1-7 G3-5 The families were based on sequence similarity (16). Based on the position on the genome, the G4 quadruplexes may be functionally distinct. Because only intronic regions were observed, it is now well known that G quadruplexes are present in the transcriptome. Developed to identify highly conserved G-quadruplex motifs in homologous nucleotide sequences, QGRS-conserve is a computational method aimed at reducing the likelihood of false-positives when predicting G-quadruplex formation. The tool evaluates not only location conservation but also the conservation of structural features of the G-quadruplex motif, such as loop lengths, number of tetrads, and the total length of the structure. The technique allows for the filtering of

motifs based on qualitative conservation and promotes accurate wide-scale analysis of G-quadruplexes within exomes, transcriptomes, and genomes. QGRS-Conserve also presents strategies for dealing with specific challenges relating to overlapping G-quadruplex motifs and the impact they have on conservation analysis.

A study investigated the conservation and evolution of G quadruplex (G4) structures across 37 genomes, from fungi to mammals (17). The study found that G4 structures have evolved with increasing complexity in genomes and species, and that the loop length of G4 motifs plays a critical role in their stability. The study also discovered that G4 structures are enriched in transcription factors, which may be involved in a variety of biological processes. The study confirmed the existence of G4 structures in cells through immunofluorescence staining and suggested that G4 structures may have a regulatory role in gene transcription. Finally, the study found an antagonistic relationship between G4 structures and DNA methylation, which may have emerged early in evolution and been maintained throughout subsequent evolutionary processes.

2.3.3 G-Quadruplex databases

Over the years, G-quadruplex identification has garnered pace and researchers have biologically and computationally identified thousands of G quadruplex forming sequences along with hundreds of G quadruplex structures. Quadbase enables researchers to query quadruplex sequences in the genomes of prokaryotes and eukaryotes (10). It contains multiple interfaces for searching the quadruplex patterns and analyze their conservation between orthologous genes across organisms. The Pattern Search interface have been used to search quadruplex sequences across 146 prokaryotes and four eukaryotes. The Orthologs Analysis interface is designed to find PG4 motifs that are conserved across organisms. The Pattern Finder interface is a tool that enables users to find quadruplex motifs in a given sequence of interest. Another database, Greglist database provides list of all human genes that have potential G-quadruplex motifs in their promoter regions (18). The database is curated using the Quadparser.

The Nucleic acid G-quadruplex structure (G4) Interacting Proteins DataBase (G4IPDB) is an useful resource for researchers studying the interaction between proteins and G-quadruplex structures forming sequences (19) . This database contains detailed information about over 200 proteins and their interaction with G-quadruplex forming sequences, including binding/dissociation constants, interacting residues in proteins, and related PDB entries. In addition, G4IPDB provides a web-based G-quadruplex predictor tool that predicts G-score for putative G-quadruplex forming sequences. This information could be beneficial for the development of therapeutics for diseases such as cancer and neurological disorders. The G4IPDB database is expected to assist researchers in developing structure-based drug design, virtual screening, molecular dynamic simulation, and docking studies for the development of therapeutics targeting nucleic acid-based diseases.

The G-quadruplex database, G4LDB, is a collection of reported G-quadruplex ligands that stabilize G quadruplex aimed ligand and drug discovery (20). G4LDB compiles a data set that covers various physical properties and 3D structures of G-quadruplex ligands, provides web-based tools for G-quadruplex ligand design, and facilitates the discovery of novel therapeutic and diagnostic agents targeting G-quadruplexes. The database currently contains over 800 G-quadruplex ligands with approximately 4000 activity records, The new version, G4LDB 2.2 includes over 3200 G4/iM ligands, 28,500 activity entries, and 79 G4-ligand docking models (21). The database also provides an online docking module. Studies on G-quadruplex ligands are at the forefront of drug discovery, and a comprehensive database such as G4LDB will benefit such studies.

2.4 Single Nucleotide Variants

Genetic variants are the differences in DNA sequences among individuals in a population. SNVs are alterations of a single nucleotide base at a specific position in the DNA sequence. These variations can occur in different forms, including single nucleotide variants (SNVs), indels, copy number variations (CNVs), and structural variations (SVs). SNVs can occur

in both protein-coding and non-coding regions of the genome. In protein-coding regions, SNVs can lead to changes in the amino acid sequence of the resulting protein, potentially impacting protein structure, stability, and function. On the other hand, SNVs in non-coding regions can also have significant effects on gene expression (22, 23), transcription factor binding(24, 25), and splicing (26), leading to disease development.

Recent advances in next-generation sequencing technologies have enabled the identification of numerous SNVs in the human genome. However, not all SNVs are deleterious and can cause disease. To identify pathogenic SNVs from benign ones, researchers have developed various bioinformatics tools that can predict the functional effects of SNVs. These tools use different algorithms, machine learning techniques, and features, and each has its own strengths and limitations.

SNVs are the most common type of genetic variation, and they can be classified as synonymous or non-synonymous depending on their effect on the resulting protein. Synonymous SNVs do not change the amino acid sequence of the protein, and they are usually considered neutral. In contrast, non-synonymous SNVs alter the amino acid sequence of the protein, potentially leading to changes in its function, stability, or interactions with other molecules.indels, short for insertion-deletion mutations, are genetic variations that involve the insertion or deletion of one or more nucleotides in the DNA sequence. These variations can cause frame-shift mutations, where the reading frame of the gene is altered, potentially leading to the formation of a truncated and non-functional protein. Copy number variations (CNVs) are genetic variations that involve the gain or loss of entire segments of DNA, which can range from a few nucleotides to entire genes. These variations can affect gene dosage, potentially leading to changes in gene expression levels,

and they have been associated with a wide range of human diseases, including cancer and neurodevelopmental disorders (27).

Structural variations (SVs) are genetic variations that involve large-scale changes in the structure of chromosomes, such as inversions, translocations, and deletions. These variations can affect the spatial organization of the genome, potentially leading to changes in gene expression, chromatin accessibility, and higher-order chromatin structures (28).

The impact of genetic variants on biology is complex and multifaceted. Some variants may have no effect on protein function or may confer beneficial adaptations, while others can have deleterious effects, leading to the development of various diseases. For example, non-synonymous SNVs in the BRCA1 and BRCA2 genes have been associated with an increased risk of breast and ovarian cancer (29). Deletion of phenylalanine (ΔF508) in the CFTR gene d can cause cystic fibrosis, a genetic disorder that affects the lungs (30). CNVs in the CYP2D6 gene can affect the metabolism of drugs, potentially leading to adverse drug reactions or therapeutic failure. Finally, SVs in the 22q11.2 region (31) have been associated with a wide range of developmental disorders, including DiGeorge syndrome and velocardiofacial syndrome.(32, 33)

2.5 Variant Databases

Variant databases are repositories of genetic variations that have been identified in individuals or populations. These variations can be single nucleotide polymorphisms (SNPs), insertions, deletions, or structural variations. Variant databases include both germline and somatic variations, and can be used for a variety of purposes, including identifying disease-causing mutations, understanding population genetics, and characterizing functional genetic elements.

Genome-wide association studies (GWAS) are one way in which variant databases are used. In a GWAS, genetic variations are compared between cases (individuals with a disease) and controls (individuals without the disease) to identify associations between specific genetic variants and the disease (34, 35). Variant databases are also used to annotate the functional effects of genetic variations, such as whether a variant occurs in a protein-coding region of the genome or in a noncoding region that regulates gene expression. These databases have become essential resources for the study of genetic variations and their implications in disease susceptibility and progression. With the increasing availability of high-throughput sequencing technologies and the development of bioinformatics tools, these databases have evolved into comprehensive repositories of genetic variation data. The use of variant databases has facilitated the discovery of new genetic associations and insights into disease pathogenesis, leading to the development of novel therapeutic strategies. As more data are generated and integrated into these databases, their utility is expected to continue to expand. Some examples of variant databases include COSMIC (36), CLINVAR (37), dbVAR (38), LOVD (39), HuVarBase (40), ncVarDB (41),1000 Genomes Project (42) among many others.

The study of somatic mutations in human cancer has been revolutionized by the availability of large-scale databases such as the Catalogue Of Somatic Mutations In Cancer (COSMIC). With almost 6 million coding mutations along with 19 million non-coding mutations across 1.4 million tumor samples, COSMIC provides a unique and detailed resource for exploring the impact of somatic mutations on cancer. These mutations, which can occur in any gene, can lead to a range of different outcomes, including changes in protein structure and function, alterations in gene expression, and the promotion of drug resistance.

COSMIC is a comprehensive and curated database, with data derived directly from scientific literature by expert manual curators, ensuring quality, accuracy, and descriptive data capture. In addition to coding mutations, COSMIC also covers all the genetic

mechanisms by which somatic mutations promote cancer, including non-coding mutations, gene fusions, copy-number variants, and drug-resistance mutations. Similarly, ClinVar is a public open-access database created by the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH) that archives human genetic variants and their associations with human health and disease. ClinVar aggregates data from various sources, making it a centralized resource that aids users in interpreting variants. The database also includes a powerful search tool that allows users to search for variants based on their clinical significance, such as pathogenic, likely pathogenic, benign, likely benign, or uncertain significance. ClinVar provides detailed information about each variant, including its genomic location, inheritance pattern, and clinical significance. The database also includes information about the evidence used to support each variant's clinical significance, such as data from functional studies, family segregation studies, and population genetics studies.

2.6 Unsupervised learning

Machine learning algorithms are a set of algorithms that enable computers to learn from data and make predictions or decisions without being explicitly programmed. These algorithms can be broadly classified into two categories: Supervised and unsupervised learning.

2.6.1 Introduction to Machine learning

Supervised learning involves training a model on labeled data, where the correct output is provided for each input. The goal is to enable the model to predict on unseen data(48). Commonly used supervised algorithms include decision trees, K-Nearest Neighors, Naïve Bayes, Artificial Neural network, Logistic Regression and Support Vector machine (SVM) among others. Classification, regression, structured prediction (tree or a sequence), ranking are the common problems used for supervised learning(49, 50) .

Unsupervised learning, on the other hand, relies on the inherent structure of the data to identify patterns and make predictions. One common approach in unsupervised learning is clustering, which involves the grouping of data points into clusters based on their similarity. Semi supervised leaning and Reinforcement learning are other methods using the approaches in combined manner.

2.6.2 Clustering

Clustering is a widely used technique in the field of data mining and machine learning. It involves the grouping of data points into clusters, with each cluster representing a distinct group or pattern in the data (53, 54). This technique is often used to identify underlying patterns and structures in complex datasets and can be applied to a wide range of applications, including image (55) and text analysis (56), customer segmentation (57), and outlier detection (58). One common use of clustering is data reduction, where the goal is to find a compact representation of the data that retains the important characteristics of the original dataset. This can be useful for reducing the storage and computational requirements of a dataset and can also make it easier to visualize and understand the underlying patterns in the data. In addition to data reduction, clustering can also be used for other purposes, such as identifying natural data types, finding useful and suitable groupings of data, and detecting unusual or outlying data points. it can be used to identify closely related sequences, such as members of a gene family or a species clade. The specific goals and applications of clustering will depend on the specific problem at hand, and the choice of clustering algorithm may vary accordingly.

2.6.3 Similarity/ Distance Metrics

Clustering algorithms rely on the notion of similarity or distance between objects, which is typically measured using a variety of metrics, such as Manhattan distance, Euclidean distance, Minkowski distance, Cosine similarity, Pearson correlation, Jaccard similarity, and Dice coefficient (59). The most common approach is to compute the Euclidean distance between data points and their nearest cluster and assign the point to the cluster with the minimum distance.

2.6.4 Types of Clustering

Clustering algorithms can be divided into two main categories (60): partition based and hierarchical.

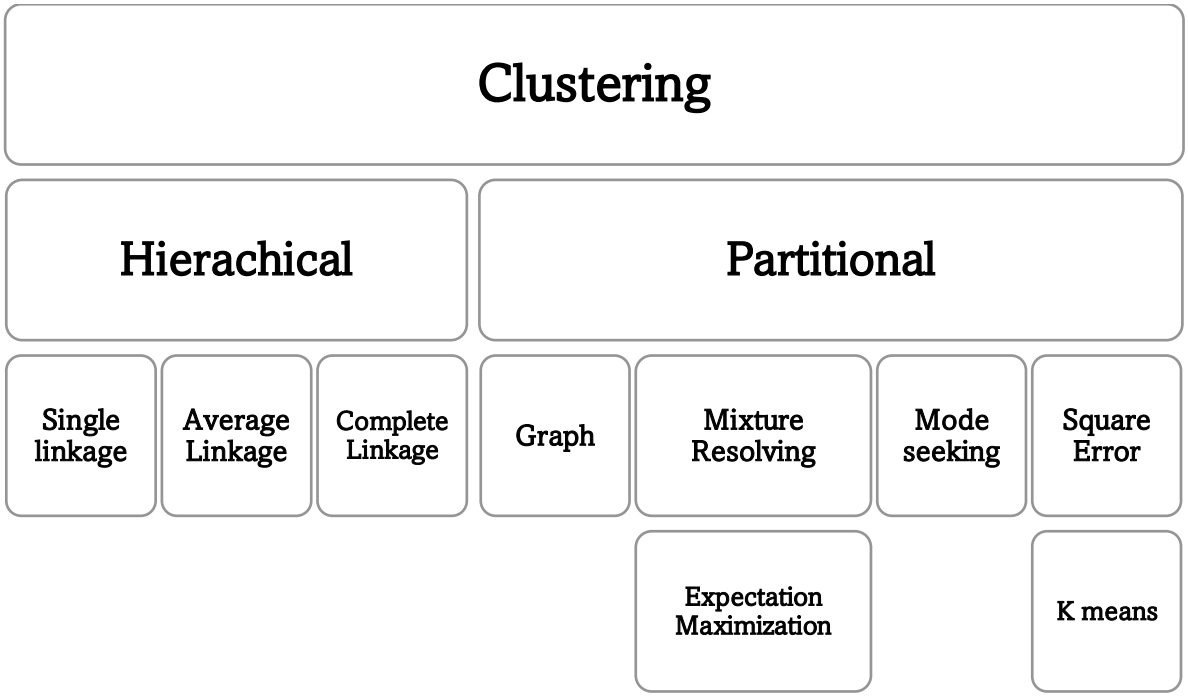


Figure 2-5 A division of different clustering algorithms (49).

2.6.5 Partition-based algorithms

Partition-based algorithms divide the data into a predefined number of clusters, using some measure of similarity (optimizing certain objective function). These algorithms can

be further divided into centroid-based and density-based algorithms, depending on the specific approach used to define and identify the clusters. K-means is one of the most common centroid based algorithms which minimizes the within group sums of squares as its optimal criteria for separating the clusters. Other examples include K-modes, PAM, CLARA, CLARANS and FCM. DBSCAN is a density-based algorithm based on differentiating the clusters based on statistical distribution of the density of each cluster with a threshold criterion (53).

2.6.6 Hierarchical clustering:

In hierarchical clustering, the algorithm starts with individual data points as clusters and iteratively combines them into larger and more inclusive clusters based on some measure of similarity. The final clusters are identified by cutting the branches of the resulting tree at a specific level. In divisive clustering, the algorithm starts with all the data points in a single cluster, and iteratively divides them into smaller and more distinct clusters based on the same measure of similarity (61).

There are many different methods for hierarchical clustering, Single linkage clustering combines clusters based on the similarity of only one element of each cluster with smallest distance, while complete linkage clustering uses the distance between the most dissimilar elements in two clusters to guide their combination. Average linkage clustering, also known as UPGMA or WPGMA, uses the average distance between the objects in two clusters to guide their combination. These two methods differ in the way the average distance is calculated. UPGMA uses a proportional averaging based on the number of elements in each cluster, while WPGMA is a simple averaging metric of elements at each step. Ward's method is an agglomerative clustering approach that seeks to minimize the

sum of squares within a cluster (62) . This criterion is also used in K-means clustering and can be used to identify the initial number of clusters for further iterative clustering (63).

Model-based algorithms which optimize the fit of the provided data and mathematical models are also used (64, 65). Mixture models hypothesize the data points provided are a mixture of probability distributions and algorithms such as MCLUST, EM, Self-Organizing Maps (SOMS)(66) are examples of such approaches to identify clusters and reducing noise.

2.6.7 Clustering in Bioinformatics

Clustering is a widely used technique in the field of bioinformatics, where it is used to group biological data into clusters based on their similarity. This technique is often applied to high-dimensional datasets, Some examples of clustering applications in bioinformatics and computational biology include gene function prediction, where the goal is to assign unknown genes to known functional categories based on their expression patterns; gene expression analysis (67), where the goal is to identify groups of genes with similar expression profiles across different conditions or treatments; and protein structure prediction, where the goal is to group protein sequences into clusters based on their structural similarity. It can also be used to identify conserved sequence motifs or regulatory regions within a set of sequences. Additionally, clustering can be used to visualize the relationships between sequences and to detect outliers or unusual sequences that may be of interest.

Evaluating the performance of clustering algorithms can be challenging, as there is no absolute "correct" way to group the data into clusters. One common approach is to use silhouette analysis, which measures the separation distance between the resulting clusters,

and provides a way to assess the quality and suitability of the clustering for a given application. Other metrics such as modularity (68), is used to select the clusters based on the intra or inter cluster distances. Overall, clustering is a powerful and versatile technique for uncovering hidden patterns and structures in complex datasets and has a wide range of applications in various fields.

2.6.8 Clustering in DNA sequences

Clustering of sequences in bioinformatics refers to grouping of biological sequences that are similar. The biological sequences include genomic, proteins, DNA, 16S ribosomal RNA, transcriptomic data. A common approach is to use a K-mer approach where kmers are taken as features for the clustering algorithm. Tools such as CD-HIT, UCLUST, DNACLUST, SpCLUST, MESHCLUST, VSEARCH (43), has been used for clustering operational taxonomic units (OTU) in microbiomics.

CD-HIT (Cluster Database at High Identity with Tolerance) is an open-source tool based on a greedy incremental heuristic algorithm of matched alignment columns and word counting avoiding the expensive memory cost of pairwise alignment. Sorted by length, the first sequence is regarded as the first cluster. Iteratively, based on the similarity threshold, additional clusters are generated as seed or added to the already defined clusters (69). UCLUST and DNACLUST are similar tools with greedy incremental heuristic algorithms but vary in the sorting (UCLUST) or the cluster representative approach (DNACLUST) to CD-HIT.

2.6.9 Markov Chains

A Markov chain, named after Andrey Markov, is a mathematical system that satisfies the Markov property, which states that, the probability of transitioning to a

particular state at any given time is dependent only on the current state and time, and is independent of the sequence of events that preceded it. This property is known as the "memoryless" property of a Markov chain. It can be thought of as the probability of transitioning to any particular state is dependent solely on the current state and time elapsed. A Markov chain is a sequence of random variables X_1, X_2, X_3, ... that satisfies the Markov property, which states that the probability of transitioning from one state to another depends only on the current state and time elapsed. This can be expressed mathematically as follows:

$$P(q_n = x \mid X_1 = q_1, X_2 = q_2, ..., X_{n-1} = q_{n-1}) = P(X_n = x \mid x_n = q_{n-1})$$

We can express the Markov chain

$Q = q_1q_2 ...q_S$ a set of s states $x = x_{11}x_{12} ...x_{S1} ...x_{SS}$ a transition probability matrix A, each x_{ij} is the probability of getting from state i to state j, such that $\sum_{j=1}^{n} x_{ij} = 1 \ \forall n$

In other words, for a first order Markov chain, the probability of transitioning to any particular state at time n+1 is dependent only on the current state at time n and is independent of the specific sequence of states that led to the current state. Markov chains have several important properties, including irreducibility (the ability to reach any state from any other state), aperiodicity (the system will return to a certain state with probability 1 after a certain amount of time), and positive recurrence (the system will visit any state an infinite number of times over time). There are various techniques for analyzing and solving Markov chains, including steady-state analysis and numerical methods. These techniques allow for the determination of long-term behavior of the system, including the probability of being in a particular state at a given time. One of the key advantages of using a Markov chain is that it allows for the prediction of future states based on current and past states.

Used to model a wide range of systems, including chemical reactions, traffic flow, communication networks and in the field of genetics, linguistics, and computer science, they have been effective in modeling systems where the future behavior is influenced by the current state, but not by the specific sequence of events that led to that state.

2.7 Hidden Markov Models

A hidden Markov model (HMM) is a statistical model that is widely used in various fields, such as natural language processing, speech recognition, and bioinformatics. In essence, an HMM is a probabilistic model that captures the time-varying behavior of a sequence of observations. At each time step, the HMM generates an observation based on a hidden state, which is not directly observable to the user. The hidden states are connected by a set of transition probabilities, which determine the likelihood of transitioning from one state to another over time.

2.7.1 Difference between Markov chain and HMM

In a Markov chain, the state of the system is fully observable at each time step. In contrast, in a hidden Markov model, the state of the system is not directly observable, but can be inferred from the observations made at each time step. The hidden variables in an HMM are referred to as "hidden states," and the observations made at each time step are referred to as "emissions." The probability of transitioning from one hidden state to another and the probability of emitting a particular observation are determined by a set of parameters known as the "transition probabilities" and "emission probabilities," respectively. The key advantages of using a hidden Markov model are that it allows for the modeling of systems in which the state of the system is not directly observable but can be inferred from the

observations made at each time step. This allows for the observations of noisy and missing data possible through the hidden states. This makes HMMs particularly well suited for modeling complex systems with hidden or unobserved variables. HMM can be thought of as a generalization of a Markov chain, where the observations are generated by the hidden states rather than being directly observable. Additionally, HMMs are computationally efficient, making them suitable for applications with large datasets. In recent years, HMMs have been applied to a variety of problems, including speech recognition, machine translation, and protein structure prediction(70). Hidden Markov models (HMMs) are a powerful tool for modeling and analyzing complex systems, such as DNA sequences. HMMs are a type of statistical model that can capture the time-varying behavior of a sequence of observations, such as the sequence of nucleotides in a DNA molecule. In an HMM, the observed data is generated by a set of hidden states, which are not directly observable to the user. The hidden states are connected by a set of transition probabilities, which determine the likelihood of transitioning from one state to another over time.

2.7.2 Multiple Sequence Alignment

Multiple sequence alignment (MSA) is a method used to align multiple biological sequences, such as DNA, RNA, or protein, in order to compare their similarities and differences (71). MSA is a widely used tool in bioinformatics applied to including phylogenetic analysis, protein structure prediction, and functional annotation. In an MSA, the aim is to align the sequences in such a way that the regions of similarity are maximized, and the differences minimized (72). This is typically done by introducing gaps into the sequences in order to align them properly. There are several approaches to constructing MSAs. One common approach is to perform pairwise alignments and add

proteins to the alignment based on a guide tree, pairwise similarity scores, or Monte Carlo optimization. Another approach is to align all proteins to a pivot structure, which might be a consensus structure or a chosen representative structure. There are a number of tools available for multiple sequence alignment, including Clustal (44), MUSCLE , and MAFFT (73). These tools use different approaches and may be more or less suitable for different types of sequences or alignments depending on the degree of similarity and length of sequences. Clustal is a popular multiple sequence alignment tool that uses a progressive alignment algorithm to align sequences. It starts by aligning the most similar sequences and then gradually adds more sequences to the alignment based on their similarity to the already aligned sequences. MUSCLE (74) is another multiple sequence alignment tool that uses a heuristic approach to align sequences. It starts by aligning the most similar sequences and then iteratively refines the alignment by aligning the sequences in small groups and then combining the alignments. MAFFT (73) is a multiple sequence alignment tool that uses a combination of progressive and iterative approaches to align sequences. It first aligns the most similar sequences and then uses an iterative refinement process to improve the alignment.

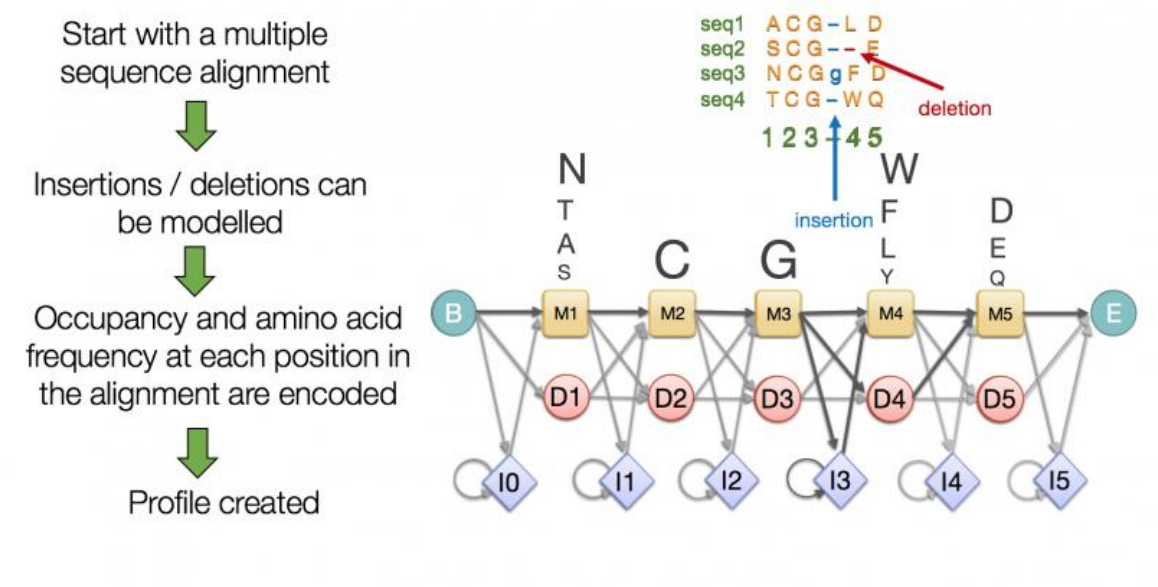


Figure 2-6 Profile HMM utilizing a multiple sequence alignment (45) (CC-BY-SA-4.0)

2.7.3 Profile Hidden Markov Models

In the context of multiple sequence alignment, an HMM can be used to model the probability of observing a particular sequence of amino acids or nucleotides in a protein (75) or DNA sequence, given the underlying hidden states that represent the alignment of the sequences. Profile HMMs (76) and profiles (77, 78) are both used for multiple sequence alignment, but there are some differences between them. One major difference is that in a profile, the penalty for gaps or insertions is the same in every position of the alignment, regardless of the level of variability in that position. In contrast, in a profile HMM, the penalties for gaps or insertions are position-desspendent and are learned from the training data. This means that positions that are more variable may have a smaller penalty for gaps or insertions compared to more conserved areas.

Additionally, profile HMMs consist of a sequence of match states, which are analogous to positions in a multiple sequence alignment, and corresponding insert and delete states. Each insert and match state has a probability distribution over amino acids, which gives

the probability of a particular amino acid given that state. The parameters of the profile HMM model are the probabilities for transitions between states and the amino acid probability distributions, and these are optimized to give high probabilities to sequences belonging to the modeled family and low probabilities to other sequences.

The HMM architecture consists of a set of hidden states, which represent the different possible alignments of the sequences, and a set of observation symbols, which represent the different nucleotides symbols (ACGT, K=4) that can be observed in the sequences. The HMM is defined by a set of transition probabilities, which specify the probability of transitioning from one hidden state to another, and a set of emission probabilities, which specify the probability of observing a particular symbol given a particular hidden state. In a profile hidden Markov model (HMM) for multiple sequence alignment, the hidden states can represent the different possible alignments of the sequences at a particular position. One common way to represent these alignments is to use three types of hidden states: match states, delete states, and insert states. A match state represents a position where all the sequences are aligned, and a nucleotide is observed in each sequence. A delete state represents a position where one or more of the sequences has a gap, or "deletion," in the alignment. An insert state represents a position where one or more of the sequences has an extra nucleotide that is not present in the other sequences (79, 80) .

Transition probabilities t_{ij} are defined for each state i moving to state j and self with the transition probabilities equalling 1 with sum of $\Sigma_j t_{ij}$ and emission probabilities $e_i(x)$, such that $e_i(x)$ over all K symbols in each state i is $\Sigma_x e_i(x) = 1$, to create an ensemble of multiple HMMs for multiple families of G quadruplex identified as discussed.

The hidden states can be represented mathematically as follows:

Match state: Mi = "match"

Delete state: Di = "delete"

Insert state: Ii = "insert"

The observation symbols O = [O_1, O_2, ..., O_l] represent the different amino acids or nucleotides that can be observed in the sequences.

The transition probabilities t_{ij} , can be denoted by $P(M_i|M_{i-1})$, which specifies the probability of transitioning from one hidden state to another. For example, the probability of transitioning from a match state to a delete state might be represented as P(Delete|Match). The emission probabilities $P(Oj|Mi)$ specify the probability of observing a particular symbol O_j given a particular hidden state S_i. For example, the probability of observing the nucleotide "G" at a match state might be represented as P("G"|Match). We use the hidden Markov model based on the assumption to identify the common tetrads and diversity in the loop. From the multiple sequence alignment, we calculated the transition and emission probabilities for each group of sequences based on the different possible alignments of the sequences and the observation symbols. Using these probabilities, the probability of a particular alignment given the observed data using the forward algorithm can be estimated or the Viterbi algorithm can be used to find the most likely alignment given the observed data. These algorithms have been used for a variety of applications, including identifying conserved regions in the sequences, predicting the function of a protein and DNA based on its sequence, and identifying relationships between different sequences.

Basically, Hidden Markov model, with HMM, $\lambda = (A, B, \pi)$, observations $O = O_1O_2...O_l$ and state sequence $S = s_1s_2...s_t$ can be used for three types of problems which can be summarized as the evaluation problem, the decoding problem, and the learning problem.

2.7.3.1 (1) The Evaluation Problem

This problem refers to the task of estimating the probability of a particular sequence of observations given a particular HMM. This probability is known as the likelihood of the observations given the HMM and is used to evaluate the fit of the HMM to a given set of data.

Hidden states in model complicate the evaluation process. This is helpful in scoring a sequence. The problem is solved using the forward or the backward algorithm.

For each new query using all multiple identified groups as individual models, we choose the best model among the competing models.

Given a finite collection $(M_1,M_2...,M_L)$ of HMM's with the same output alphabet O, for any output sequence $O = (O_1, O_2,...,O_L)$ of length L, find which model M_ℓ is most likely to have generated O. In this project, we use hidden Markov models and the forward algorithm to identify the DNA sequence family most similar to a given query sequence. This involves building multiple HMMs to represent different DNA sequence families, and then using the forward algorithm to identify the query sequence that is most similar to each of these families. The forward algorithm is a method for computing the probability of an observed sequence given a particular HMM. It works by iteratively updating the probability of being in each hidden state at each time step, based on the probabilities of being in the previous hidden states and the probability of emitting the observed sequence at each time step.

There are several algorithms that can be used to solve the evaluation problem for HMMs, including the dynamic programming based forward algorithm and the backward algorithm. The forward algorithm estimates the probability of a particular sequence of observations given the HMM by iteratively updating the probabilities of the states as it moves forward through the sequence. The backward algorithm estimates the probability of a particular sequence of observations given the HMM by summing over all possible sequences of observations that could have been generated by iteratively updating the probabilities of the states as it moves backward through the sequence. This could be useful for a variety of applications, such as identifying relationships between different DNA sequences or predicting the function of a particular DNA sequence based on its similarity to known sequences. The forward algorithm is discussed further below.

2.7.3.2 Forward Algorithm

To calculate the joint probability of observing the first t characters and being in state s at length l, we can write this as:

$$f_s(l) \; = \; P(\pi_l \; = \; s, x_1, .., x_l \,)$$

With the exponential number of paths for length l, we utilize dynamic programming employing forward algorithm with the Markov property.

$$f_s(l) = \sum P\,(x_1, .., x_l, \pi_l \; = \; s, \pi_{l-1} = t) \; = \; \sum_t P\,(x_1, .., x_{l-1}, \pi_{l-1}, \pi_{l-1} = t) \; *$$

$$P(\, x_l, \pi_l | \, \pi_{l-1})$$

We write the above equation in terms of $f_t(l\text{-}1)$ and transition and emission probabilities,

$$f_k(l) = e_s x_l \sum_t P\,(f_t(l-1) \; * a_{ts}$$

$$P(x_1, .., x_n) = \sum P\,(x_1, .., x_n, \pi_l \; = \; t\,) = \; \sum_t f_t\,(N)$$

Input:

$x = x_1 \ldots x_N$

Initialization:

$f_0(0) = 1$

$f_s(0) = 0$, for all s>0

Iteration:

$f_s(i) = e_s(x_i) \times \sum_j a_{js} f_j\ (i - 1)$

Termination:

$$P(x, \pi^*) = \sum_s f_s(N)$$

2.7.3.3 The Decoding Problem

Given a model and a sequence of observations, the decoding problem determines the most likely state sequence in the model that produced the observations It is formally defined as follows: given an HMM M = (Q, O, π, A, B), for any observed output sequence O = (O_1, O_2,…,O_L) of length L, find a most likely sequence of states S = (s_1, s_2,...,s_m) that produces the output sequence O. This problem is also known as the "maximum a posteriori" (MAP) estimation problem, as it involves finding the hidden state sequence that maximizes the posterior probability of the hidden states given the observations.

The decoding problem is important in a variety of applications, including multiple sequence alignment, gene prediction, and protein structure prediction, where the goal is to identify the most likely alignment or structure given a set of observed sequences.

There are several algorithms that can be used to solve the decoding problem for HMMs, including the Viterbi algorithm and the posterior decoding algorithm. The Viterbi algorithm is a dynamic programming algorithm that finds the most likely hidden state

sequence by recursively computing the maximum likelihood of each hidden state at each position in the sequence, given the observations and the HMM. The posterior decoding algorithm finds the most likely hidden state sequence by computing the posterior probability of each hidden state at each position in the sequence, given the observations and the HMM. Currently, in this project we do not deal with the decoding problem. However, with identification of more structures, syn and anti-conformation of guanines in the model will prove helpful for additional classification.

2.7.3.4 The Learning Problem

In Hidden Markov Models (HMM), the learning problem involves estimating the model parameters from a set of observed sequences. transitions between states and the probability distributions over different emissions probabilities. The goal of the learning problem is to find the optimal values for these parameters that maximize the likelihood of the observed sequences. Given a model and a sequence of observations , how should the model parameters be adjusted in order to maximize Given a set (O_1,O_2..., O_L) of output sequences on the same output alphabet O, usually called a set of training data, given Q, find the optimal values for parameters π, A, and B for an HMM M that produces all the sequences in the training set, in the sense that the HMM M = (Q, O, π, A, B) is the most likely to have produced the sequences in the training set. The technique used here is called expectation maximization, or EM. It is an iterative method that starts with an initial triple π, A, B, and tries to improve it. Baum-Welch Algorithm, frequently known as the forward backward algorithm is used.

In a HMM, the parameters of the model include the probabilities for transitions between states and the probability distributions over different emissions (such as nucleotides or

amino acid). As the number of parameters in the model increases, more information is needed to produce a useful model. To reduce the number of free parameters, the emission probabilities for the insert states can be set equal to or to some background frequency.

CHAPTER 3 STRUCTURAL AND FUNCTIONAL CLASSIFICATION OF G-QUADRUPLEX FAMILIES WITHIN THE HUMAN GENOME

3.1 SUMMARY

G quadruplexes are short secondary DNA structures located throughout genomic DNA and transcribed RNA. Although G4 structures have been shown to form in vivo, no current search tools are known to exist to examine these structures based on previously identified G quadruplexes, much less filter them based on similar sequence, structure, and thermodynamic properties. We present a framework for clustering G quadruplex sequences into families using the CD-HIT, MeShClust and DNACLUST methods along with a combination of Starcode and BLAST. Utilizing this framework to filter and annotate clusters, 95 families of G quadruplex sequences were identified within the human genome. Profiles for each family were created using hidden Markov models to allow for identification of additional family members and generate homology prob-ability scores. The thermodynamic folding energy properties, functional annotation of genes associated with the sequences, scores from different prediction algorithms and transcription factor binding motifs within a family were used to annotate and

Neupane, A.; Chariker, J.H.; Rouchka, E.C. Structural and Functional Classification of G-Quadruplex Families within the Human Genome. Genes 2023, 14, 645. https://doi.org/10.3390/genes14030645

compare the diversity within and across clusters. The resulting set of G quadruplex families can be used to further understand how different regions of the genome are regulated by factors targeting specific structures common to members of a specific cluster.

Keywords: G-quadruplex; G4; clustering; hidden Markov models; DNA structures

3.2 INTRODUCTION

G-quadruplexes are stranded secondary structures of nucleic acids rich in guanine containing four runs of at least three guanines. These runs are separated by short loops, typically 2-7 nucleotides in length, which can potentially fold into an intramolecular G-quadruplex structure. The tetrad guanine structure is stacked on top of each other and held together by mixed loops of DNA forming Hoogsteen base pairing giving a four-stranded structure with nucleobases on the inside and a sugar phosphate backbone on the outside (Figure 3-1). Metal ions (typically K+ or Na+) sitting internally to the Hoogsteen bases stabilize the base pairing. Stacking occurs through the O6 atoms of guanines facing the center creating a tubular space able to function as an ion channel. The presence of a metal cation in this channel allows for interaction with the eight O6 atoms of the guanine quartet.

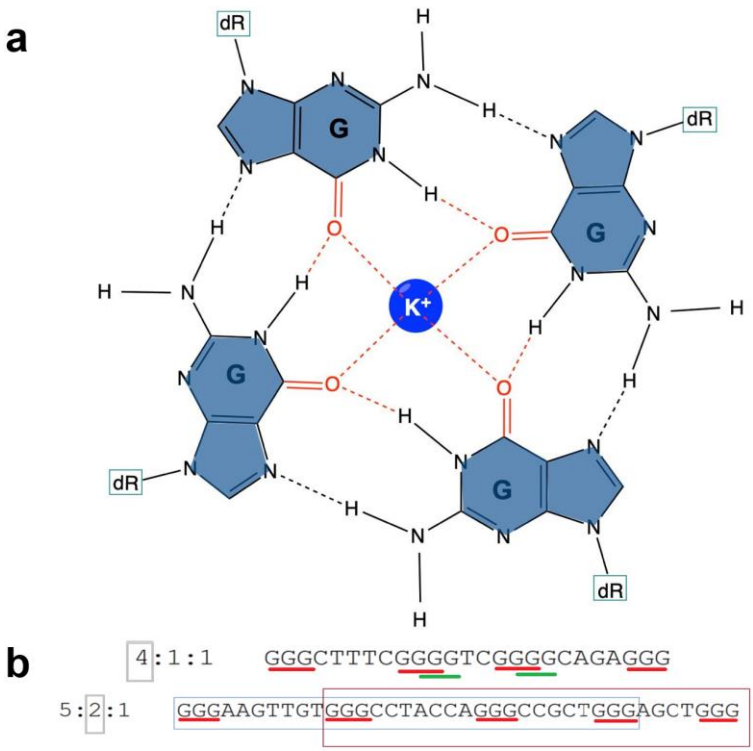


Figure 3-1 (a) G-tetrad structure forming G quadruplexes. Hydrogen bonds between the guanine from different tetrads form a planar ring. (b) Sequence of G4 with multiple guanine tetrads. Here, 4:1:1 and 5:2:1 refers to the result from Quadparser separated out as the number of tetrads: total G4 sequences: non-overlapping G4 sequences.

3.2.1 Roles of G-quadruplexes

Over the past three decades, guanine rich quadruplex sequences have been implicated as key structural regulators of gene expression, cellular differentiation and transcription factors and their cell line and tissue specificity (46). Similarly, elevated levels of G quadruplexes have been identified across cancer tissues including breast (47), stomach (48), liver (49), as well as neurodegenerative diseases (50). Computational analyses of G quadruplex patterns have identified the prevalence of G quadruplexes in oncogenic promoters, introns, splice sites, intergenic and telomeric ends. Initially the secondary

structures were thought to act as a physical obstacle to RNA polymerase for transcription as identified through G4 specific antibodies (51, 52) and chemical probing (53, 54). Further evidence suggests the varied tissue specific functionality of these structures are affected by the cross talk of additional transcription factors (55), proteins and physiological conditions. Additionally, G4 structures have a role in genomic instability and are associated with higher rates of double strand breakage in nucleosome depleted regions of highly expressed cancer genes. High- and low-density bands of G4 across both chromosomal strands have been observed showcasing a role of G quadruplex in pairing of homologous chromosomes during meiosis (56). Further, recent evidence shows that G4 formation is highest during DNA replication at the S phase and lowest during G2 and M phase, consistent with phases of transcription, replication, and chromatin accessibility (57).

3.2.2 Characteristics of G4s

Sequence characteristics such as sequence length (58), base composition (59, 60) and loop length (61-65) are important parameters for defining the secondary structure and stability of G quadruplexes. Molecular dynamics show that telomeric G4 repeats (TTAGGG) in the presence of a K+ cation form a structure with three single nucleotide loops in a parallel fashion. Increasing the loop length by a single base causes the sequences to adopt a mixture of parallel and antiparallel folded structures (66). The conformation and stability of G quadruplexes has been used to study to effect of transcription factor binding and altered mRNA expression of several genes. Examples include nucleolin (67) and Ewing's Sarcoma proteins (68) which preferentially bind to

structures with longer loop length. Computationally, G4s are defined by the pattern $G_xN_{1-7}G_xN_{1-7}G_xN_{1-7}G_x$ where x ≥3 (length of guanine repeats). The guanine tracts are separated by loops of any base composition of length 1-7 bases. This pattern is the basis for regular expression-based tools such as Quadparser (4) and QGRSmapper (7). With experimental data, it is known that different intermolecular structure, long loops, and non-canonical structures with G tracts containing two guanines exist (69-72). Methods such as G4screener (73), PQSfinder (74) and G4Catchall (75) allows the search of G quadruplexes for variable quartet and larger loop sequences. G4Hunter (5) provides a score for guanine skewness which is based on predefined values, with a score based on the number of consecutive Gs. G4RNAscreener (73) uses a machine learning algorithm trained with experimental RNA sequences from the G4RNA database (76) and incorporates a threshold using metrics from tools such as G4Hunter (5), cG/cCscore, and G4 Neural Network score for G4 prediction. RNAfold (77) has an option to predict the thermodynamic parameters for G quadruplex formation. DSSR (78) and ElTetrado (79) use the tertiary structure of each G quadruplex for annotating and classifying different base pairs and tetrad structures. 3D-NuS (80) allows visualization of 3D DNA structures including duplex, triplex, and quadruplexes. 3D NuS visualizes the G quadruplex structure and its strand orientation, loops and G quartets based on the energy minimization of G4 structures using experimental data.

G4 structure was found to be evolutionarily conserved in seven yeast species (81). While G quadruplex regions are significantly enriched in regulatory regions of eukaryotes, short loops of G4 are conserved in different species. Protozoa and fungi have limited diversity

of G4 while an increase in diversity has been observed across invertebrates and vertebrates (82). However, the evolutionary mechanism for this structure or the relationship of these structures at an evolutionary scale is not known.

Sequencing read fragments utilizing a customized approach that introduces stabilizing and destabilizing conditions (K+, Li+, PDS) allows for high throughput sequencing of G4 locations (83, 84) with a method known as G4 seq. Versions of this method have been used to identify 1,420,841 G quadruplexes in 12 species. Using a similar method, 161 and 168 G4 sites were identified in the genomes of Pseudomonas (85) and Escherichia (86), respectively.

Over 100,000 G4 sequences have been mapped in vivo to the human genome. Several proteins such as FUS, TAF15, TARDBP, PCBP1 have been determined to be enriched at G4 loci using artificial G4 binding (87). SP2, a transcription factor (TF) encoded by a subfamily of the Sp/XKLF family, is a sequence specific TF that has a strong association with G quadruplex affinity. SP2 binds to the CCAAT motif independent of the zinc finger domain necessary for binding to GC rich motifs (88). It was shown in vitro that the SP1 TF was able to bind to a DNA sequence lacking the consensus motif and was able to form G quadruplex sequences (89). Luciferase expression studies show sequences of G4s in the KIT promoter mutated through site directed mutagenesis were able to create a modulation (on/off) system for KIT expression through SP1 binding (90). Additionally, G quadruplex structures can bind to G quadruplex sites in other promoter locations (91) mediating cis (92) and trans (93) acting regulation of transcriptional and translational

processes respectively, implying that G quadruplex sequence and structural diversity is a key factor for biological functions.

3.2.3 G4 families

Previously, a small family of G quadruplexes labelled Pu27 was identified based on sequence homology (91). The parent G quadruplex is a 27 nucleotide (nt) G4 formed in the nuclease hypersensitive element (NHE) region of the c-MYC promoter associated with different forms of cancer, and predominantly involved in the regulation of expression of c-MYC gene (94). c-MYC is an oncogene that regulates genes in cell cycle and molecular metabolism. Rezzoug et al identified seventeen potential G quadruplex forming sequences homologous to the Pu27 G4 which has been shown to selectively bind to the NHE region of c-MYC promoter (91). In addition, G4 regions regulating VEGF genes have been shown to have an additional G-tract to act as a spare tire for formation of G quadruplex sequence upon oxidative damage upon the guanine tracts (95). Similar sequences have been identified for c-MYC, KRAS (96), BCL2 (97), HIF-1α and RET genes. This highlights the presence of sequence specific G quadruplexes able to form, bind and regulate gene expression. Further, over the past decade numerous G quadruplex stabilizing and destabilizing ligands have been identified that recognize and interact selectively to these G4 sequences. Different classes of these heteroaromatic polycyclic, macrocyclic and aromatic compounds have been designed to target the diversity of G4 structure. The subtle differences in grooves, loop composition and loop length allow for structural variability in these sequences. DNA aptamers that can form G4 are used for binding nucleolin (98). More than 50 transcription factors with overlapping binding sites

to G4 region have been identified (46, 99). Folding, misfolding and unfolding of G4 structures have been implicated in different biological processes (100, 101).

3.2.4 Detection of G4 families

The prediction of G quadruplexes across genomes can be useful to identify the location of similarly structured G quadruplexes, which can in turn be used to develop profiles of independent families based on conservation of a variety of factors. We present a framework to predict G quadruplex sequences and identify similar sequences using trained profile hidden Markov models (HMMs) (102). We identify pG4 sequences across the human genome, cluster these sequences using sequence clustering tools, CD-hit (103), MeShClust (104) and DNACLUST (105) as well as starcode (106) and BLAST (107). These approaches utilize average weighted clustering to identify the quartet and loop patterns. We then further train HMM models using these clusters for creation of families. Despite the short length of G quadruplex sequences, position dependent insertion and deletion within loops offers insight into the loop characteristics.

3.3 Materials and Methods

3.3.1 Dataset preparation

Since there are no current families or experimental similarities of G4 structure, we start with putative G4s and apply sequence-based methods for clustering. Later, these clusters are used as initial seeds for identifying G quadruplexes in experimental datasets. Initially, we focused on the G4s identified from Quadparser (4) on the human GRCh38 genome. The following process is followed for all groups of sequences based on the number of GGG tetrads (Figure 3-2).

1. CD-hit, MeShClust, and DNAclust and a combination of Starcode and BLAST with hierarchical clustering are utilized for the initial clustering of G-quadruplex sequences.

2. Steps (3)-(7) are repeated separately for each clustering method.

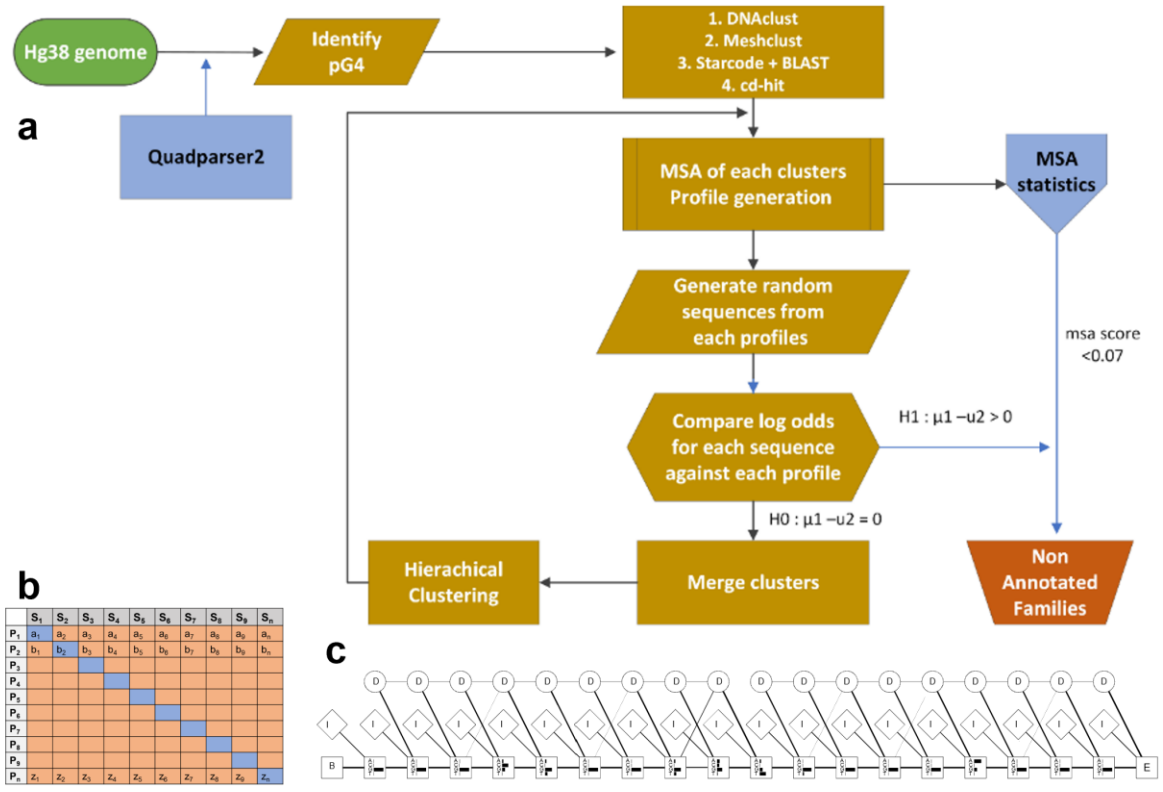


Figure 3-2: Process for identifying and characterizing G quadruplex families. (a) Workflow diagram for identifying distinct G quadruplex families. (b) Process for identifying appropriate profile for a specific family. In this case, $S_1,\ldots,S_n$ represents the list of sequences generated from HMM profile $P_1,\ldots,P_n$ respectively. We compare the average log odds for input S_1 over profile $P_1 \ldots P_n$ and recursively measure for all the profiles. For each row, the diagonal element is compared with non-diagonal values(log-odds) using a Wilcoxon rank sum test with null and alternate hypothesis, H_0: $T_1 - T_2 = 0$, H_1: $T_1 - T_2 > 0$. (c) Profile HMM derived from a selected G4 alignment. Match states are represented as rectangles with four residue emission probabilities indicated as black bars,

insert states (I) as diamond, and delete states as circle. The start and end states are B (begin) and E (end) respectively. Delete states are silent states with no emission probabilities and weighed lines represent the transition probabilities between states.

3. A multiple sequence alignment (MSA) of each cluster of sequences is carried out in R using the DECIPHER package (108). The StaggerAlignment and AdjustAlignment functions are used to separate regions of alignment and gaps are shifted to improve the alignment.

4. Clusters with fewer than four sequences are filtered out. An MSA score for each cluster is calculated as the average number of gaps in each column of an alignment divided by the length using MStatX (109).

5. Each alignment is trained as a model profile HMM using HMMER 3.0 (110) and the aphid package (111) in R version 3.4.1 independently. The transition and emission probability matrices are estimated based on the plan7 PHMM model based on Durbin (102). An example of a profile HMM stating match, insert and delete state is shown in Figure 3-2b). There are seven outgoing transitions based on the match, insert and delete states, i.e. $I_n \rightarrow I_n$, $M_n \rightarrow I_n$, $M_n \rightarrow M_{n+1}$, $M_n \rightarrow D_{n+1}$; $D_n \rightarrow M_{n+1}$, $D_n \rightarrow D_{n+1}$; $I_n \rightarrow M_{n+1}$ where n represents each position of the alignment (except the final position). The observed counts of emissions and state transitions are converted into probabilities.

6. The sequences in each cluster are used as input for all the profiles and the log-odds scores are generated using the forward algorithm.

7. A pairwise Wilcoxon rank sum test is carried out to compare each profile using the log-odds between the profile HMM through which the sequences were generated and all other profiles Figure 3-2c). If a profile is diverse (p value < 0.05) against all profiles compared, has a probability of 0.99 for the tested sequences, and has a gap score less than a threshold of 0.10, the profile is saved as a family. For the sequences that are non-significant (p value > 0.05) the sequences are input to the MSA and are merged and/or clustered using agglomerative clustering. Alignments with a gap score of 0.6 after merging are filtered. The process is iterated for a maximum of n=100 times.

8. The group of sequences obtained from all the methods is combined and checked for redundancy using a modification of step (7) utilizing a threshold score of log-odds 5 and Akaike weight of 0.7 for identifying the final families, which are additionally manually checked and filtered.

9. The alignment and profile HMM are manually verified resulting in 95 clusters referred to as families. Experimentally validated G quadruplexes were obtained from processed peaks mapped to hg19 from GEO, accession GSE63874 (83) using bedtools (112) and quadparser2 after conversion to human genome hg38 coordinates by liftover. The models are used as a trained classifier for identifying additional sequences. G4 sequences from experimental G4 seq was tested against the cluster HMMs. The likelihood that a query sequence fits the model of an individual family is calculated using the forward algorithm (113) and the normalized Akaike weights (114, 115) is calculated. The maximum Akaike weight of the query given a particular model is selected as the

nearest family of the query sequence. The families are manually verified and the variability of sequences in the families is further analyzed based on annotation of the G4, thermodynamic scores (folding energy), G4hunter scores and literature. The steps below highlight the method for the combination of Starcode and BLAST with hierarchical clustering.

a. Levenshtein distance is used to identify the nearest group of sequences which are then filtered based on the length of the sequence and the number of G tetrads. Starcode (106) utilizes a modified Needleman-Wunsch dynamic programming approach known as the poucet algorithm for determining the initial and nearest groups of sequences. Sequences below a fixed Levenshtein score are used to identify the groups and each group is filtered by length of the sequence and loop sequence content. Using specific Levenshtein distance as a constraint through this algorithm, one or two nucleotide mismatches can be identified in short DNA sequences.

b. The remaining sequences from step (1) that are not in any group are passed through BLAST for pairwise all vs all BLAST. -log(E value) is used as the similarity metric.

c. Hierarchical clustering is applied comparing the agglomerative, Ward, complete, and divisive methods of clustering. The number of clusters is calculated based on the sum of the within- cluster inertia. The optimal number of clusters is the maximum difference from two successive clusters between the groups, i.e. max (I_m/I_{m+1}). The mode of the number of clusters was selected as the optimal cluster.

d. Pairwise alignment of sequences of individual clusters obtained from steps (a) and (c) is carried out using the Pairwisealignment function in the Biostrings (116) package. Hierarchical clustering of the sequences is performed based on the pairwise distance. Consensus of Silhouette (117), Frey index, Macclain Index, Cindex, and Dunn index were used for identifying the

Table 3-1 Cluster summary based on different clustering techniques.

Method	Number of sequences	No of clusters	No. of sequences in 2 largest clusters	HMM clusters (sequences)	HMM families, 1st iteration (sequences)	Final families selected (sequences)
Starcode + BLAST with hierarchical clustering	29,112	2,717	419, 323	95 (842)		
DNAclust	9,610 (4,664)	587	142, 126	31 (1,165)		
Cd-hit (kmer 8)	6,335	786	182, 115	30 (389)		
Meshclust	14,222	508	1,720, 1,410	72 (1,843)		
		Total			220 (3,888)	95 (2,174)

optimal number of clusters. The metrics are calculated using the NbClust package (118) in R.

3.4 Results

In the preliminary step, a combination of Starcode and BLAST was used with hierarchical clustering to identify 2,717 clusters of G quadruplexes with 29,112 sequences. Using DNACLUST, 587 clusters with 4,664 sequences were identified. A total of 786 clusters with 6,335 sequences were identified with Cdhit with a k-mer of 8. MeShClust with an identity threshold of 90% and k-mer size of 9 was able to identify 508

clusters. Any clusters with fewer than four sequences were discarded. The two largest clusters had 1,720 and 1,410 sequences, respectively. The overall clustering summary is provided in Table 3-1.

The HMMs for the identified clusters were utilized to predict additional G quadruplex sequences. In addition, the MSA was used to detect transcription factor site motifs found within each family. The G4 families suffered from redundancy of motifs because of the high percentage of guanine bases. To identify unique motifs, a pipeline was created to merge and re-cluster the families. Overall, the Starcode and BLAST pipeline identified 95 clusters of G-quadruplex genomic DNA sequences. The MeShClust pipeline identified 72 clusters, while DNACLUST And CD-HIT identified 31 and 30, respectively. The final iteration of the clustering and merging sequences across profiles from the various clustering approaches resulted in 95 distinct families.

3.4.1 G quadruplex families

The resulting 95 families were created from 1,739 distinct individual G4s identified from 2,145 distinct regions of the hg38 human genome. Given the short sequence length and guanine composition, many of the G4 sequences are not unique. One of the largest families identified, Family 23 is comprised of 163 regions with 118 distinct G4s occurring over 122 genes (Appendix Table A 1). Similarly, Family 79 has 130 regions with 99 distinct G4s occurring over 128 genes distributed across all chromosomes (Appendix Table A 2). We identified multiple sequence repeats capable of forming multiple G4 structures with different conformation in Families 46, 62, 88, 89 and 90 based on the available guanines (Appendix Table A3). Smaller families 2 and 3 have 7

and 6 distinct sequences occurring in proximity to 8 and 7 genes respectively (Appendix Table A 4 & Appendix Table A 5). A summary of the predicted of G4 sequence families is present in Table 3-2.

We analysed the clusters for their sequence characteristics, functional annotation, and structural features as presented below. We highlight some of the clusters that have strong biological significance with related biological and molecular processes, including Family 4 (Appendix Table A 10), Family 32 (Appendix Table A 11), Family 75 (Appendix Table A 12), and Family 80 (Appendix Table A 13).

3.4.2 Categorical enrichment of select families

Family 4 consists of nine sequences distributed over nine genes and seven chromosomes (Appendix Table A 6). Figure 3-3 illustrates the dot bracket notation of the consensus of the family along with thermodynamic characteristics. While this family is relatively small, the associated

Table 3-2 Summary of count of G4 sequences identified using predictive models pHMM across different clusters, genes, and chromosomes.

	TRAINING					PREDICTED			
Family	G4s	Chrs	Distinct Sequences	Associated Genes	Consensus using training sequences	G4s	Chrs	Distinct Sequences	Associated Genes
1	15	12	5	14	GGGGTGGGTGGGGAGGG	643	24	118	468
2	8	5	7	8	--GGGARKGGSCTGGGACAGGG	25	13	25	29
3	10	6	6	7	GGGAGGGGGCTGCWGGGATGGGGG	270	22	257	219
4	9	7	8	9	-GGGCTGGG-GMGGGAAGGAGAGGG	106	22	106	95
5	8	6	7	7	GGGKKGGGGWGAATRGGGCAYGGG-	355	23	341	271
6	8	6	6	7	-GGGGKCTCAGGGGCTGGGCAGRGGG	213	23	200	183
7	7	7	7	7	-GGGC-CCSKGGGCDGSGRGGMRGGG	636	24	614	564
8	7	7	7	7	GGG-MCTTGGGGGTKGGGASAA--GGG-	376	23	369	311
9	10	9	8	10	-GGGSTGGGGAGGGTGGG	350	23	136	276
10	20	15	10	20	GGGGTGGGGGTGGGAGGG	261	23	107	187
11	15	12	8	15	GGGRGKKKGGTGGGAGGG	164	23	84	132

12	17	9	17	17	GGGGC-CWGGG-TGGGA-AAGGG-	347	24	330	289
13	64	20	62	62	---GG-RWGGGCYKGG-GGGCWGGG	143	22	125	125
14	52	20	50	51	-GGGRCGGGGCAGGGG-TG-GGG	163	24	153	140
15	13	9	13	13	GGGRRAWRGGGTGGGAGGG	151	22	116	121
16	8	7	8	8	GGGGATKDG-GGGAGGGAGGG	152	23	134	113
17	23	11	16	23	GGGAAGGG---TCAGGG-CCAGGG	312	22	293	286
18	14	11	12	14	GGGTGGGTGGGGKMAGGG	439	23	242	345
19	8	8	8	8	GGGCCMMGGGCTGGGGCAGGG	59	19	59	63
20	8	6	7	8	GGGWDGGSMRGGGCM--CAAGGG	421	23	414	343
21	7	6	7	7	GGGGC-AGGGGCAGGGDGTGAGGGG	130	23	120	101
22	8	6	8	8	-GGGCYAGGGT-TGGGWRAGGG	60	22	49	44
23	163	23	118	122	--GGGTKG--GKGRWG-GGRTGGGGG	794	24	555	603
24	35	19	34	35	GGGGGYRGGGSWGGGGWGGG	107	21	91	84
25	39	18	32	37	GGGRR-GGG-RTGGGG--CCKGGGG	434	23	418	365
26	9	7	9	9	-GGGGBWGGGGKSAGGGWGGG	69	19	67	49
27	11	9	11	11	-GGG-GCTGGGRMCWGGGCWGGG	113	22	107	98
28	79	21	79	79	GGGGA-WGGGMARGGGY-RGGG	87	21	83	67
29	18	15	17	18	GGGSHWGGGGGGKGGGRGGG	108	21	103	98
30	12	6	12	12	GGGGKRKGGGKMWGGGKGGG	209	23	180	174
31	44	18	43	44	GGGGMRGGGGKKGGGGTGGG	107	23	94	88
32	90	23	85	88	GGGSTGGGGKKGGGGSWGGG	164	22	146	130
33	111	22	102	108	GGGCTG-------GGGCKGGG--SCWGGG	210	22	184	160
34	9	6	8	9	GGGAATGGGGGGTGGGGG-GGGG	101	22	98	70
35	25	16	25	25	-GGGCA---GG-GGAGGGMYAGG-----GG	179	22	173	148
36	52	20	46	48	-GG--GCCTKGGGG---WGGGAGGG-	540	23	497	439
37	7	6	5	7	-GGGSCAGGGCCAGGGCCAGGG	137	22	125	124
38	7	7	7	7	GGGGYGGGGGR-CAGGGCCAGGG	207	23	200	199
39	12	8	11	12	GGGGAGRGTGGG-MAGGGTGGG	145	24	143	111
40	21	13	16	20	GGGYTGGGRA-TGGGTGGG	489	23	289	348
41	11	8	10	11	GGGM-CAGGGYKSSGGSSAGGG	100	22	99	88
42	17	13	17	17	GGGA-GGGAGGGRAACYYSRGG-	534	23	522	415
43	17	11	17	17	GGGGCCYGGGCCTGGGGAGGG	68	22	64	73
44	9	6	9	9	GGGC-YAGA-GGGTGGGYWGGG	151	22	141	125
45	28	12	28	28	-GGGSKK-KGGGCAGGGG--CAGGGG-	207	23	196	151
46	8	7	8	8	-GG-GKTGGGGGMWGGGRGGRGGG	83	21	77	61
47	21	13	17	20	--GGGGTGGGA--GGGATGGYGGGG-	134	21	118	101
48	21	13	21	21	-GG-GRTTGGGGGT-GG-GG-RTGGGG	776	24	724	547
49	29	10	12	12	-GGGGGCAGGGCYGGG-GCTGGG	54	21	44	43
50	32	19	30	32	-GGGAGAGGGT--TKGGKGR--AGGG	271	23	252	221
51	12	7	9	10	-GGGGTGGGCAGGGMAGMYTGGG	141	24	136	118
52	9	8	9	9	GGGCCCCSGGGGCGGGGCGGG	265	24	264	309
53	56	19	54	50	--GGGDGT-G-G-GSGG-AGGGAGGG--	155	22	145	127
54	33	18	31	33	GGG-CTCR-GG-RMAGGG-CTGGG	214	24	206	196
55	21	16	21	21	-GGGYR-GGGGTGG-GGGGC---RGGG	111	23	110	112
56	9	7	9	9	-GGGGTGGGGTKGGGG-GKRGAGGG	332	24	319	258
57	14	11	9	14	GGGSC-GGGGCGGGCGGGG	314	23	164	328
58	27	9	15	14	-GGGCTGGGKGRGGGGA-GCAGGG	155	23	132	110

59	44	16	44	44	GGG-SAGGGC-KGGGADRGGGG	265	23	247	226
60	8	7	8	8	-GGGGGTGGGGG--RRWGGGSAGGG	124	21	115	98
61	10	1	9	8	GGGACTYRTGGGCTTTGGGCCAAGGG--	106	21	105	106
62	10	4	8	6	GGGGAGACTGGGGAGGCCGGGGYRGAAGGGG	73	20	64	45
63	97	24	1	97	GGGAGGGAGGGAGGG	313	23	1	204
64	31	9	16	12	-GGGGTGKG-GGGGGGRMSGGGG	54	17	42	29
65	16	11	9	15	GGG-GARTGGGCYGGGATGGG-	97	21	86	72
66	58	21	49	53	-GG-----STGGG--CCYTG--GGK-TG--GGG	268	23	260	236
67	6	4	6	6	GGGGTGGG-CATGGGAG-GCAGGG-	214	23	200	171
68	13	1	12	12	-GGGGAGG-GGGGTGCCCTGGGTTGGG-	138	20	118	119
69	11	7	8	11	GGGCAW-GAGGG-A-G-GGKTGGG	129	22	119	99
70	19	11	14	16	GGGRGKTGGGTGGGGGTGGG	202	23	155	161
71	6	5	5	5	GGGGAAGGGACAGGGGMMRGGG	162	23	157	157
72	10	8	8	10	GGGSWG-CAGGG---AGGGCTGGG-	206	22	188	158
73	12	10	11	12	GGGTG-GGGTGGGGK-KRGATGGG-	947	23	917	664
74	12	8	12	11	GGGTGGGGRCAAGGGTRGGG	142	22	129	119
75	18	10	13	16	-GG-GGTGGGA-GGGCMKGGG	343	23	180	265
76	6	4	6	6	GGGGTGGGTGGGG-RATGAGGGG	451	24	420	329
77	19	13	19	19	-GGRRWGGGGRA--ARGAGGGAGGG	296	23	290	223
78	10	9	10	10	GGGGAMT-TGGGGGKGGGG-GGG	329	24	321	268
79	130	22	99	128	-GGGMGGGG-CGGGGCG--GGG	712	24	400	677
80	21	12	21	21	GGG-GCGGGSC---SSGGGGGMGGG-	406	23	389	418
81	13	10	13	13	GGGGRAGGG-T-GGGCTTTGGGG	347	23	329	270
82	38	20	13	34	GGGCAGGGCAGGG-CAGGG	391	24	211	284
83	10	5	8	10	GGGT-CTGGGT--CTGGGTCWGGG-	116	23	111	102
84	6	4	5	5	-GGGGCCGGGGTGGGARGYGGG	66	21	64	62
85	12	8	12	10	-GGGKY-AGGGCCAGGGTGGGGG--	53	21	50	42
86	8	3	4	5	GGGAGGGTCCWGGGGYTGGG	129	22	116	103
87	9	6	9	7	GGGSBCWGGGWS-AGGGAGGG	73	20	69	67
88	11	7	11	11	-GGGRGRCYTGGGTGGGGGGG-	120	22	107	103
89	11	9	6	11	-GGGGTGGGGGTGGGGGGG	43	20	12	40
90	10	8	3	9	-GGGGTGGGGTGGGGGGG	112	23	13	81
91	9	9	9	9	-GG-GGWGGGAGGGAARACKGGG-	75	21	70	70
92	13	7	11	13	GGGKT-GGGGAGGGGAWTWRGGG	451	23	428	367
93	9	8	7	9	GGGCCTGGGCYTGGGCYDGGG-	26	16	25	25
94	12	10	12	12	GGGAMAGGGGGSAGGGGCRGGG	86	20	86	80
95	8	7	8	8	----GGGGACAGGGRCA-GGGVCAGGG	120	21	88	79

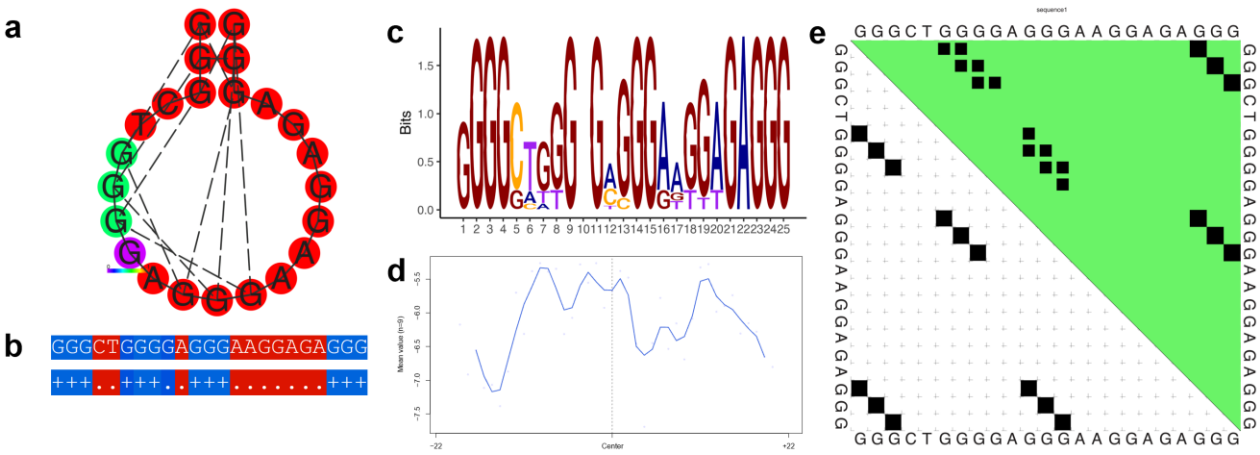


Figure 3-3 Thermodynamic properties for Family 4. (a) Centroid secondary structure with a minimum free energy of -9.64 kcal/mol using the consensus sequence of the family. (b) Dot-bracket notation showing the secondary structure. (c) Sequence logo representing the per base information content. (d) Electrostatic potential generated from all the sequences of the family using 10 flanking bases on either side of the identified G4. (e) Dot plot showing the substructures with the highest probabilities.

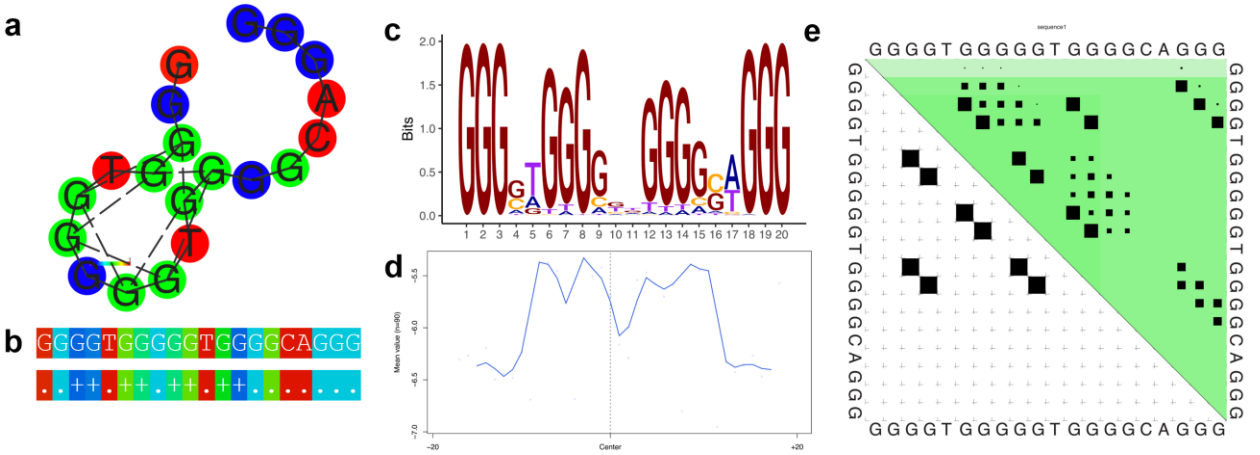


Figure 3-4Thermodynamic properties for Family 32. (a) Centroid secondary structure with a minimum free energy of -18.0 kcal/mol using the consensus sequence of the family. (b) Dot-bracket notation showing the secondary structure. (c) Sequence logo representing the per base information content. (d) Electrostatic potential generated from all the sequences of the family using 10 flanking bases on either side of the identified G4. (e) Dot plot showing the substructures with the highest probabilities.

genes are related, showing an enrichment of terms related to neural cells (e.g. glia guided migration, synapse assembly, dendritic spine development, and gliogenesis) (, Appendix Table A 10).

Family 32 contains 90 G4 sequences annotated with 85 genes (Appendix Table A 7) The thermodynamic properties are illustrated in Figure 3-4. The genes associated with Family 32 G4s are enriched for cellular organization (e.g. positive regulation of cell projection organization and positive regulation of cellular component organization), axonal development (e.g. neuron projection guidance, axon guidance), mitochondrial localization (e.g. regulation of protein targeting to mitochondrion and regulation of establishment of protein localization to mitochondrion) and size regulation (e.g., regulation of anatomical structure size and regulation of cell size) (Appendix Figure A 7, Appendix Table A 11).

Family 75 is represented by 18 G4 sequences distributed over 10 chromosomes and 16 genes (Appendix Table A 8). Enriched GO: BP terms are highly related to immune differentiation and adhesion (e.g. positive regulation of T cell differentiation, positive regulation of lymphocyte differentiation, positive regulation of leukocyte cell-cell adhesion) (Appendix Figure A 8, Appendix Table A 12).

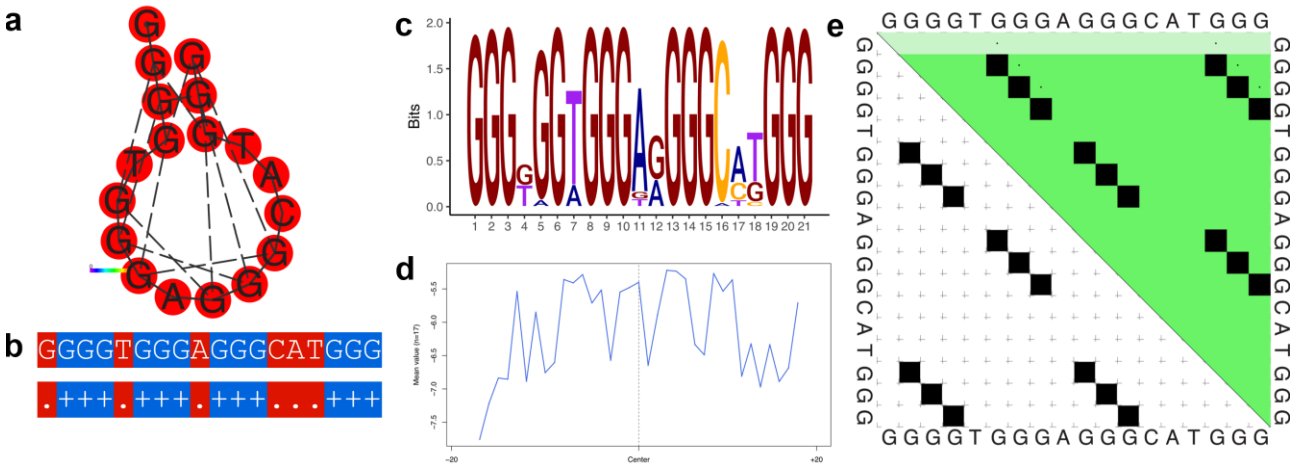


Figure 3-5 Thermodynamic properties for Family 75. (a) Centroid secondary structure with a minimum free energy of -22.82 kcal/mol using the consensus sequence of the family. (b) Dot-bracket notation showing the secondary structure. (c) Sequence logo representing the per base information content. (d) Electrostatic potential generated from all the sequences of the family using 10 flanking bases on either side of the identified G4. (e) Dot plot showing the substructures with the highest probabilities.

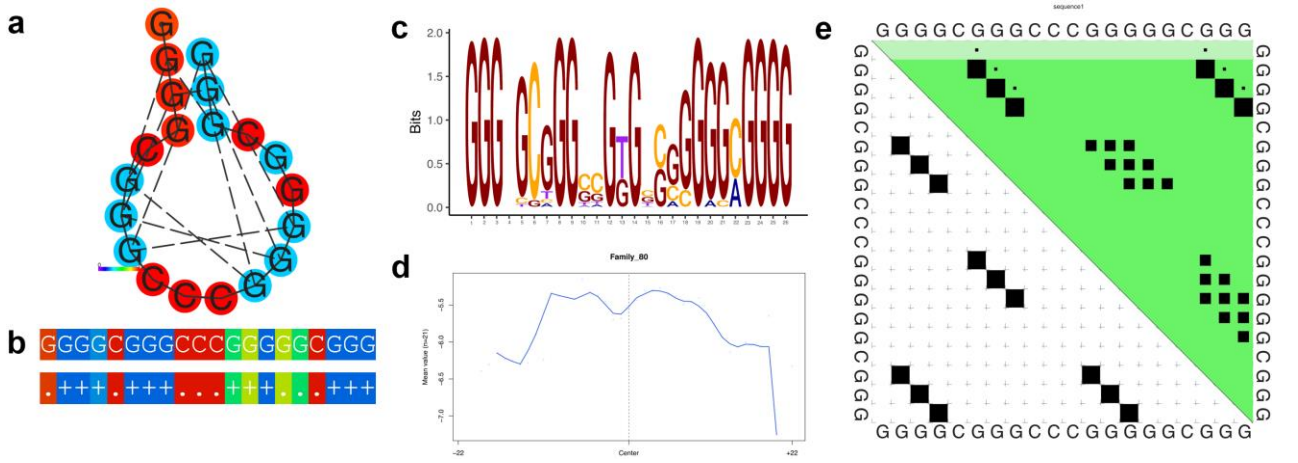


Figure 3-6 Thermodynamic properties for Family 80. (a) Centroid secondary structure with a minimum free energy of -17.38 kcal/mol using the consensus sequence of the family. (b) Dot-bracket notation showing the secondary structure. (c) Sequence logo representing the per base information content. (d) Electrostatic potential generated from all the sequences of the family using 10 flanking bases on either side of the identified G4. (e) Dot plot showing the substructures with the highest probabilities.

For Family 80, we identified 21 sequences distributed over 12 chromosomes and 21 genes (Appendix Table A 9). Genes associated with this family appear to be localized to cellular components membranes. Enriched GO:CC categories for the genes include cytoplasmic side of membrane, plasma membrane, cytoplasmic side of plasma membrane, plasma membrane region, cell projection membrane, ficolin-1-rich granule membrane, side of membrane, cell periphery, ruffle membrane, secretory granule membrane, leading edge membrane, actin filament, extrinsic component of cytoplasmic side of plasma membrane, ruffle, membrane, extrinsic component of plasma membrane, intrinsic component of membrane, intrinsic component of endoplasmic reticulum membrane, plasma membrane protein complex, ficolin-1-rich granule, and tertiary granule (Appendix Table A 13, Appendix Table A 13). A summary of enriched GO terms as determined from GOprofiler and simplifyEnrichment for selected families is present in

Figure

Figure 3-7.

3.4.3 Thermodynamic properties of select families

The free energy of the thermodynamic ensemble for the consensus sequence of Family 1 was calculated to be -28.11 kcal/mol. The frequency of the MFE structure was 50.62% with an ensemble diversity of 0, suggesting a strict conformation of tetrads for formation of a G4 structure. The minimum free energy for the family was calculated to be -27.69 kcal/mol. This family consists of six training sequences that have a single length loop with T-T-A loops (represented by 1-1-1 loops). For Family 11, the free energy of the

thermodynamic ensemble was calculated to be -20.22 kcal/mol. The frequency of the MFE structure in the ensemble is 25.23% and the ensemble diversity is 0, suggesting once again a strict conformation of tetrads for G4 formation. Family 63 is identified with the sequence G3AG3AG3AG3 and is found across 24 chromosomes and 97 genes distributed among intronic, intergenic and promoter regions. The free energy of the thermodynamic ensemble for Family 63 was calculated to be -36.00 kcal/mol while the frequency of the MFE structure in the ensemble is 100% and the ensemble diversity is 0.00. Figures 3-3a, 3-4a, 3-5a and 3-6a represent the base pairing of each base in the G quadruplex sequence. Figures 3-3b, 3-4b, 3-5b and 3-6b highlight the centroid secondary structure in dot-bracket notation. A base pairing probability matrix is used to identify added information about the ensemble G4 secondary structure.

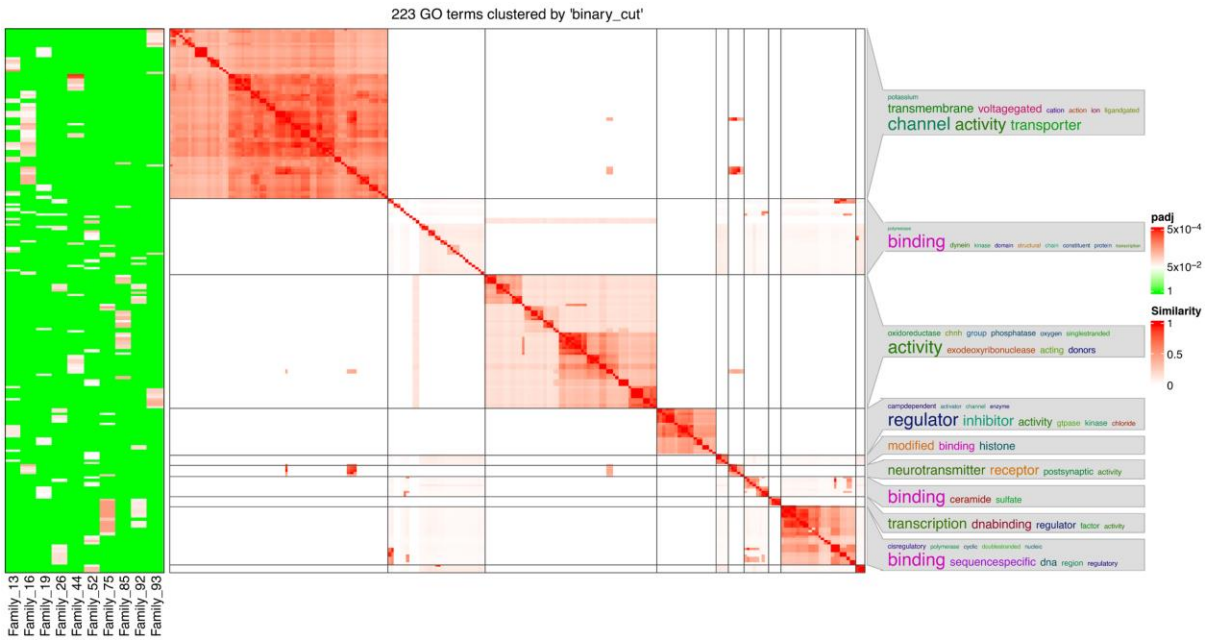


Figure 3-7 Summary of enriched GO terms for select families as determined by the GOProfiler and simplifyEnrichment R packages.

Applied initially to identify different secondary structures of RNA sequences, dynamic programming provides efficient computation of base pairing probabilities for secondary structure formation. The MFE secondary structure highlighting encoding positional entropy (Figures 3-3c, 3-4c, 3-5c and 3-6c) is calculated using the consensus sequence of the G4 cluster as predicted by RNAfold. DNA shape features such as the minor groove width and electrostatic potential (Figures 3-3d, 3-4d, 3-5d and 3-6d) depend upon the charge distribution of nucleotides in a DNA sequence and affect the folding into secondary structure and transcription factor binding in these locations (119). The difference in stacking energies causing the varying hydrogen bonding patterns can be predicted in each dinucleotide step and can be used to infer minor groove width (120). The guanine amino group repeats in G quadruplexes affect charge distributions in the minor and major groove of helical DNA leading to rotation of the tetrads. We use it to

annotate the different families of G quadruplex identified here. A dot plot of the structure with MFE is shown in Figures 3-3e, 3-4e, 3-5e, and 3-6e for each of the selected families.

When DNA is bent around in secondary structures such as helical or G quadruplex structures, the bend is separated based on dinucleotide sequences. Propeller twist is defined as the twist along the axis making two bases “non-coplanar” (121). Previous studies have provided evidence for the flexibility nature of the GG and GC dinucleotides with low propeller twist while AA shows the highest. The flexible nature of such a structure favors G quadruplex sequences. Low propeller twist is related with the ability for the nucleotides to slide on each other and stack in a stable manner. For each cluster, we calculated dinucleotide frequency normalized by individual length of G quadruplex, minimum free energy, minor groove width, propeller twist, helical twist, roll, and electrostatic potential with -10 and +10 region around the identified clusters of G quadruplex using DNAshapeR (122). These features address the shape, thermodynamic stability, and flexibility of rotation of the guanine amino groups, and transcription factor recognition site.

3.4 Classification of experimentally validated G4 sequences

Using the sequences from peaks mapped from a G4 seq experiment (GEO accession GSE63874), and identified using Quadparser2, we found all possible pG4 sequences with four tetrads and used it as query the model classifier. We classified 18,340 individual G4s identified from 22,226 distinct regions of the hg38 human genome into 95 families. Based on the clustering for experimental sequences, the major families represented are

Family 73 (917 unique G4s related to 664 genes), Family 2 (25 unique G4s, 29 genes), and Family 93 (26 unique G4s, 25 genes). Family 63 has a distinct G4 sequence $G_3AG_3AG_3$ that is repeated throughout the genome, occurring 313 times over 23 chromosomes and 204 genes.

3.5 G4 repeat and loop length characteristics

For genes with repeats of G4 sequences (i.e. more than four tetrads), multiple G4 sequences with a variable loop length are possible

Figure 3-8 Example sequences with multiple tetrads. (a) G-quadruplex sequence from chr19:43,479,561-43,479,598 overlapping the PHLDB3 gene with guanines labelled in red. (b) Three possible alternate G4 regions for the PHLDB3 region. (c) MFE structure for the PHLDB3 region. (d) G-quadruplex sequence from chr17:81,432,609-81,432,932 overlapping the BAHCC1 gene with guanines labelled in red. (e) MFE structure for the BAHCC1 gene.). We identify all possible linear combinations of G tetrads for such sequences and classify all combinations of the sequences into families. This provides a way to identify multiple conformations forming G quadruplexes. One example gene with a variable length sequence is BAHCC1, a chromatin regulator known to interact with transcriptional repressors to ensure gene silencing through recognition and bind to PRC2 complex mediated H3K27me3 through chromatin compaction and histone deacetylation (123, 124). Within a single G4 region, we identified repeats of 13 different sequences (length of G4 repeat: 314 bases), with each sequence being distinct enough to occur in a separate family. We also find 29 G4

sequences in NRD2, with most of the sequences occurring in Family 17, with one each also occurring in Family 7 and 10.

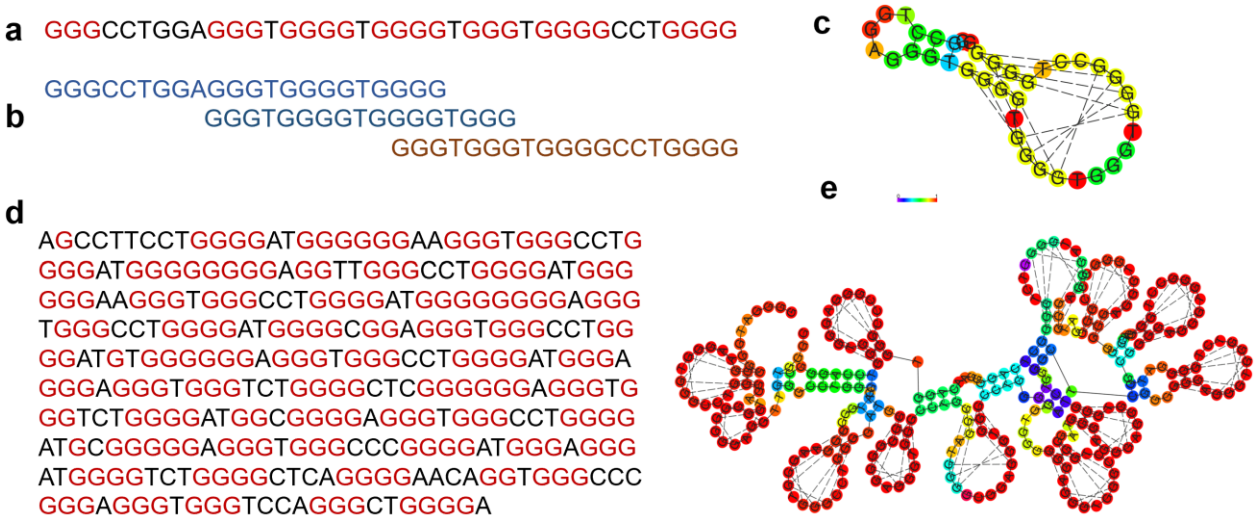


Figure 3-8 Example sequences with multiple tetrads. (a) G-quadruplex sequence from chr19:43,479,561-43,479,598 overlapping the PHLDB3 gene with guanines labelled in red. (b) Three possible alternate G4 regions for the PHLDB3 region. (c) MFE structure for the PHLDB3 region. (d) G-quadruplex sequence from chr17:81,432,609-81,432,932 overlapping the BAHCC1 gene with guanines labelled in red. (e) MFE structure for the BAHCC1 gene.

We identify similar repeats of five distinct sequences spanning an intronic region in PLOD1, which codes for lysyl hydroxylase and is involved in collagen synthesis. A 45 nucleotide G quadruplex sequence present in the promoter region of tyrosine hydroxylase (TH) can regulate transcription and has been linked with neurological and psychological disorders such as Parkinson's and schizophrenia (125, 126). We found two additional G quadruplex sequences in the opposite strand across promoter and intronic regions of TH which have matches to Family 14 and 37, respectively.

Semaphorins are a group of membrane spanning proteins that bind to Plexin (PLXNA and PLXNB) receptors to regulate axon cue signaling, cytoskeletal development and cell adhesion (127, 128). The regulation and signaling of SEMA proteins with the plexin family has been a topic of study, and we identified 39 and 37 distinct G quadruplex forming sequences in the SEMA family and PLXN family respectively, with similar G4 loops present in both genes. The prediction identified multiple G4 sequences present in SEMA6C, SEMA6D, and PLXND1 with the highest match to Family 48 (Table 3-3). Similarly, SEMA4D, SEMA4B, and PLXNA4 shared sequences occurring in Family 17. These findings suggest that multiple regions can form G quadruplex in these genes, resulting in multiple conformations that might allow for differentiation for methylation in a pattern specific manner.

Table 3-3 G4 sequences identified in the genic regions associated with the plexin and semaphorin gene families with high similarity to G4 Families 17, 48 and 79.

Location	Sequence	Log odds	Akaike weight	Strand	Gene ID	Gene symbol	Family
chr15:90204178-90204199	GGGAGGGCACTAGGGCCCTGGG	8.987	0.617	+	10509	SEMA4B	17
chr3:126991053-126991092	GGGCAGGGCAGGCAGGGAAGGG	10.584	0.892	+	5361	PLXNA1	17
chr9:89440465-89440503	GGGTAGGGCTCAGGGGCCAGGG	14.015	0.996	-	10507	SEMA4D	17
chr1:151141755-151141776	GGGATGGGGGTTGGGGGGTGGG	13.6	0.828	-	10500	SEMA6C	48
chr15:47662210-47662233	GGGGTGGGGGGTGAGGGGATGGGG	11.857	0.994	+	80031	SEMA6D	48
chr3:129567938-129567973	GGGTTGGGGTGGGGGGTGGGG	12.652	0.772	-	23129	PLXND1	48
chr3:129588350-129588372	GGGTGTCGGGGGTGGGGGAGGGG	9.599	0.787	-	23129	PLXND1	48
chr3:122983446-122983465	GGGCGGGGACGGGGCGGGG	12.301	0.981	-	54437	SEMA5B	79
chr3:129606851-129606910	GGGCGGGGCCGGGGCGGGG	14.216	0.916	-	23129	PLXND1	79
chr3:50276050-50276067	GGGAGGGTCGAGGGCGGG	6.415	0.677	+	7869	SEMA3B	79

The PDB structures 22AG, 2KF8, 5LQG, and 5YEY represent telomeric quadruplex DNA forming a range of conformations with antiparallel topology based on varying physiological conditions. These telomeric G4 sequences are determined to have the highest likelihood of matching Family 22. They have a similar loop size to structure 2KM3 (70), which has a variant of CTAGGG repeat instead of TTAGGG repeats. The 2KM3 structure forms a chair type G quadruplex in K+ solution and is most similar to Family 33. Based on the sequence characteristics, these differences in structure which are caused by a one or two bp change can affect the overall prediction of the glycosidic conformation. This in turn can be used to help understand the structure based on the local environmental and interacting conditions.

The 2LXQ G4 structure is found upstream of pilin expression locus in Neisseria gonorrhoeae, a human pathogen 5'-$G_3TG_3TTG_3TG_3$ sequence is implicated in pilin antigenic variation (129). Known to form an all-parallel stranded topology, the sequence was predicted to have the highest likelihood score with Family 40. A highly conserved G4 sequence at NHE III_1 upstream of promoter 1 been studied and identified to silence transcription of c-MYC (94, 130-133) and other short loop G4 sequences that form a similar topology. $TAG_3AG_3TAG_3AG_3T$ was predicted to belong to Family 52 as well as Family 1. Despite following the same 1:2:1 pattern as the 2LXQ structure, the presence of adenosine in place of thymidine as the linker loops is considered as a different family.

Experimental evidence shows that G4s with short loop sequences favor a parallel topology while structures with longer loops tend to form hybrid or antiparallel structures (134). Sequences with thymine compared to adenine as a single length loop have been found to have higher melting point than a single A base (135). Given our clustering scheme, multiple sequences with short loops can show high log-odds for multiple families. In these cases, the Akaike weight can help guide the context and identify multiple families containing such sequences.

3.4.4 G4 in enhancers

Potential regulatory roles of G4 families were analyzed by looking at the overlap between G4s and enhancers. The overlapping enhancers were then used as input into the Gene-Enhancer link correlation (http://compbio.mit.edu/epimap/) to determine if any of the overlapping enhancers were correlated with gene expression, and if so, in what cell type. We then performed hierarchical clustering of the intersecting G4s based on the correlations. Two main groups of interest result.

In the first group, 102 G4 sequences are found in 158 genes, belonging to 57 distinct G4 families. GO:BP analysis of this group results in terms associated with immune system processes (e.g. T cell receptor signaling pathway, regulation of leukocyte proliferation, interleukin-10 production and regulation of cytokine production involved in immune response) or signaling cascades (e.g. positive regulation of ERK1 and ERK2 cascade, calcium ion transmembrane import into cytosol, and Fc receptor signaling pathway) (Appendix Figure A 10, Appendix Table A 14).

The second group had ubiquitous high correlation with all cell types in the dataset (Appendix Figure A 12). We identified 234 genes in this group with 107 distinct G4s belonging to 55 distinct families and found enrichment of terms relating to immune responses (e.g. defense response to virus, cytokine-mediated signaling pathway and regulation of defense response), regulated cell death (e.g. apoptotic signaling pathway, extrinsic apoptotic signaling pathway via death domain receptors, and positive regulation of programmed cell death), lipid biosynthesis (e.g. regulation of lipid biosynthetic process and response to fatty acid), and migration (e.g. positive regulation of protein localization and positive regulation of mononuclear cell migration) (Appendix Figure A 11, Appendix Table A 15).

Based on the enriched terms from two groups, it appears as though the G quadruplex functions across multiple pathways in different cell types. It is possible that tissue specific conditions control the actual G4 formation, leading to tissue specific functional regulation. The results of the enhancer-gene correlation related to the presence of G4 sequences in enhancer regions in group 1 are more likely to affect genes in thymus, T cell and lymphoblastoid cells.

3.5 Discussion

Our clustering methodology presented here has allowed for the construction of families of G quadruplexes based on sequence similarity, loop length and composition, and thermodynamic properties. Further analysis of these families uncovers that many of these families have functional enrichments, indicating they are potentially regulated by common mechanisms since they have structural similarities. Comparing our results to

the only previously studied family, Pu27, shows a high agreement, with 12 of the 18 Pu27 members belonging to Family 1 (Table 3-4).

Table 3-4. Family prediction for previously identified Pu27 family of G4 sequences.

Overall sequence	name	Minimum G4 sequence	length	Log odds	Akaike weight	family
TGGGGAGGGTGGGGAGGGTGGGGAAGG	Pu27-c-MYC	GGGGAGGGTGGGGAGGG	17	6.99	0.89	1
		GGGTGGGGAGGGTGGGG	17	5.7	0.59	40
		GGGGAGGGTGGGGAAGG	17	4.95	0.45	1
TGGGAGGTGGGGAGGAGGGTTGGGAAGG	Pu1--PLEKHG5	GGGAGGTGGGGAGGAGGGTTGGG	23	7.42	0.53	48
TGGGAGGTGGGGAGGAGGGTTGGGAAGG		GGGAGGAGGGTTGGGAAGG	19	6.93	0.94	15
TGGGGAGGGTGGGGAGGCCGGG	Pu1-2-MYBPHL	GGGGAGGGTGGGGAGG	16	2.41	0.53	1
TGGGGAGGGTGGGGAGGGTGGG	Pu3---	GGGGAGGGTGGGGAGGG	17	6.99	0.89	1
		GGGTGGGGAGGGTGGG	16	7.33	0.9	9
TGGGGAGGGTGGGGAGGGCGGGG	Pu3-SOX2	GGGGAGGGTGGGGAGGG	17	6.99	0.89	1
		GGGAGGGTGGGGAGGG	16	5.62	0.74	1
TGGGGAGGGTGGGGAGGGTGGTGAGGGTGGGGAGGGGGAAGG	Pu5-GRM6	GGGGAGGGTGGGGAGGG	17	6.99	0.89	1
		GGGAGGGTGGGGAGGG	16	5.62	0.74	1
		GGGGAGGGTGGTGAGGGTGGGG	22	7.53	0.26	76
		GGGTGGTGAGGGTGGGGAGGGGG	23	7.47	0.82	73
TGGGGAGGGTGGGGAGGGTGGGGAGGG	Pu7-SDK1	GGGGAGGGTGGGGAGGG	17	6.99	0.89	1
		GGGTGGGGAGGGTGGGG	17	5.7	0.59	40
GGGTGGGGAGGGTGGGGAAG	Pu9---	GGGTGGGGAGGGTGGGG	17	5.7	0.59	40
GGGGAGGGTGGGGAGGGGATGGAA	Pu9-2BC022036	GGGTGGGGAGGGGATGG	17	5.85	0.37	40
GGGAGGGTGGGGAGGGTGGGGAGGG	Pu10-1--	GGGAGGGTGGGGAGGG	16	5.62	0.74	1
		GGGTGGGGAGGGTGGGG	17	5.7	0.59	40
		GGGGAGGGTGGGGAGGG	17	6.99	0.89	1
GGGTGGGGAGGGTGGGGAAGG	Pu10-2--	GGGTGGGGAGGGTGGGG	17	5.7	0.59	40
		GGGGAGGGTGGGGAAGG	17	4.95	0.45	1
GGGGAGGAAGGGGAGGGTGGGGAGGG	Pu11NAV2	GGGGAGGGTGGGGAGGG	17	6.99	0.89	1
		GGGAGGGTGGGGAGGG	16	5.62	0.74	1
GAGGGTGGGGAGGGTGGATGAGGAAGG	Pu14SPTLC2	GGGTGGGGAGGGTGG	15	3.19	0.63	9
TGGGGAGGGTGGGGAGGGTGG	Pu16--	GGGGAGGGTGGGGAGGG	17	6.99	0.89	1
		GGGAGGGTGGGGAGGG	16	5.62	0.74	1
GAGGGTGGGGAGGGTGGGGA	Pu17--	GGGTGGGGAGGGTGGGG	17	5.7	0.59	40
GGGGAGGGTGGGGAGGGAGCTGGGGA	Pu20-CDH4	GGGGAGGGTGGGGAGGG	17	6.99	0.89	1

		GGGTGGGGAGGGAGCTGGGG	20	4.01	0.49	51
TGGGGAGGGTGGGGAGAGGCGGGGTGGGGAGGG	PuX-TM4SF2	GGGAGGGTGGGGAGAGG	17	3.41	0.83	18

Multiple transcription factors can bind to the alternative motifs present in G quadruplex regions (99) in response to environmental conditions in response to stimuli. These conditions trigger the folding and unfolding of G4 structures. We identify Family 40 as an alternate conformation in these sequences as multiple tetrads allow the alternate guanine bonds for stable structure. Nucleoside diphosphate kinase (NM23-H2) (136, 137) has been previously identified to unfold Pu27 causing the increase of c-MYC transcription while nucleolin (138) has been identified to stabilize the G4 structure. The mechanism of TF binding and control of expression of the expression of c-MYC gene is poorly understood and is beyond the scope of prediction through this model. However, this process sheds light upon the collection of multiple conformation of structures in equilibrium which can alter the change in binding grooves for transcription factors and further downstream process. Failing to take the dynamic nature of Pu27 and other G quadruplex sequences in the genome into account could limit the effectiveness of any therapeutic compounds designed to target it.

Several G4 ligands are currently being considered for their therapeutic value. For instance, CX-5461 is utilized for treatment of BRCA1/2 deficient tumors through topoisomerase II inhibition (139, 140), Melanoma cell lines have been treated with G4 ligand RHPS4 that targets the MYC gene (141) among others. G4 ligands such as APTO-253 (142), TMPyP4 (143), telomestatin (144) have been tested for their effect on

leukemia. Despite showing promising results and inhibition of cell growth, telomerase shortening and senescence was observed with some of the G4 ligands in different leukemia cells (145). With the information of G4 formation and binding of specific ligands to multiple G4 structures, identification of G4 clusters can provide additional information about DNA damage occurring or novel binding motifs of specific G4 ligands.

G4 structures contribute to genomic instability and the proliferative nature of different cancers. The context and location of individual G4 can serve as a roadblock for many oncogenes, but the presence of G4 in the vicinity of a tumor suppressor gene can have the opposite effect. To understand the intended consequence of these targets for all the G4 ligands, it is important to characterize the thousands of G4 structure present in the genome and classify these structures based on their structure, function, or localization.

This study identifies related families of G quadruplex sequences within the human genome and presents them as clusters described by both an MSA and HMM. The approach described here can easily be applied to other model organisms where G4s are known to play regulatory roles. Many of these clusters were functionally annotated, allowing for a more complete understanding of these structures as well as identification of multiple targets for testing of G4 ligands. As more information on experimentally validated G4 regions becomes available, refinement of clustering methodologies will yield more informative G4 families.

.

CHAPTER 4 ANALYSIS OF NUCLEOTIDE VARIATIONS IN HUMAN G-QUADRUPLEX FORMING REGIONS ASSOCIATED WITH DISEASE STATES

4.1 SUMMARY

While the role of G4 G quadruplex structures has been identified in cancers and metabolic disorders, single nucleotide variations (SNVs) and their effect on G4s in disease contexts have not been extensively studied. The COSMIC and CLINVAR databases were used to detect SNVs present in G4s to identify sequence level changes and their effect on alteration of G4 secondary structure. 37,515 G4 SNVs in the COSMIC database and 2,115 in CLINVAR were identified. Of those, 7,236 COSMIC (19.3%) and 416 (18%) of the CLINVAR variants result in G4 loss, while 2,728 (COSMIC) and 112 (CLINVAR) SNVs gain a G4 structure. The gene ontology term "GnRH (Gonadotropin-releasing hormone) secretion" is enriched in 21 genes in this pathway that have a G4 destabilizing SNV. Analysis of mutational patterns in the G4 structure show a higher selective pressure (3-fold) in the coding region on the template strand compared to the non-template strand. At the same time, an equal proportion of SNVs were observed among intronic, promoter and enhancer regions across strands. Using GO and pathway enrichment, genes with SNVs for G4 forming propensity in the coding region are enriched for Regulation of Ras protein signal transduction and Src homology 3 (SH3) domain binding.

4.2 INTRODUCTION

G-quadruplexes are stranded secondary structures of nucleic acids rich in guanine. These nucleic acid sequences are characterized by four runs of at least three guanines separated by short loops, which can potentially fold into an intramolecular or intermolecular G-quadruplex structure (146). The tetrad structure of guanine is stacked on top of each other and held together by mixed loops of DNA giving a four-stranded structure that has nucleobases on the inside forming Hoogsteen base pairing and the sugar phosphate backbone on the outside (Figure 1). They are found in G-rich sequences of both DNA and RNA and are stabilized by metal cations such as potassium (K+) or sodium (Na+) (147). The binding energy is held through the H bonding between the guanines, stabilized by π-π interactions and charge interactions between the sixth position of oxygen (O6) and cations (K+, Na+) between the stacks. The structural architecture of a G-quadruplex is quite diverse and can form different topologies based on factors such as the chemical environment, loop length (134, 148), and localization in the sequence or structure molecularity (149). The stacking of the guanine tetrads is bound by the loops of nucleotide bases of variable sizes which determine the folding of the secondary structure.

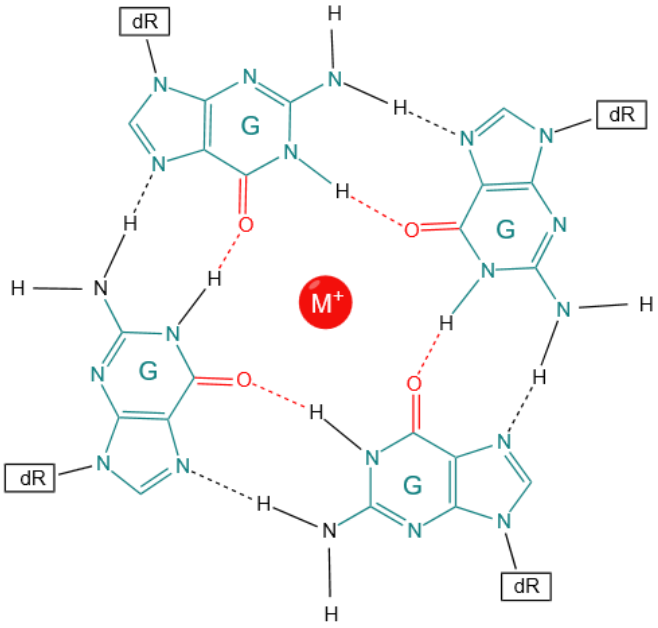


Figure 4-1 Guanine tetrad formed by Hoogsteen bond formation

The structural architecture of a G-quadruplex is quite diverse and can form different topologies based on factors such as the chemical environment, loop length (83, 84), and localization in the sequence or structure molecularity (85). The stacking of the guanine tetrads is bound by the loops of nucleotide bases of variable sizes which determine the folding of the secondary structure.

4.2.1 Functional role of G4 regions

G4 sequences do not always form G4 structures, which can additionally be dependent upon physiological conditions and methylation patterns guided by chromatin structure for their formation (57). However, when they do, they can alter several functional roles. One such perturbed function is transcription which is affected by stalling the replication fork (150-152). In cells that do not have the normal DNA repair machinery, this causes down regulation of several genes and cell cycle arrest (153).

Additionally, G4, G4 stabilizing agents and double-stranded breaks (DSB) facilitate the homologous recombination repair pathway affecting genome instability. Based on the size of the G quadruplex, thermodynamically stable short loop structures within the G4 have been extensively studied to cause instability in replication dependent processes (64). Alteration of DNA polymerase function and helicases in sites of G4 formation has been well established and is used in identification of G quadruplexes in vivo (154, 155).

While some ligands have shown binding affinity towards G quadruplex structures for treatment of cancer specific cells and transcriptional alteration (156), binding of other ligands that stabilize G4 lead to multiple DNA damage (156), micronuclei formation, delayed replication fork progression (157), and telomeric defects (158-160).

4.2.2 Mutations within G4 regions

DNA lesions can be mutagenic or lethal, and when they are found in G quadruplex regions, they can alter the secondary structure by changing the guanine tract base pairing or altering the composition of the loop region. A single nucleotide mutation in the G4 present in the promoter region of c-MYC has been shown to change transcription in vivo (132). Mass spectroscopy studies using single nucleotide substitution in the central block of parallel G4 forming sequencing found a deleterious effect of G quadruplex stability and association rate (161). A trinucleotide CGG repeat expansion in the untranslated region of the FMR1 gene has been linked with ataxias and Fragile X Syndrome (162). A T→C SNP at the GC rich region of Apolipoprotein E (APOE) is known to vary G quadruplex structure and has been linked to onset of Alzheimer's Disease (163). It has been proposed that specific helicases promote genomic stability by actively resolving G4 structures which can be altered by the addition of G4 stabilization ligands presence of specific DSBs (153, 154, 164). Baral et al. identified several eQTL variants in potential G-quadruplex regions (165). Changes in loops of G quadruplexes and stability led to a significant alteration in gene expression among individuals further fueling the structural role of G4s in regulation and binding of transcription factors (164).

Selective mutation of the G rich region to disrupt the G4 structure has been found to alter transcription. The mutation further can hinder the recruitment of transcription factors that overlap the G rich region and function as recognition motifs or bind to the G quadruplex region. Siddiqui-Jain et al. demonstrated that a single G→A mutation destabilizes the folding of G4 in the Pu27 region of MYC which is otherwise repressed, resulting in a threefold increase in transcriptional activity of the gene in tumor cell lines (132). Studies

related to 8-oxoguanine in the G quadruplex established the presence of G-A and guanine abasic lesions in G quadruplex structures based on the position in the sequence which can destabilize the secondary structure leaving the unfolded sequence prone to cleavage, leading to further instability in the telomere region (166).

4.2.3 Study motivation

Given the roles that G4 regions and mutations within them play in transcriptional and translational control, we set out to identify the impacts of mutations in G-quadruplex regions and patterns associated with the variants. This was aided by looking at variants annotated in the COSMIC (167) and CLINVAR (168) databases, which represent mutations associated with cancers (COSMIC) or other clinical relevance (CLINVAR). We identified somatic and germline variations representing SNVs occurring within G quadruplex sequences. Because of their high stability and increased cellular uptake, G quadruplex sequences have interesting diagnostic and therapeutic functions. Understanding how known variants in the genome confer stability or disrupt the G quadruplex sequences will allow a better understanding of G4 structure and function.

4.3 MATERIAL AND METHODS

4.3.1 Putative and validated G4 identification

Quadparser version 2 (4) with the default parameters was used to identify 175,778 putative G quadruplex regions in the human genome hg38 assembly across both strands. Experimentally validated G4 regions were obtained from an experiment utilizing a method called G4 Seq (GEO accession GSE63874) previously performed by Chambers, et al. (83). The intersection between the putative and experimental G quadruplexes was found using BEDTOOLS (112).

4.3.2 SNP identification

Cancer-specific curated somatic mutations from the COSMIC database (167) were used for the analysis. COSMIC contains 22,996,215 distinct single nucleotide variants (SNVs) (19,721,019 non-coding variants (NCV) and 5,977,977 coding) from 1.4 million tumor samples. An additional 550,239 germline SNVs from other clinically relevant diseases and disorders were obtained from CLINVAR (168) version (clinvar_20200203.vcf.gz).

For both sets of data, a two-pass analysis was performed. In the first pass, overlaps between the SNVs and putative G quadruplex regions were found to determine potential loss of a G quadruplex structure due to mutations. In the second pass, mutations leading to a G in regions with flanking guanines that could result in the gain of a G quadruplex were detected. In each case, a variant call format (VCF) file describing the coding and non-coding mutations was obtained from COSMIC (167) and CLINVAR (168). Using the VCF, SNVs were filtered using bcftools (169), with insertion and deletion events (INDELs) removed.

4.3.3 Identification of SNPs affecting G4 formation

A window 30 bases upstream and 30 bases downstream of each variant was used to search for putative G quadruplex sequences. Prospective G4 regions were compared with the Vienna Package RNAfold v2.4.8 to determine changes in G quadruplex stability as a result of the variant (77). The values of ΔMFE (minimum free energy) and ΔED (ensemble diversity) were used as the determining metrics. MFE calculates the stability of the sequence structure based on the binding propensities while centroid distance to ensemble provides the diversity of the sequence structure and alternate structures it can form. G4hunter was also used to compare the G4 scores and the formation of pG4 (5).

Based on the location of a specific SNV inside a G4 region, the relative location of the mutation was calculated as the position of the SNV in the G4 divided by the total length of the sequence. In terms of multiple potential G4 regions, the whole region was used as a single sequence and the relative location of the mutation was calculated. Each SNV was converted into a 3-mer based on its context and changes in the k-mer resulting in a broken GGG quad structure were calculated. For each 3-mer, the number of changes was calculated using one base before and after the location of the variant, respectively. In addition, the SNV in the context of loop and guanine tetrads was analyzed based on the trinucleotide context. The R package annotatr was used for randomized background counts for each annotation (170).

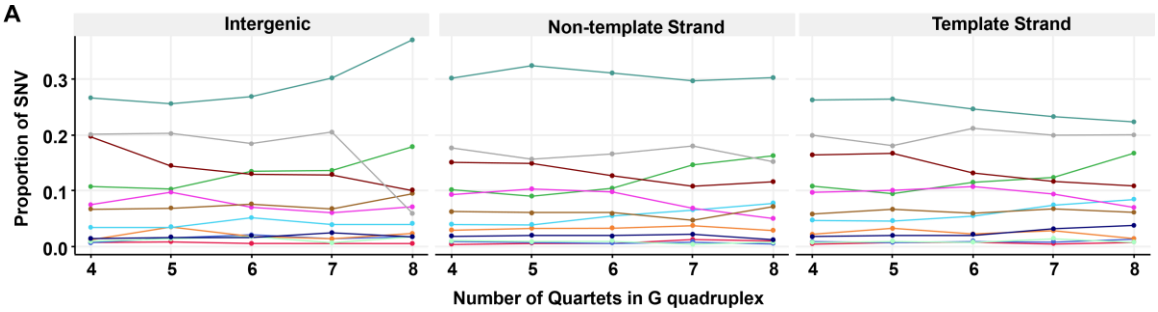
4.3.4 Enrichment analysis

Based on the G4 identified, the hg38 coordinates of the G4 were used to find the enrichment of transcriptional factors using Remap (171) for Hep-G2, K562, HEK293 and HEK293T cell lines. Further, enrichment analysis of the genes with individual mutations were selected based on the number of SNV per gene, effect of SNVs on the G quadruplex, G4 per gene and samples as specified in the result. Functional annotation enrichment of genes was carried out using DAVID functional annotation (172) while the enrichment analysis of TFs involved was carried out using STRING database (173). In order to analyze the disruption of motifs by each SNV, the R package motifbreakR (174) was used.

4.4 RESULTS

4.4.1 COSMIC somatic mutations

Using the COSMIC database, 37,515 (0.16% of all COSMIC mutations) distinct single nucleotide somatic mutations were identified within 26,504 pG4 regions from 9,693 genes, 8,998 of which were determined to be protein coding according to ENSEMBL hg38 annotations. The remaining genes were identified as lncRNA (n=540) or miRNA (n=111). The most frequently observed mutation observed in the COSMIC filtered dataset was the transition event G→ A (28%) followed by the transversion event T→G (18%) (Figures 2A and 2B). The variants were expected to be high in number for G→A and G→T (15%) mutations; however, we also identified the T→G transversion to be high in these regions compared with A→G transitions. Comparatively, higher G/C→A/T variants in intragenic CpG islands has been observed due to the spontaneous deamination to the cytosine hypermethylated CpGs within these regions (175, 176). However, the effect of these mutations is less studied across G4 regions. We found a lower transition:transversion ratio (p = 0.00001) occurring in the G4 region (1.02), compared to the overall mutations in COSMIC database (1.146) (Appendix Table B 1 & Appendix Table B 2).



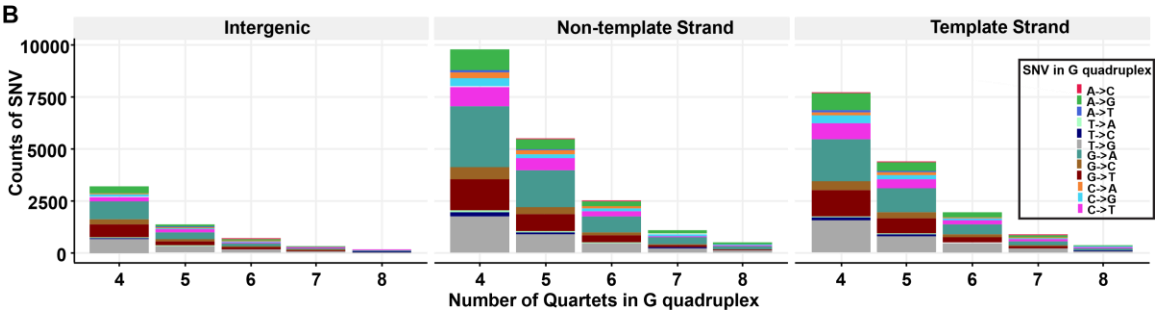


Figure 4-2: Composition of SNVs in G4 regions from the COSMIC database. Shown is (A) proportion and (B) count of selected SNVs

Based on the G4Hunter (5) and RNAfold (77) results, we compared the number of SNV events that breaks the G4 structure and changes in the thermodynamic stability based on the minimum fold energy of each sequence. We found 7,236 (19.2% of variants in G4) of the SNVs within the G4Hunter identified G4s result in the loss of a G quadruplex, while 2,728 SNVs led to the gain of a new G quadruplex (Figure 3A, Appendix Table B 3).

4.4.2 CLINVAR germline mutations

Using the CLINVAR database, 5,026 SNVs were identified in pG4 regions out of which 2,155 intersected with experimental G4. Most of these G4 mutations occurring in exons (50%, n=2,559). The remaining variants are found in introns (24%, n=1,251), promoters (11%, n=554), and transcription termination regions (3.5%, n=179). Overall, 13.92% (700 variants) were associated with non-coding RNA, and 84% (n=4,265) SNVs occur in protein coding regions (Figures 3B-3D, Appendix Table B 4).

4.4.3 Change to G4 stability

RNAfold was used to differentiate the impact of the variant on the stacking. Variants were classified based on the change in stability and formation of available guanines for stacking by combining the sequence pattern analysis of G4Hunter with thermodynamic parameters from RNAfold (Figures 4A-4F). The majority of the SNVs (81%) did not

affect the GGG stacking in such a way that the formation of tetrads of guanines was not possible. Though complete breakage of structure does not occur, we found a decrease in the stability of the G quadruplex structure in 40% of these variants. This is due to the presence of additional guanines in the loop that aid the conformational diversity of G quadruplex which can act as extra base for stacking (Figure 4E). We found 10,435 SNVs across the combined COSMIC and CLINVAR mutations that increase the stability (lower the MFE relative to the reference sequence) while 12,061 SNVs brought no change to the MFE. An additional 15,019 variants destabilize the G4. Transversions were more likely to change the structure of the G quadruplex region without disrupting the G stacks and increasing the thermodynamic stability of the structure (17%) compared to transitions (10%). Additionally, transition mutations were found to destabilize the G4 structure at a higher rate (22%) compared to transversions (17%) (Table 4-1).

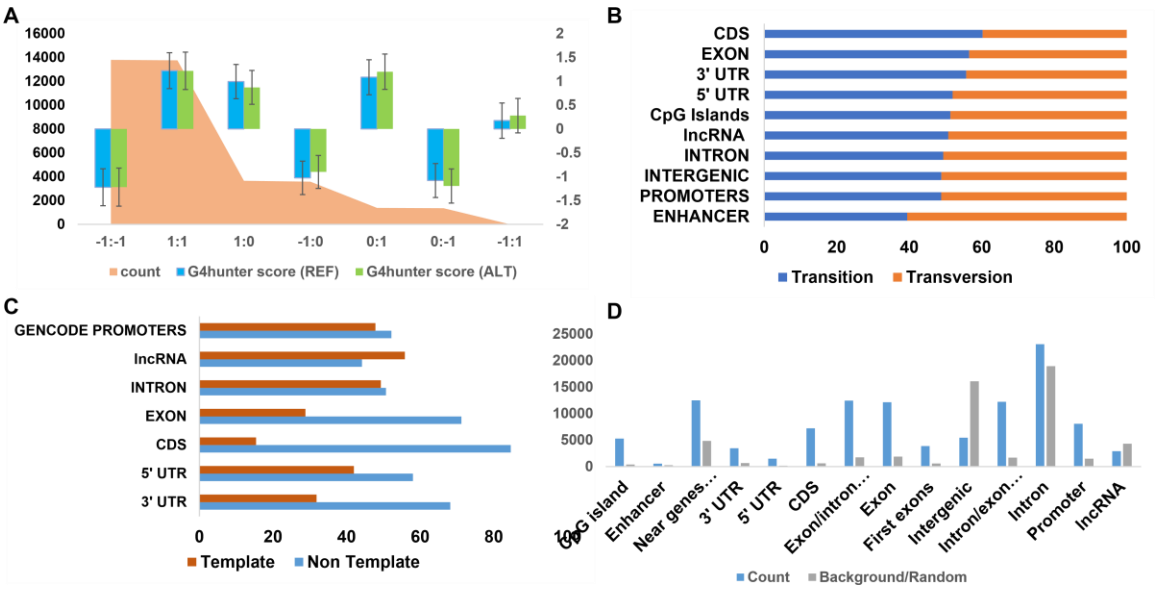


Figure 4-3 Identified G4 variants relative to functional annotations. Shown is (A) count of change in pG4 with G4Hunter score across both strands before and after mutation (0: absence of pG4; 1: presence of G4 in the forward strand; -1: presence of G4 in the reverse strand); (B) percentage of the type of mutation across annotations from the COSMIC database; (C) percentage of SNVs that occur in a G4 region across the template and non-template strand for functional annotation groups; and (D) count of variants in

functional annotations against randomized background count of variants in the human genome.

4.4.4 Variants in transcript regions

We find comparatively higher number of mutations in G4 forming exonic regions in 5'UTR, 3' UTR and CDS regions of protein coding genes when the G4 in formed in the strand opposite the transcribed gene (Figure 3C). The count of SNV around G4 forming regions in intron and promoter regions were proportionate with the transcript opposite or in same strand as the transcript. This shows selection pressure of variants around exon regions as compared to the non- coding regions. Previously, it has been hypothesized the formation of G4 in either strand within the transcribed region, along with nascent RNA would lead to formation of DNA:RNA hybrid R loops in the G quadruplex which results in physically halting the polymerase movement inhibiting further rounds of transcription (177). Additionally, G4 formed on the non-template strand could interfere with the reannealing of the DNA strands increasing the stability of the R loop hybrid.

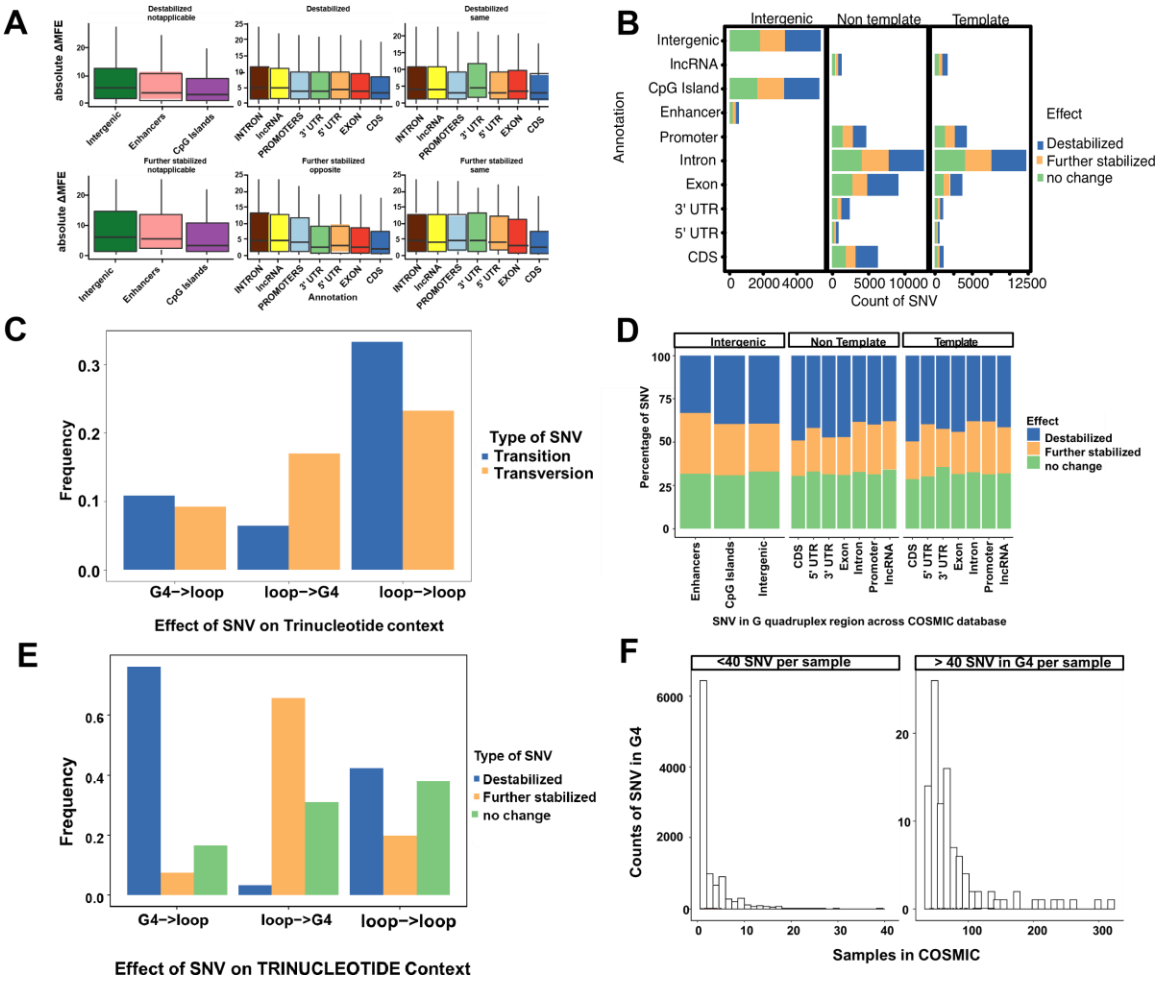


Figure 4-4Thermodynamic changes associated with variants in various genomic features. (A) Non-zero delta MFE of G4 across different annotation for destabilized and further stabilized effect by SNV. (B) Count of variants across different regions of the genome and in strand specific or alternative to the coding gene. (C) Proportion of effect on G4 based on the type of mutation. (D) Proportion of variants across different annotation. (E) Proportion of trinucleotide context based on the type of effect on the G4 sequence. (F) Histogram of variants by sample

4.4.5 Gene component variants

Comparing mutations in different functional groups, G→A mutations are elevated in exons (35.18%) and decreased proportion in promoter region (26.87%). We find a lower percentage of T→G mutations in G4 regions occurring in exons (11.74%) compared to intron, promoter (18%), enhancers (29.84%) and intergenic regions (18%). This pattern of low T→G variants coincides with counts in the CDS region while the 5' UTR show

increased T→G variants (16%). G→A SNVs are found less in enhancers (19%) which are distant from the transcription site and deamination occurring in upstream of transcription site does not affect the G4 region but comparatively have the highest proportion of T→G (29.84%) mutations (Table 4-2).

Previously, higher counts of C→T over G→A variants were identified in the non-template strand, which was hypothesized due to cytosine deamination in the nearby 2kb downstream of 5' end of genes due to higher exposure of single stranded DNA (178). However, we predict the implication of these variants occuring within G quadruplex regions and cause a conformational shift in its structure leading to alteration in expression and binding patterns across these regions. Additionally, 8-oxoguanine formation in G quadruplex binding Sp1 proteins is an important regulator for adipose tissue development and GC rich promoter region with transcription factor sites activating proportional to increasing 8 oxo-G abundance (179).

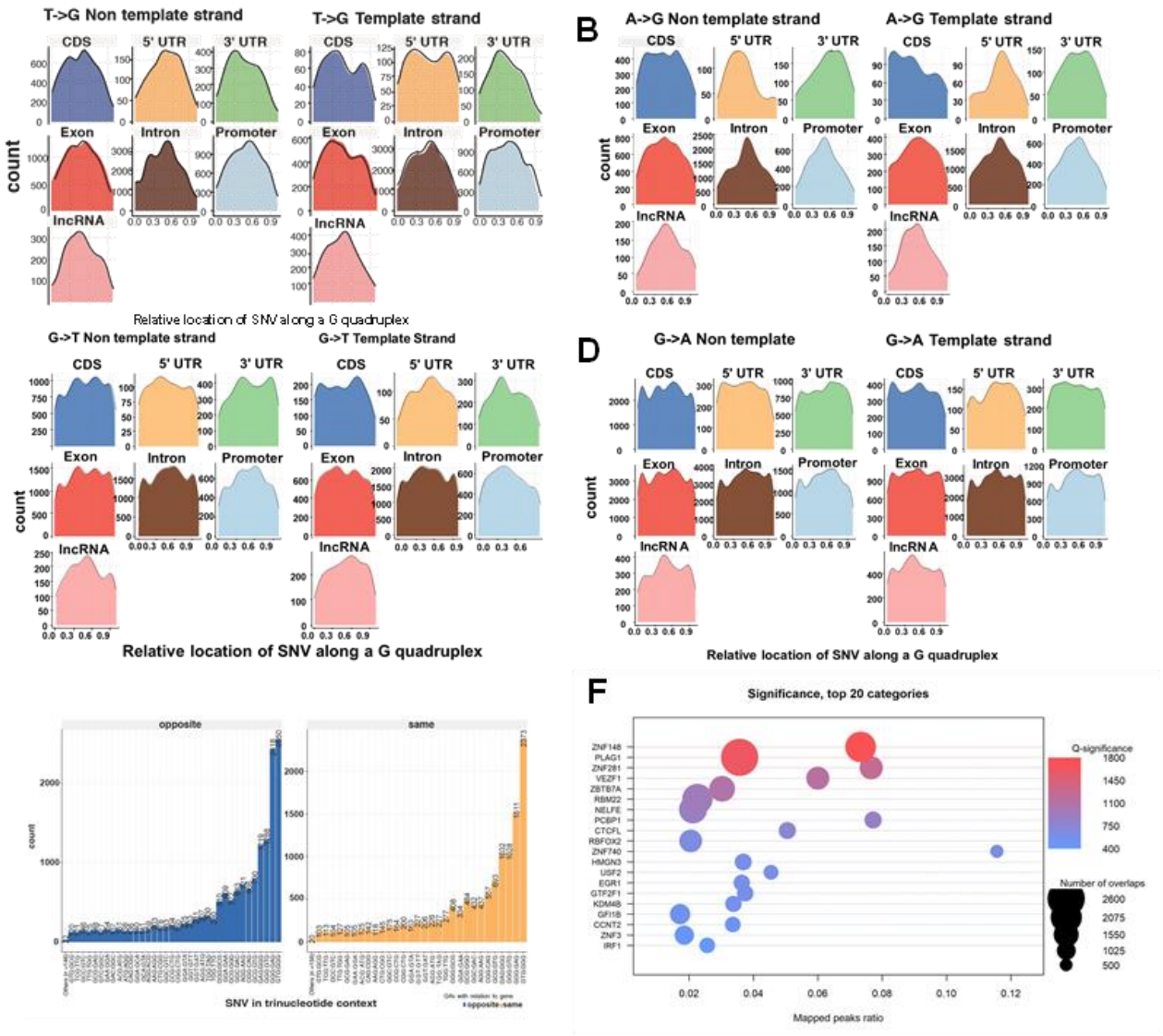


Figure 4-5 . Distribution of SNVs across the G4 regions on the non-template and template strand. Shown are the results for (A) T→G variants; (B) A→G variants; (C) G→T variants; and (D) G→A variants. (E) Distribution of SNVs in trinucleotide contexts relative to the opposite or same strand as the corresponding gene. (F) Significance of the top 20 transcription factors and their genome-wide binding sites.

4.4.6 Enrichment analysis

4.4.6.1 Gene Ontology.

Gene Ontology (GO) enrichment analysis was performed for biological processes (GO:BP) and cellular components (GO:CC). A total of 424 GO:BP categories were determined to be significant (FDR ≤ 0.05) overall (

Appendix Figure B 1;

Appendix Table B 5), while 425 significant GO:BP

enrichments were found for COSMIC alone (Appendix Figure B 2: Enriched GO:BP terms for G4 mutations.) and 48 were found for CLINVAR (Appendix Table B 6). When this was further broken down into mutations resulting in a loss of a G4, we found 205 significant GO:BP overall (

GO ID	GO Description	Universe	COSMIC and CLINVAR	Adjusted P-value
GO:0032502	developmental process	4584	1138	3.23E-24
GO:0048856	anatomical structure development	4152	1043	2.82E-23
GO:0007399	nervous system development	1648	488	1.15E-22
GO:0009653	anatomical structure morphogenesis	1871	539	2.18E-22
GO:0048731	system development	2976	784	9.30E-22
GO:0007275	multicellular organism development	3266	842	1.00E-20
GO:0032501	multicellular organismal process	5319	1270	1.40E-20
GO:0030154	cell differentiation	2860	742	3.10E-18
GO:0048699	generation of neurons	969	309	3.31E-18
GO:0048869	cellular developmental process	2879	745	5.24E-18
GO:0016043	cellular component organization	5370	1259	6.78E-17
GO:0022008	neurogenesis	1096	335	1.06E-16
GO:0030182	neuron differentiation	923	292	1.76E-16
GO:0071840	cellular component organization or biogenesis	5539	1277	2.38E-14
GO:0023051	regulation of signaling	2733	694	3.06E-14
GO:0010646	regulation of cell communication	2727	692	4.16E-14
GO:0048666	neuron development	729	236	4.22E-14
GO:0000904	cell morphogenesis involved in differentiation	495	175	7.97E-14
GO:0048468	cell development	1355	385	9.16E-14
GO:0048513	animal organ development	2176	567	3.84E-13
GO:0023052	signaling	5190	1196	9.50E-13
GO:0032989	cellular component morphogenesis	549	186	1.02E-12
GO:0007154	cell communication	5221	1200	2.02E-12
GO:0031175	neuron projection development	656	212	2.53E-12
GO:0048667	cell morphogenesis involved in neuron differentiation	381	140	3.85E-12
GO:0034330	cell junction organization	518	176	4.81E-12
GO:0000902	cell morphogenesis	718	226	6.91E-12
GO:0048812	neuron projection morphogenesis	441	155	9.70E-12
GO:0007010	cytoskeleton organization	1309	365	1.42E-11
GO:0009966	regulation of signal transduction	2464	621	1.59E-11
GO:0009887	animal organ morphogenesis	582	190	2.87E-11
GO:0007155	cell adhesion	1216	342	3.26E-11
GO:0120036	plasma membrane bounded cell projection organization	1144	325	3.84E-11
GO:0120039	plasma membrane bounded cell projection morphogenesis	455	157	3.85E-11
GO:0048858	cell projection morphogenesis	459	157	8.90E-11
GO:0030030	cell projection organization	1164	328	9.33E-11
GO:0032990	cell part morphogenesis	469	159	1.42E-10
GO:0030029	actin filament-based process	717	219	5.92E-10
GO:0061564	axon development	323	118	1.06E-09
GO:0007409	axonogenesis	298	111	1.25E-09
GO:0050808	synapse organization	275	104	2.31E-09
GO:0050793	regulation of developmental process	1810	465	4.44E-09
GO:0051128	regulation of cellular component organization	1929	490	6.77E-09
GO:0007165	signal transduction	4776	1078	4.27E-08
GO:0016477	cell migration	1210	325	7.79E-08
GO:0048870	cell motility	1362	359	8.30E-08
GO:0098609	cell-cell adhesion	743	217	8.39E-08

GO:0051716	cellular response to stimulus	5982	1316	9.92E-08
GO:0035556	intracellular signal transduction	2168	534	1.26E-07
GO:0048583	regulation of response to stimulus	3327	777	1.74E-07
GO:0051239	regulation of multicellular organismal process	2117	522	1.84E-07
GO:0099537	trans-synaptic signaling	501	157	2.02E-07
GO:0099536	synaptic signaling	522	162	2.28E-07
GO:0009888	tissue development	1239	329	2.43E-07
GO:0098916	anterograde trans-synaptic signaling	495	155	2.81E-07
GO:0007268	chemical synaptic transmission	495	155	2.81E-07
GO:0065007	biological regulation	10721	2221	3.01E-07
GO:0072359	circulatory system development	729	211	3.69E-07
GO:0065008	regulation of biological quality	2937	692	5.65E-07
GO:0050794	regulation of cellular process	9523	1993	7.40E-07
GO:0007267	cell-cell signaling	1269	333	7.84E-07
GO:0003008	system process	1358	351	1.64E-06
GO:0007417	central nervous system development	584	173	2.68E-06
GO:0009987	cellular process	14783	2936	4.03E-06
GO:0040011	locomotion	1075	285	5.65E-06
GO:0050789	regulation of biological process	10085	2089	8.77E-06
GO:1905114	cell surface receptor signaling pathway involved in cell-cell signaling	388	123	1.03E-05
GO:0030036	actin cytoskeleton organization	637	183	1.15E-05
GO:0050804	modulation of chemical synaptic transmission	252	88	1.18E-05
GO:0030048	actin filament-based movement	113	49	1.44E-05
GO:0099177	regulation of trans-synaptic signaling	253	88	1.46E-05
GO:0006812	cation transport	886	240	1.55E-05
GO:0030001	metal ion transport	678	192	1.58E-05
GO:0010975	regulation of neuron projection development	288	97	1.61E-05
GO:0060047	heart contraction	187	70	1.80E-05
GO:0006811	ion transport	1187	307	1.99E-05
GO:0048518	positive regulation of biological process	5294	1158	2.05E-05
GO:0034329	cell junction assembly	341	110	2.35E-05
GO:0032879	regulation of localization	1615	400	2.54E-05
GO:0044057	regulation of system process	392	122	3.78E-05
GO:0007411	axon guidance	169	64	4.59E-05
GO:0097485	neuron projection guidance	169	64	4.59E-05
GO:0051179	localization	4343	964	4.59E-05
GO:0048522	positive regulation of cellular process	4704	1036	5.30E-05
GO:0007507	heart development	338	108	5.46E-05
GO:0006936	muscle contraction	260	88	6.00E-05
GO:0055085	transmembrane transport	1060	276	6.09E-05
GO:0034220	ion transmembrane transport	816	221	6.22E-05
GO:0042391	regulation of membrane potential	319	103	6.34E-05
GO:0048523	negative regulation of cellular process	3896	872	6.65E-05
GO:0003015	heart process	193	70	7.53E-05
GO:0007166	cell surface receptor signaling pathway	2271	536	8.29E-05
GO:0003012	muscle system process	305	99	8.73E-05
GO:1902531	regulation of intracellular signal transduction	1429	356	9.28E-05
GO:0061061	muscle structure development	407	124	1.07E-04
GO:0040012	regulation of locomotion	842	225	1.46E-04
GO:0098655	cation transmembrane transport	656	182	1.81E-04
GO:0048646	anatomical structure formation involved in morphogenesis	751	203	2.65E-04
GO:0048519	negative regulation of biological process	4381	964	2.73E-04
GO:0048729	tissue morphogenesis	348	108	2.78E-04
GO:0007416	synapse assembly	126	50	2.85E-04
GO:0022603	regulation of anatomical structure morphogenesis	700	191	3.10E-04
GO:0051960	regulation of nervous system development	257	85	3.10E-04
GO:0030334	regulation of cell migration	767	206	3.51E-04
GO:0050905	neuromuscular process	73	34	3.51E-04
GO:0031589	cell-substrate adhesion	290	93	3.99E-04
GO:0050896	response to stimulus	7117	1503	4.02E-04
GO:0031344	regulation of cell projection organization	467	136	4.37E-04
GO:0003013	circulatory system process	443	130	5.12E-04
GO:0065009	regulation of molecular function	2121	498	5.42E-04
GO:0006810	transport	3620	807	5.65E-04
GO:0120035	regulation of plasma membrane bounded cell projection organization	453	132	6.34E-04
GO:0008015	blood circulation	363	110	7.79E-04
GO:0030111	regulation of Wnt signaling pathway	274	88	7.79E-04
GO:0007420	brain development	384	115	8.24E-04

GO:0051130	positive regulation of cellular component organization	844	221	9.67E-04
GO:2000145	regulation of cell motility	818	215	1.04E-03
GO:0006996	organelle organization	3147	708	1.04E-03
GO:0070252	actin-mediated cell contraction	86	37	1.15E-03
GO:0051056	regulation of small GTPase mediated signal transduction	238	78	1.44E-03
GO:0050877	nervous system process	755	200	1.46E-03
GO:0051234	establishment of localization	3775	834	1.55E-03
GO:0060048	cardiac muscle contraction	111	44	1.57E-03
GO:0051094	positive regulation of developmental process	967	247	1.64E-03
GO:0098662	inorganic cation transmembrane transport	572	158	1.65E-03
GO:0051963	regulation of synapse assembly	58	28	1.69E-03
GO:0060322	head development	402	118	1.74E-03
GO:0001667	ameboidal-type cell migration	333	101	2.14E-03
GO:0048638	regulation of developmental growth	167	59	2.18E-03
GO:0007167	enzyme-linked receptor protein signaling pathway	795	208	2.19E-03
GO:0008016	regulation of heart contraction	164	58	2.57E-03
GO:0051049	regulation of transport	1327	324	2.69E-03
GO:0006941	striated muscle contraction	142	52	2.76E-03
GO:0045595	regulation of cell differentiation	1129	281	2.82E-03
GO:0097435	supramolecular fiber organization	710	188	3.17E-03
GO:0035637	multicellular organismal signaling	128	48	3.20E-03
GO:0098660	inorganic ion transmembrane transport	622	168	3.22E-03
GO:0001505	regulation of neurotransmitter levels	143	52	3.47E-03
GO:0048598	embryonic morphogenesis	308	94	3.66E-03
GO:0010647	positive regulation of cell communication	1374	333	3.80E-03
GO:0006816	calcium ion transport	313	95	4.16E-03
GO:0009967	positive regulation of signal transduction	1255	307	4.32E-03
GO:0099565	chemical synaptic transmission, postsynaptic	51	25	4.41E-03
GO:0009719	response to endogenous stimulus	1076	268	4.62E-03
GO:0099587	inorganic ion import across plasma membrane	112	43	5.31E-03
GO:0098659	inorganic cation import across plasma membrane	112	43	5.31E-03
GO:0009790	embryo development	505	140	5.37E-03
GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	195	65	5.93E-03
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	511	141	6.44E-03
GO:0023057	negative regulation of signaling	1067	265	6.57E-03
GO:1903522	regulation of blood circulation	188	63	6.84E-03
GO:0050807	regulation of synapse organization	124	46	7.05E-03
GO:0001508	action potential	117	44	7.54E-03
GO:0023056	positive regulation of signaling	1380	332	7.60E-03
GO:0051962	positive regulation of nervous system development	147	52	8.38E-03
GO:0198738	cell-cell signaling by wnt	343	101	8.47E-03
GO:0010631	epithelial cell migration	261	81	8.86E-03
GO:0043269	regulation of ion transport	458	128	9.01E-03
GO:0016358	dendrite development	155	54	9.18E-03
GO:0050803	regulation of synapse structure or activity	129	47	9.61E-03
GO:1903115	regulation of actin filament-based movement	32	18	9.75E-03
GO:1901888	regulation of cell junction assembly	144	51	9.80E-03
GO:0040008	regulation of growth	416	118	9.86E-03
GO:0007612	learning	53	25	1.00E-02
GO:0022607	cellular component assembly	2547	575	1.03E-02
GO:0010648	negative regulation of cell communication	1060	262	1.06E-02
GO:0042221	response to chemical	2943	656	1.07E-02
GO:0086001	cardiac muscle cell action potential	73	31	1.16E-02
GO:0071495	cellular response to endogenous stimulus	974	243	1.17E-02
GO:0090130	tissue migration	267	82	1.17E-02
GO:0090132	epithelium migration	263	81	1.19E-02
GO:0090596	sensory organ morphogenesis	119	44	1.21E-02
GO:0048738	cardiac muscle tissue development	138	49	1.34E-02
GO:0007269	neurotransmitter secretion	91	36	1.45E-02
GO:0099643	signal release from synapse	91	36	1.45E-02
GO:0016055	Wnt signaling pathway	339	99	1.47E-02
GO:0040007	growth	515	140	1.52E-02
GO:0048589	developmental growth	273	83	1.55E-02
GO:0098703	calcium ion import across plasma membrane	36	19	1.76E-02
GO:0099003	vesicle-mediated transport in synapse	121	44	1.91E-02
GO:0014706	striated muscle tissue development	144	50	2.13E-02
GO:0016310	phosphorylation	1444	342	2.13E-02
GO:0007517	muscle organ development	179	59	2.24E-02

GO:0034762	regulation of transmembrane transport	385	109	2.30E-02
GO:0006836	neurotransmitter transport	137	48	2.39E-02
GO:0044087	regulation of cellular component biogenesis	774	197	2.41E-02
GO:0086003	cardiac muscle cell contraction	62	27	2.50E-02
GO:0007611	learning or memory	115	42	2.61E-02
GO:0009968	negative regulation of signal transduction	1009	248	2.74E-02
GO:0090257	regulation of muscle system process	157	53	2.86E-02
GO:0035249	synaptic transmission, glutamatergic	66	28	3.09E-02
GO:0060560	developmental growth involved in morphogenesis	135	47	3.44E-02
GO:0002009	morphogenesis of an epithelium	275	82	3.61E-02
GO:0007264	small GTPase mediated signal transduction	389	109	3.61E-02
GO:0002027	regulation of heart rate	84	33	3.80E-02
GO:0043542	endothelial cell migration	194	62	3.82E-02
GO:0051668	localization within membrane	570	150	4.30E-02
GO:0086091	regulation of heart rate by cardiac conduction	41	20	4.31E-02
GO:0060828	regulation of canonical Wnt signaling pathway	211	66	4.44E-02
GO:0051965	positive regulation of synapse assembly	35	18	4.46E-02
GO:0048588	developmental cell growth	129	45	4.73E-02
GO:0099504	synaptic vesicle cycle	114	41	4.75E-02
GO:0042127	regulation of cell population proliferation	1218	291	4.82E-02
GO:0050890	cognition	156	52	4.83E-02
GO:0007158	neuron cell-cell adhesion	16	11	4.94E-02

), 75 for COSMIC (Appendix Table B 8) and 25 for CLINVAR (Appendix Table B 9). Among the COSMIC enrichments were synapse organization, axonogenesis, neuron projection guidance, axon guidance, cell-substrate adhesion, neuromuscular process, regulation of neuron projection development, and xenobiotic glucuronidation. One example gene is the App transcript, which is involved in synapse formation and function in the developing brain. The App transcript is transported to neuronal dendrites, where the transmembrane APP protein plays an integral role in synapse formation and function. However, the translation of App is repressed by the binding of the Fragile X Mental Retardation protein (FMRP) to G-quadruplexes in the App coding region. This repression is thought to occur through direct interaction with the ribosomes, resulting in stalled ribosomal progression on the mRNA (180). Past studies have also shown that this repression can be relieved by synaptic activation of metabotropic glutamate receptors, specifically mGluR5 receptors. This results in the release of FMRP and an increase in APP translation (181).

The CLINVAR enrichments included a number of muscular-related processes, such as striated muscle contraction, neuromuscular process, actin-mediated cell contraction,

cardiac conduction, cardiac muscle cell action potential, cardiac muscle cell contraction, membrane depolarization, regulation of actin filament-based movement, muscle tissue morphogenesis, muscle organ morphogenesis, regulation of heart rate, regulation of action potential, cardiac muscle cell action potential involved in contraction, cell communication involved in cardiac conduction, regulation of striated muscle contraction, sensory perception of sound, regulation of heart rate by cardiac conduction, musculoskeletal movement, multicellular organismal movement, transmission of nerve impulse, cardiac muscle tissue morphogenesis, skeletal muscle contraction, and ventricular cardiac muscle cell action potential. Variants leading to a gain of a G4 result in 115 GO:BP enrichments overall (Appendix Table B 10), 22 for COSMIC (**Error! Reference source not found.**) and 2 for CLINVAR (Appendix Table B 11). Among the COSMIC enrichments from genes gaining G4 due to the variants are positive regulation of transcription by RNA polymerase II and actin cytoskeleton organization while the CLINVAR enrichments genes based on loss of G4 include system development, action potential, and cardiac muscle cell action potential. Loss of G4 using COSMIC resulting in similar enriched GO terms as did G4 loss in CLINVAR, included muscle contraction, muscle system process, cardiac muscle contraction, striated muscle contraction, heart contraction, heart process, cardiac muscle cell contraction, cardiac muscle cell action potential involved in contraction, actin-mediated cell contraction, actin filament-based movement, cardiac muscle cell action potential, regulation of heart contraction, action potential, and multicellular organismal signaling.

Among the enriched categories detected were PDZ domain proteins (GIPC2, GRIDZIP, LIMK2, PDLIM7, PDZD7, WHRN, SIPA1L3, PRX, MYO1BA, MAGI2, and MAST)

with G4 in coding regions and variants affecting the RGG (arginine-glycine-glycine) domain or G quadruplex stability negatively. Proteins with RGG repeats have been known to bind to G4 structures. Variants in these regions affecting the G4 stability further could affect downstream binding.

GO:CC enrichments yield 128 significant categories overall (129 for COSMIC and 14 for CLINVAR) (Appendix Figure B 4; Appendix Figure B 5, Appendix Figure B 6, Appendix Table B 12). Among the enriched GO:CC categories detected in COSMIC are collagen containing extracellular matrix, and cell-cell contact zone indicating mutations in these genes affect the adhesion of cells to the extracellular matrix. Other enriched GO:CC terms in CLINVAR include I band, sarcolemma, and myofilament Z disc. Enriched GO:CC terms from loss of G4 using CLINVAR database include collagen trimer and PCSK9-LDLR complex.

4.4.6.2 KEGG metabolic pathways.

KEGG enrichment yielded 96 significant pathways overall as well as 91 COSMIC and 11 CLINVAR (Appendix Figure B 7, Appendix Figure B 8, Appendix Figure B 9). Those leading to a loss of G4 yielded 33 significant categories, including 12 and 5 for COSMIC and CLINVAR, respectively (Appendix Table B 17, Appendix Table B 18 & Appendix Table B 19). Among the enriched categories for genes with loss of G4 within CLINVAR are hypertrophic cardiomyopathy, dilated cardiomyopathy, arrhythmogenic right ventricular cardiomyopathy, adrenergic signaling in cardiomyocytes, and acute myeloid leukemia. KEGG enrichments for a gain of G4 resulted in 31, 3, and 0 for overall, COSMIC and CLINVAR respectively (Appendix Table B 20 & Appendix Table B 21). The enriched

terms from gain of G4 in CLINVAR variants are melanoma, phospholipase D signaling pathway and cocaine addiction.

4.4.6.3 INTERPRO protein domains.

INTERPRO enrichment yielded 23, 23, and 2 enrichments overall, for COSMIC and CLINVAR respectively. (Appendix Table B 22,Appendix Table B 23 & Appendix Table B 24). Included were Src homology-3 domain (n=69 FDR=3.73E-03) and Pleckstrin homology-like domain (PH) (n=147, FDR=1.30E-10). The binding affinity of PH domains with the exception for some binding phosphoiniositides with high affinity, majority have unique recognition domains and are known for functional plasticity (182, 183).

4.4.6.4 Transcription Factors.

We identified the enrichment of transcription factors (TFs) including NFKB1, ZFX, MBD3, ASX1, SUZ12, NCOR1, HMGN3, USF2, EGR1, GTF2F1, KDM4B, HNRNPH1, HNRNPL, NONO, TARDBP, NFATC3, KDM3A, and HOXA3 among others (Appendix Figure B 10, Appendix Figure B 11, Appendix Figure B 12; Appendix Table B 25). The majority (92%) of these had at least a G quadruplex in their gene structure working in a feed forward regulation of genes. We identified these variants break the motifs for transcription binding sites.

4.4.7 Trinucleotide context mutation in G quadruplex sequence

Based on the nucleotide context one base pair before and after the mutation, we identified 79% of the variants to be affecting the loop region and 23% of the SNV after the change leads to the formation of GGG in regions with G(A|C|T)G. We find 36% (n=6,810) of the transversion mutations are T→G, while 21% (n=4,070) of the transitions are A→G. This change becomes more prominent, T→G mutations occurring in context of GTG→GGG

occurs in 14% of the SNVs leading to formation of stable G tetrad while GAG→GGG occurs as 6% of the variants. Interestingly, the destabilization of GGG region occurs by GGG→GAG transition in 11% of SNVs (Figure 4-5e). Previously, it has been reported that the GGG exhibits context dependent specific mutational patterns that preserve the potential for G4 formation (184). We find G→A mutations to be approximately 29% of the total SNVs in the selected G quadruplex regions, with 26% (n=4,144; 11% of the total) of those variants occurring in a context of GGG→GAG with implications of alteration mechanism for G quadruplex sequences (). We observe these patterns throughout different noncoding annotations, except exonic regions and CDS regions. We identify an increased propensity to be able to form stable multiple conformations with de-stabilized structures for 25% of the sequences with the variants while 14% of the variants incurred no change to the stability of the structure (Appendix Table B 26). This approach of analyzing the probable base pairing alternatives for additional guanine Hoogsteen base pairing can help identify the effects of variants within the G4 structure and hence predict the structure change and functionality of G quadruplex in various molecular processes.

Based on the position of the mutation in the G quadruplex from the starting point and the length of the sequence, the normalized position for each variant in the G quadruplex was calculated. The relative location of a variant in a G4 is defined as the position of variant divided by the length of the G4. For single nucleotide variants mutating to G either from A or T, we find similar elevated patterns in the center of the G quadruplex. T|A→G mutations show conservation of guanine in the center region with the exception of the CDS and exon in both template and non-template strand across both COSMIC and CLINVAR databases (Figure 4-5, Supplemental Figures 13-16). These changes are stricter for SNVs

within the 5'UTR across the template and non-template strand in the CLINVAR database, where we observe mutations in the relative center of the G quadruplex for T→G variants as compared to the 5' UTR COSMIC mutations where we observe mutations across the two extreme loops compared to the center. A→G mutations are observed in a higher proportion at the beginning of G quadruplex in CDS region which provides evidence for mutation pressure in the coding region preferentially protecting the coding sequence. G quadruplexes in UTRs have been reported to be under selection pressure and variants in G4 can account for instability in G4 and diseases (185).

4.5 DISCUSSION

4.5.1 Variants involved in oxidation

High occurrences of oxidized guanine in G quadruplex structures compared to duplex DNA has been previously established (186). The mutation has been suggested to occur around the external tetrads compared to the central tetrad due to radical trapping antioxidants that slow the efficiency of mutation (187). We identify an increase in counts at the middle stacking of the G quadruplex for A|T→G, implying the functional impact of types of specific variants towards the conformation of the G quadruplex. The observed elevation in counts can be accounted for the presence of tetrads available for variants from guanine to A, C and T. G quadruplex with spare tires can also form alternate structures or exclusion of certain guanosines in case of lesions or substitution in one tetrad region. Base excision repair with APP1 and OGG1 at the promoter of VEGF has shown this mechanism for formation of G quadruplex and this suggests formation of G quadruplex for other genes through a similar mechanism (188, 189). Oxidative stress occurring due to the reactive oxygen species (ROS) affects the genome stability and promote mutagenesis, senescence,

and other age-related diseases (190). Mutations in GGG regions can destabilize the stacking of guanines, altering the ionization potential affecting the ability of the G region to be further oxidized. G→A, T or C mutations can disrupt the stacking while mutations to G can further stabilize the G quadruplex or allow additional conformations for the stacking. We investigated the change of each type of SNV in each annotation to have the highest change. Based on absolute ΔMFE based on the change, we find pG4 in CDS region and CpG region are least prone to the variants while enhancers and Intergenic G4 are prone to higher stabilizing and destabilizing due to the variants (Figure 4-4A, Appendix Figure B 17). G quadruplexes in 3' UTR in the same strand of coding genes along with introns are prone to the variants and are highly stabilized or destabilized by the variants occurring in COSMIC.

We investigated which SNVs in each annotation had the highest change. The 3' UTR has a higher incidence of T→G versus A→G SNVs. This implies that T→G mutations are more likely to stabilize G quadruplexes found in the 3'UTR. Putative G4s in CDS and CpG regions are least prone to variants while enhancers and intergenic G4 are show higher changes in stabilization (both stabilizing and destabilizing) due to the SNVs (Appendix Figure B 17A, Supplemental Figure 17).

4.5.2 Role of location of SNVs in G4s

The relative position of G→T substitutions along G quadruplex sequences is shown in Figure 4-5A. The location of this mutation at the beginning of the G quadruplex can disrupt the structural formation; however, further elevated peaks at varying locations leading to additional guanines across the G4 may introduce additional tetrads in introns and exons (Figure 4-5 C and D).

The observation of an increased number of G quadruplex stacks resulting from G→T|A substitutions that break up longer runs of G's is consistent with studies that oxidation of the multiple G's occur at the start of the G quadruplex tetrads. Our results help establish that the location of mutations and the type of mutation in G rich regions alter the shape and stability of the G quadruplex structure. Previously it has been established that the most sensitive sites are located at the center tetrad (191). For mutations in CLINVAR, we observe a higher mutation rate at the start of G4. The A→G mutations associated with COSMIC variants show a considerable difference in their location relative to the G4 position (Figure 4-5). The escape of 8-oxoG from DNA repair during DNA replication can cause the misincorporation of adenine opposite 8-oxoG leading to the addition of T in place of G. For instance, a sequence with GTTAGGG with 8-oxoG at its fifth position, a mis-incorporation of the A occurs opposite G. Due to the presence of consistent Gs in the region, the true proportions of change in these regions can be hard to monitor over a range of replications. Methylation of cytosine leads to formation of 5-methyl cytosine which are residues for spontaneous transitions. Cytosine deamination might be the primary cause of C→T transition. Further, based on the context, a high proportion of T→G mutations lead to a GTG→GGG structure, supporting the stability of the G quadruplex. It presents a question of whether T→G mutations confer additional stability of G4 in cancer cells. Past studies have highlighted the conditional impact of OG mutations in base pairing with A in mutagenic MutY homolog harboring increased G→T transversions in MUTYH leading to a higher incidence rate of colorectal cancer (192-194). Thymine glycol are non-mutagenic lesions which are highly mutagenic and in regions of DSBs, are cytotoxic. In vitro studies have shown it to block replicative and repair DNA polymerases (195). The OG, thymine

glycol and abasic sites formed are repaired by the excision repair pathway. The difference in repair of oxoG sites have been observed in NEIL glycolyases which have been known to remove guanidinohydantoin (Gh) and spiroiminodihydantoin (Sp) from G quadruplex structures in promoter region over parallel conformation (196). However, the glycolysases were not able to remove the oxoG structure from the telomeric G quadruplex or the same G quadruplex structure in antiparallel structures.

4.5.3 TERT G4 mutations

A study has highlighted that the entire 67 bp G quadruplex associated with the TERT promoter was found to be completely protected from DNase cleavage while the version containing G→T variants was found to be degraded into discrete segments (197, 198). Additionally, this region folds into a compact G4 structure without any hairpins in between the G quadruplex stacks. However, based on DMS footprinting studies, formation of hairpins has been predicted (199). Overall, we identified 52 possible SNVs in 39 base pair locations in this 67 bp G quadruplex. The SNV chr5:1,295,113 (G→T) located in the TERT region is present around a G quadruplex in the non-template strand. The SNV was associated with more than twenty-two cancer types including central nervous system, liver, bladder, ovarian, breast, kidney lung, bone, pancreatic, among others. Many of these SNVs destabilize G quadruplexes. Further, with nine tetrads (GGG repeats present) multiple G4 can potentially be formed. With a SNV (G→A), We find the G quadruplex stability with the variant to differ if alternate G4 tetrads are used for the stacking.

4.5.4 Transcription factor binding

Transcription factor proteins (TFs) known to bind G rich regions including SP1 (200), NF-κB (201), CREB (202), and the methyl-CpG binding domain MBD of methyl-CpG binding

protein 2 (MeCP2) (203) had decreased association constants up to 10 fold for transcription factor sites with change of guanine to 8-oxoguanine in model duplex DNA with the donor acceptor pattern change on the imidazole ring in guanine compared to OG. The structure change for guanine for association with CREB was found have a role in epigenetic repression (202). This is supported with our results highlighting reversal of these sequences to a stabilized G4 by change through T→G region in cancer cells. For instance, the variant chr10:122,143,482: G→A significantly affects the binding sites of TFs NHLH1, FOXO3, TAL1, TP53, HES5, HES7, USF2, EGR3, ZNF740, and SP1 among others (Appendix Table B 25). We observe similar observations for an additional 424 SNVs which occur in at least five cancer types in the COSMIC database and disrupt the TF binding site with an average of 15.1 TF per variant (Appendix Table B 27).

Local network cluster (STRING) analysis of the enriched TFs yielded terms related to PRC1 complex (4/12 FDR 0.00049) and PcG protein complex (6/25, FDR 6.22e-06), PcG protein complex, and positive regulation of histone H3-K27 methylation (11/59 FDR 1.82 e-10). Polycomb repressive Complex (PRC1) engage in transcriptional control through chromatin modification with histone 2A through a protein ligase Ubiquitylation (204, 205). Although the mechanism of PRC1 is under active investigation, recent evidence suggests role of G tracts to selectively remove PCR2 complex from genes during gene activation (206). Polycomb complexes have been associated with repression to maintain cell identity but are associated with actively transcribed loci, and this evidence suggest direct role of G quadruplexes across cell types to regulate expression through structural variation. GO cellular component analysis for the TFs found enriched terms related to Brahma complex,

(3/3 FDR 0.00079), Ino80 complex (5/15 FDR 5.35e-05) which are different complexes associated with chromatin remodelling.

Different repair mechanisms including BER, and mismatch repair are required for protecting non-canonical or mismatch base pairs due to polymerase error. Neurogenerative disorders occurring through expansion of CAG→CTG repeats have been associated with MutSβ, a heterodimer involved in mismatch repair. Though the involvement of G quadruplexes in gene transcription and telomere regulation has been studied and proven, the mechanism of base excision repair by DNA glycosylases in G quadruplex and other non-canonical structures is poorly understood. We identified G quadruplexes with SNVs in the genes of CHRNG, GRIN2C, CHAT, ADCY1, GABRG3, CACNG3, PPFIA3, LRTOMT, VAMP2, TSPOAP1, MAPK3, GABRR2, KCNJ6, PICK1, and STX1A, among others. These genes have been associated with several psychiatric disorders, schizophrenia, Bipolar Disorder, Tobacco Use Disorder, Parkinson's disease, and autism.

Previous research has shown the presence of G quadruplex sequences in various untranslated dendritic mRNAs suggesting the role of G quadruplexes as a neurite localization signal. Deletion of different putative G quadruplex sequence led to severe loss of signal in neurites. It has been hypothesized that the G quadruplex structure being sensitive to cationic, can function in correlation to the neuronal activity in localization and transport as activity dependent changes. Cationic sensitivity could influence the stability and structure and regulate the binding of trans-acting factors (207).

4.6 CONCLUSION

G quadruplexes are formed because of an intricate balance between the folding energy by a nick in the DNA, methylated guanines, and guanines available for stacking. The balance between the hypomethylated and hypermethylated G rich regions near promoters (despite

cytosine deamination and cytosine methylation) results in the preserved regions of CpG islands are observed across mammalian genome (208). These previously identified regions as CpG islands can be the preserved G quadruplex regions. Further, methylated guanines CpG islands have been identified within the genes (209) and the methylation susceptibility constraints the G quadruplex formation. We hypothesize these methylation and oxidation patterns are one mechanism by which G quadruplexes can preserve their sequence conformation and the variants occurring in these regions alter the molecular functions downstream.

With the introduction of next generation techniques for identification of G quadruplexes, analysis of variants in these complex region and mechanism of formation of G-quadruplex in different cell types remains uncertain. Our study points out a subset of different genes and G quadruplexes sequences which are affected in cancer cells and consequences of secondary structure forming regions with a nucleotide level investigation. G quadruplex formed in genomic regions participate in gene regulatory pathways to alter gene expression and downstream pathways. Based on large accumulation of published studies, we identify the possible effects of these single nucleotide variants occurring on coding and non-coding regions on the stability of G quadruplexes.

Table 4-1 Count/Proportion of Effect of type of mutation on stability of G4 (COSMIC database)

Type of SNV	Effect of SNV on MFE	Freq	Percentage
Transition	Destabilized	8600	22.93
Transversion	Further stabilized	6603	17.60
Transition	no change	6552	17.46
Transversion	Destabilized	6419	17.11
Transversion	no change	5509	14.68
Transition	Further stabilized	3832	10.21

Table 4-2 Proportion of SNV by annotation.

SNV	3' UTR	5' UTR	CDS	CpG Islands	Enhancers	EXON	Intergenic	INTRON	lncRNA GENCODE	PROMOTERS
G→A	34.82	31.23	39.34	27.58	19.01	35.18	27.84	26.74	28.27	26.87
G→T	18.3	14.82	15.84	12.01	12.08	16.6	17.15	14.56	14.89	13.64
T→G	12.67	16.48	8.59	16.78	29.84	11.74	18.38	19.91	18.15	18.95
C→T	11.13	8.24	12.19	12.23	5.68	11.35	7.94	9.06	10.46	8.89
A→G	7.62	10.83	6.91	9.56	12.61	8	11.38	11.49	9.89	10.98
G→C	5.16	6.05	4.69	6.27	7.28	5.09	6.89	6.3	6.23	6.51
C→G	3.66	4.85	4.47	7.69	6.75	4.33	3.47	4.59	4.56	6.34
C→A	2.76	2.99	4.11	3.85	1.42	3.65	1.95	2.6	2.73	3.3
T→C	2.15	1.66	1.79	2	2.13	1.96	1.67	2.1	2.13	2.07
T→A	0.7	0.86	0.84	0.67	0.89	0.81	1.27	1	1.17	0.82
A→T	0.59	1.2	0.72	0.75	1.6	0.76	1.12	0.94	1	0.85
A→C	0.45	0.8	0.51	0.62	0.71	0.52	0.94	0.72	0.53	0.77

CHAPTER 5 G4-SAMUHA

The identification of G quadruplex sequences in DNA is of great interest in bioinformatics and has become an important field of research. G quadruplexes are secondary structures formed by guanine-rich DNA sequences, which play important roles in various biological processes such as transcriptional regulation, replication, and telomere maintenance. Several tools have been developed to predict G quadruplex sequences from DNA sequences, but most of these tools are based on sequence patterns and lack specificity. In this chapter, we present a bioinformatics tool named "G-samuha" which is developed to identify putative G quadruplex sequences present in DNA sequences and compare against existing G quadruplex sequences to identify network of similar G4 sequences. This tool is developed using the R Shiny web browser and is based on Hidden Markov Model (HMM) profiles.

5.1 Methodology

G4-samuha is a user-friendly web-based tool that takes in input a DNA sequence or a G quadruplex sequence as a string or fasta file. The tool named "G4-samuha" is designed to identify putative G quadruplex sequences in DNA sequences using a combination of regular expression pattern and Hidden Markov Model (HMM) profiles. The tool takes an input DNA sequence or fasta file and searches it against a regular expression pattern to identify potential G quadruplex sequences. Then, each putative G quadruplex sequence is input into the tool, which uses HMM profiles, developed in an earlier chapter, to obtain log odds scores for each sequence against each profile. The HMM profiles are designed to identify the loop sequences and patterns that are present in G quadruplex structures. The tool then provides the nearest families of each input based on the loop sequence of the G quadruplex and pattern. The output is presented as a table that displays the predicted G

quadruplex sequences along with their loop sequences, patterns, and their corresponding HMM profiles. To develop the HMM profiles, a training set of G quadruplex sequences was used. The training set was obtained from multiple clustering tools (CD-HIT,DNACLUST, MESHCLUST, and a combination of Starcode and BLAST with hierarchical clustering), which were further trained based on the arrangement of Guanine tetrads and loops in each cluster. The sequences present in the database in experimentally annotated based on G4 identified from from GEO, accession GSE63874 (83).The sequences were aligned using DECIPHER package, and the loop sequences and patterns were extracted using custom scripts. The HMM profiles were then built using the aphid package in R.

5.2 Results and Discussion

The tool was also able to identify putative G quadruplex sequences similar to the training families in the input sequences. For G quadruplex sequences with short loops, the families of identified sequences could be redundant, and hence a confidence score comparing the sequence homology of the specific sequence against all families is provided which can support the uniqueness of each sequence to one or multiple families.

Based on the Hidden markov model generated using the multiple sequence alignment, users can upload fasta file or directly input DNA sequences to identify putative G quadruplex sequences based on sequence parsing of repeats of guanine interspersed by nucleotide loops. For all the potential G-quadruplex identified, the tool searches through all the model, and identifies the model with highest log odds score.

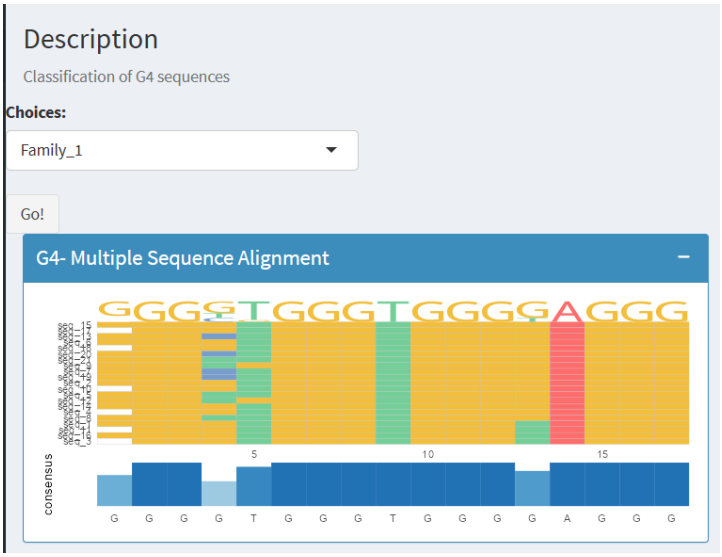


Figure 5-1 Screenshot of Multiple sequence alignment of Family 1 in the tool. User can search for specific families based on the training model

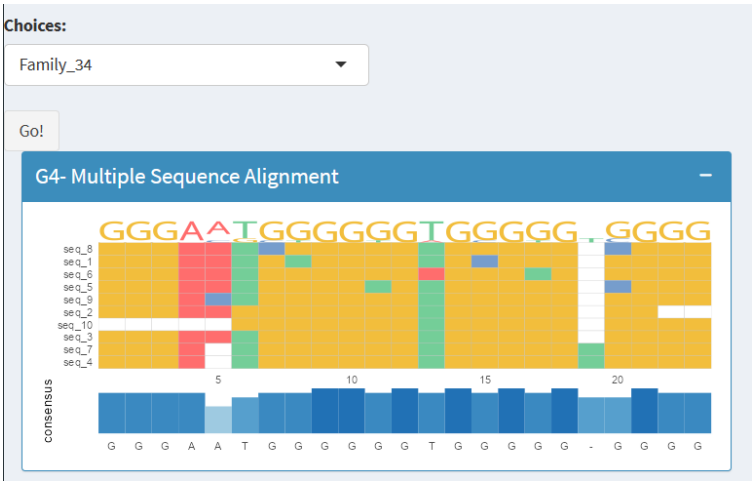


Figure 5-2 Screenshot of Multiple sequence alignment of Family 1 in the tool. User can search for specific families based on the training model

The HMM profiles used by G-samuhaare specific to G quadruplex loops and patterns, which increases the specificity of the tool. The use of HMM profiles also allow the tool to identify G quadruplexes with varying loop lengths and patterns. The tool is also user-friendly and can be used by researchers with limited bioinformatics experience.

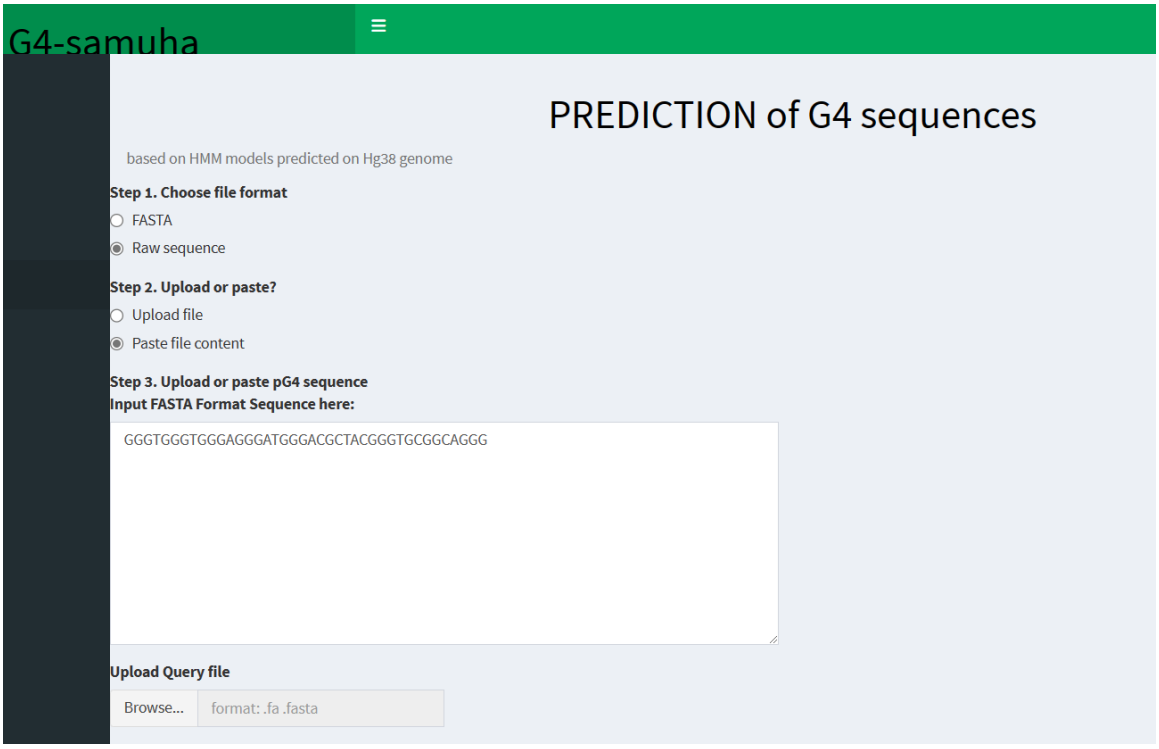


Figure 5-3 Screenshot using an example input of putative G-quadruplex repeat in G4 Samuha

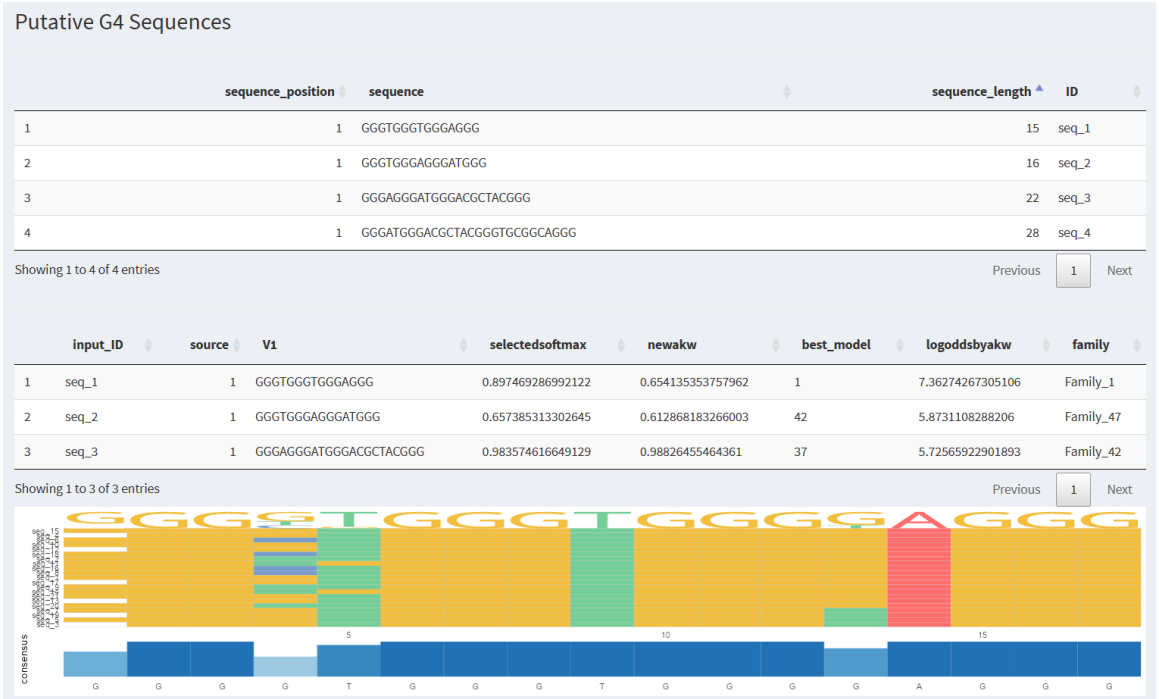
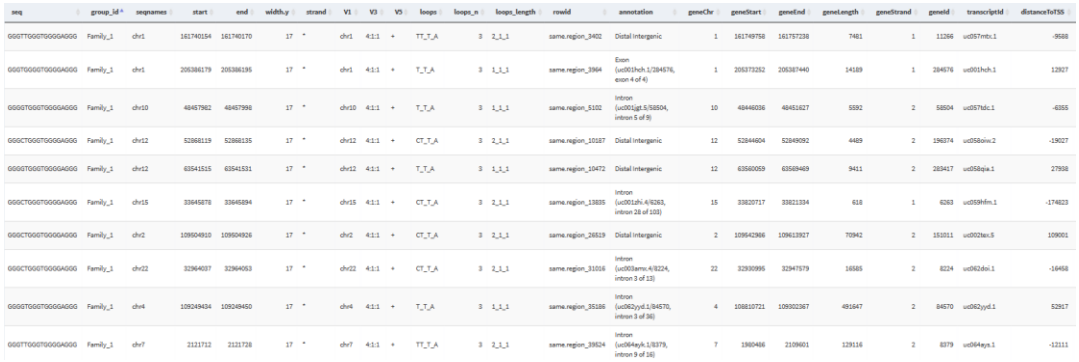


Figure 5-4 Screenshot showing results of putative G-quadruplex and log odds score for each sequence for a family in G4 Samuha



seq	group_id	seqnames	start	end	width.y	strand	V1	V3	V5	loops	loops_n	loops_length	rowId	annotation	geneChr	geneStart	geneEnd	geneLength	geneStrand	geneId	transcriptId	distanceToTSS
GGGTTGGGTGGGGAGGG	Family_1	chr1	161740154	161740170	17	*	chr1	4:1:1	+	TT_T_A	3	2_1_1	same.region_3402	Distal Intergenic	1	161749758	161757238	7481	1	11266	uc057mtc.1	-9588
GGGTGGGGTGGGGAGGG	Family_1	chr1	205386179	205386195	17	*	chr1	4:1:1	+	T_T_A	3	1_1_1	same.region_3964	Exon (uc001hch.1/284676, exon 4 of 4)	1	205373252	205387440	14189	1	284676	uc001hch.1	12927
GGGGTGGGTGGGGAGGG	Family_1	chr10	48457962	48457998	17	*	chr10	4:1:1	+	T_T_A	3	1_1_1	same.region_5102	Intron (uc001jgt.5/58504, intron 5 of 9)	10	48446036	48451627	5592	2	58504	uc057tdc.1	-6355
GGGCTGGGTGGGGAGGG	Family_1	chr12	52868119	52868135	17	*	chr12	4:1:1	+	CT_T_A	3	2_1_1	same.region_10187	Distal Intergenic	12	52844604	52849092	4489	2	196374	uc058oiw.2	-19027
GGGGTGGGTGGGGAGGG	Family_1	chr12	63541515	63541531	17	*	chr12	4:1:1	+	T_T_A	3	1_1_1	same.region_10472	Distal Intergenic	12	63560059	63568469	9411	2	283417	uc058qia.1	27938
GGGCTGGGTGGGGAGGG	Family_1	chr15	33645878	33645894	17	*	chr15	4:1:1	+	CT_T_A	3	2_1_1	same.region_13835	Intron (uc001zhi.4/6263, intron 28 of 103)	15	33820717	33821334	618	1	6263	uc059hfm.1	-174823
GGGCTGGGTGGGGAGGG	Family_1	chr2	109604910	109604926	17	*	chr2	4:1:1	+	CT_T_A	3	2_1_1	same.region_26819	Distal Intergenic	2	109642966	109613927	70942	2	151011	uc002tec.5	109001
GGGCTGGGTGGGGAGGG	Family_1	chr22	32964037	32964053	17	*	chr22	4:1:1	+	CT_T_A	3	2_1_1	same.region_31016	Intron (uc003amv.4/8224, intron 3 of 13)	22	32930995	32947579	16585	2	8224	uc062doi.1	-16458
GGGGTGGGTGGGGAGGG	Family_1	chr4	109249434	109249450	17	*	chr4	4:1:1	+	T_T_A	3	1_1_1	same.region_35186	Intron (uc062yyd.1/84570, intron 3 of 36)	4	108810721	109302367	491647	2	84570	uc062yyd.1	52917
GGGTTGGGTGGGGAGGG	Family_1	chr7	2121712	2121728	17	*	chr7	4:1:1	+	TT_T_A	3	2_1_1	same.region_39624	Intron (uc064eyk.1/8379, intron 9 of 16)	7	1980406	2109601	129116	2	8379	uc064eyk.1	-12111

Figure 5-5 Screenshot showing results of G4 Samuha for specific families identified. More than 50 fields relating to each families along with gene information, distance to Transcription Start Site (TSS) and thermodynamic profiles obtained from RNAfold for all training and predicted sequences is present

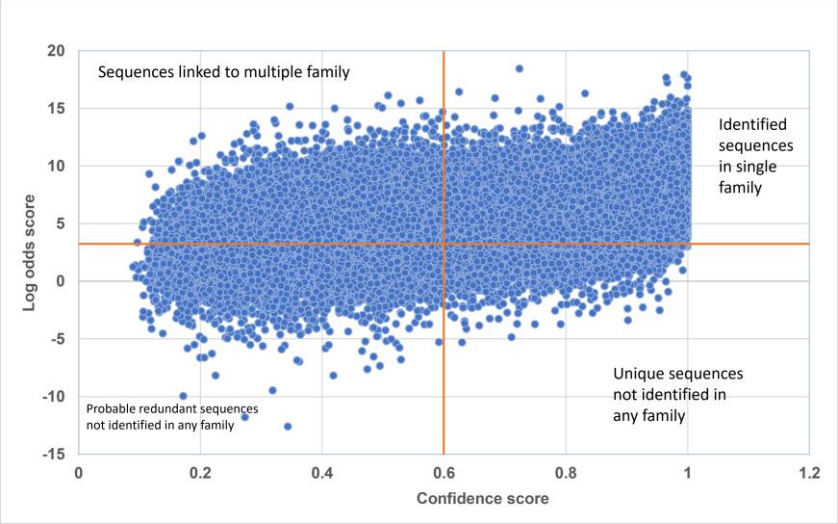


Figure 5-6 Illustration of pG4 seqeuences in human genome (hg38) with Log odds score and Confidence(Akaike) score The points above the abscissa represent the G4 sequences with families identified by the 95 families present in the model. The right side of the ordinate represent the unique sequences with the first quadrant highlighting the unique

sequences identified by the models provided through this method. The fourth quadrant represents the sequences unique but not identified by the models

5.3 Conclusion

In conclusion, we have developed a bioinformatics tool named "G-samuha" that can identify putative G quadruplex sequences present in DNA sequences. The tool is based on HMM profiles and is specific to G quadruplex loops and patterns. The tool is user-friendly and can be used by researchers with limited bioinformatics experience. The use of this tool will enable researchers to identify G quadruplex sequences in their DNA sequences, which can be used for further analysis and experimental validation.

CHAPTER 6 CONCLUSION AND FUTURE WORK

In conclusion, the project on G quadruplexes has successfully identified clusters of G quadruplexes with the aid of experimental evidence. This has led to the development of models of G quadruplexes, which have been incorporated into a R shiny tool called G-Samuha. This tool provides log odds scores for the nearest identified families along with annotations of existing families. Furthermore, an exciting avenue for future research involves the identification and analysis of G quadruplex structural variants and their impact on the structure. This research holds the potential to provide us with a deeper understanding of the regions of potential mutation across the G quadruplex region. In turn, such knowledge can inform the development of drugs and treatments that target G quadruplexes, which could have far-reaching implications for various fields. Structural evidence of binding with transcription factors and G-quadruplex complexes, such as minor groove width and electrostatic potential, can also be utilized to better understand the properties and functions of G quadruplexes. Moving forward, a promising avenue for further research is the application of the Viterbi algorithm to identify more G4 structures. This algorithm can be trained on structural parameters, such as minor groove width and electrostatic potential, to identify new G quadruplex structures with high accuracy. With continued advancements in our understanding of G quadruplexes, we can further explore their potential in various fields, including drug design and disease treatments. G quadruplexes are becoming increasingly recognized as important genomic and epigenetic elements, and their role in biological processes is being actively researched. The identification of G quadruplex structural variants and their impact on the structure has the potential to inform the development of drugs and treatments that target G quadruplexes, which could be used to treat a wide range

of diseases, including cancer and neurological disorders. Developing drugs that target G quadruplexes could potentially lead to the discovery of new therapies for cancer, as well as a range of other diseases. The unique properties of G quadruplexes make them ideal for building new nanostructures, which could be used in applications such as drug delivery and biosensors. My work on G quadruplex can support to improve the target of such drugs based on the identified families.

REFERENCES

1. Openstax, in "OpenStax Anatomy and Physiology". (2018).
2. R. E. Franklin, R. G. Gosling, Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. Nature 172, 156-157 (1953).
3. R. Lorenz et al., RNA Folding Algorithms with G-Quadruplexes. BSB 7409, 49-60 (2012).
4. J. L. Huppert, S. Balasubramanian, Prevalence of quadruplexes in the human genome. Nucleic acids research 33, 2908-2916 (2005).
5. A. Bedrat, L. Lacroix, J.-L. Mergny, Re-evaluation of G-quadruplex propensity with G4Hunter. Nucleic acids research 44, 1746-1759 (2016).
6. M. Yano, Y. Kato, Using hidden Markov models to investigate G-quadruplex motifs in genomic sequences. BMC Genomics 15, S15-S15 (2014).
7. O. Kikin, L. D'Antonio, P. S. Bagga, QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic Acids Research 34, W676-W682 (2006).
8. J. Eddy, N. Maizels, Gene function correlates with potential for G4 DNA formation in the human genome. Nucleic Acids Research 34, 3887-3896 (2006).
9. H. M. Wong, O. Stegle, S. Rodgers, J. L. Huppert, A toolbox for predicting g-quadruplex formation and stability. Journal of nucleic acids 2010, 564946-564946 (2010).
10. V. k. Yadav, J. K. Abraham, P. Mani, R. Kulshrestha, S. Chowdhury, QuadBase: Genome-wide database of G4 DNA - Occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. Nucleic Acids Research 36, 381-385 (2008).
11. J. Hon, T. Martínek, J. Zendulka, M. Lexa, pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. Bioinformatics (Oxford, England) 33, 3373-3379 (2017).
12. J.-M. Garant, J.-P. Perreault, M. S. Scott, Motif independent identification of potential RNA G-quadruplexes by G4RNA screener. Bioinformatics (Oxford, England) 33, 3532-3537 (2017).
13. A. B. Sahakyan et al., Machine learning model for sequence-driven DNA G-quadruplex formation. Scientific Reports 7, 1-11 (2017).
14. E. Belmonte-Reche, J. C. Morales, G4-iM Grinder: when size and frequency matter. G-Quadruplex, i-Motif and higher order structure search and analysis tool. NAR Genom Bioinform 2, lqz005 (2020).
15. H. B. Cagirici, H. Budak, T. Z. Sen, G4Boost: a machine learning-based tool for quadruplex identification and stability prediction. BMC Bioinformatics 23, 240 (2022).
16. A. K. Todd, S. Neidle, Mapping the sequences of potential guanine quadruplex motifs. Nucleic Acids Res 39, 4917-4927 (2011).

17. F. Wu et al., Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into G4 evolution. Communications Biology 4, 98-98 (2021).
18. R. Zhang, Y. Lin, C. T. Zhang, Greglist: a database listing potential G-quadruplex regulated genes. Nucleic Acids Res 36, D372-376 (2008).
19. S. K. Mishra, A. Tawani, A. Mishra, A. Kumar, G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins. Sci Rep 6, 38144 (2016).
20. Q. Li et al., G4LDB: a database for discovering and studying G-quadruplex ligands. Nucleic Acids Res 41, D1115-1123 (2013).
21. Y. H. Wang et al., G4LDB 2.2: a database for discovering and studying G-quadruplex and i-Motif ligands. Nucleic Acids Res 50, D150-D160 (2022).
22. K. A. Gan, S. Carrasco Pro, J. A. Sewell, J. I. Fuxman Bass, Identification of Single Nucleotide Non-coding Driver Mutations in Cancer. Front Genet 9, 16 (2018).
23. F. Chen, Y. Zhang, C. J. Creighton, Systematic identification of non-coding somatic single nucleotide variants associated with altered transcription and DNA methylation in adult and pediatric cancers. NAR Cancer 3, zcab001 (2021).
24. S. Abramov et al., Landscape of allele-specific transcription factor binding in the human genome. Nat Commun 12, 2751 (2021).
25. J. Zhao, D. Li, J. Seo, A. S. Allen, R. Gordan, Quantifying the Impact of Non-coding Variants on Transcription Factor-DNA Binding. Res Comput Mol Biol 10229, 336-352 (2017).
26. A. Kahles et al., Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell 34, 211-224 e216 (2018).
27. M. Zarrei, J. R. MacDonald, D. Merico, S. W. Scherer, A copy number variation map of the human genome. Nat Rev Genet 16, 172-183 (2015).
28. H. J. Abel et al., Mapping and characterization of structural variation in 17,795 human genomes. Nature 583, 83-89 (2020).
29. G. Federici, S. Soddu, Variants of uncertain significance in the era of high-throughput genome sequencing: a lesson from breast and ovary cancers. J Exp Clin Cancer Res 39, 46 (2020).
30. N. Sharma, G. R. Cutting, The genetics and genomics of cystic fibrosis. J Cyst Fibros 19 Suppl 1, S5-S9 (2020).
31. C. Souchay, M. Padula, M. Schneider, M. Debbane, S. Eliez, Developmental trajectories and brain correlates of directed forgetting in 22q11.2 deletion syndrome. Brain Res 1773, 147683 (2021).
32. J. Cortes-Martin et al., Deletion Syndrome 22q11.2: A Systematic Review. Children (Basel) 9, (2022).
33. L. Verges et al., An exploratory study of predisposing genetic factors for DiGeorge/velocardiofacial syndrome. Sci Rep 7, 40031 (2017).
34. M. D. Gallagher, A. S. Chen-Plotkin, The Post-GWAS Era: From Association to Function. Am J Hum Genet 102, 717-730 (2018).
35. E. Cano-Gamez, G. Trynka, From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Front Genet 11, 424 (2020).

36. S. Bamford et al., The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. British Journal of Cancer 91, 355-358 (2004).
37. M. J. Landrum et al., ClinVar: improving access to variant interpretations and supporting evidence. Nucleic acids research 46, D1062-D1067 (2018).
38. L. Phan et al., dbVar structural variant cluster set for data analysis and variant comparison. F1000Res 5, 673 (2016).
39. M. Pan et al., Novel LOVD databases for hereditary breast cancer and colorectal cancer genes in the Chinese population. Hum Mutat 32, 1335-1340 (2011).
40. K. Ganesan, A. Kulandaisamy, S. Binny Priya, M. M. Gromiha, HuVarBase: A human variant database with comprehensive information at gene and protein levels. PLoS One 14, e0210475 (2019).
41. H. Biggs, P. Parthasarathy, A. Gavryushkina, P. P. Gardner, ncVarDB: a manually curated database for pathogenic non-coding variants and benign controls. Database (Oxford) 2020, (2020).
42. L. Clarke et al., The 1000 Genomes Project: data management and community access. Nat Methods 9, 459-462 (2012).
43. T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahe, VSEARCH: a versatile open source tool for metagenomics. PeerJ 4, e2584 (2016).
44. J. D. Thompson, D. G. Higgins, T. J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 4673-4680 (1994).
45. E.-E. T. Online. (2020).
46. S. Lago et al., Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome. Nature communications 12, 1-13 (2021).
47. R. Hänsel-Hertsch et al., Landscape of G-quadruplex DNA structural regions in breast cancer. Nature genetics 52, 878-883 (2020).
48. G. Biffi, D. Tannahill, J. Miller, W. J. Howat, S. Balasubramanian, Elevated levels of G-quadruplex formation in human stomach and liver cancer tissues. PloS one 9, e102711 (2014).
49. G. Liu et al., RNA G-quadruplex regulates microRNA-26a biogenesis and function. Journal of Hepatology 73, 371-382 (2020).
50. E. Wang, R. Thombre, Y. Shah, R. Latanich, J. Wang, G-Quadruplexes as pathogenic drivers in neurodegenerative disorders. Nucleic Acids Research 49, 4816-4830 (2021).
51. G. Biffi, D. Tannahill, J. McCafferty, S. Balasubramanian, Quantitative visualization of DNA G-quadruplex structures in human cells. Nature Chemistry 5, 182-186 (2013).
52. H. Fernando et al., Genome-wide analysis of a G-quadruplex-specific single-chain antibody that regulates gene expression. Nucleic Acids Research 37, 6716-6722 (2009).
53. F. Kouzine et al., in G-Quadruplex Nucleic Acids. (Springer, 2019), pp. 369-382.
54. B. Ruttkay-Nedecky et al., G-quadruplexes as sensing probes. Molecules 18, 14760-14779 (2013).

55. A. K. Todd, S. Neidle, The relationship of potential G-quadruplex sequences in cis-upstream regions of the human genome to SP1-binding elements. Nucleic acids research 36, 2700-2704 (2008).
56. J. H. Chariker, D. M. Miller, E. C. Rouchka, Computational analysis of G-quadruplex forming sequences across chromosomes reveals high density patterns near the terminal ends. PloS one 11, e0165101 (2016).
57. R. Hänsel-Hertsch et al., G-quadruplex structures mark human regulatory chromatin. Nature genetics 48, 1267-1272 (2016).
58. A. Risitano, K. R. Fox, Influence of loop size on the stability of intramolecular DNA quadruplexes. Nucleic Acids Research 32, 2598-2606 (2004).
59. G. Sattin et al., Conformation and stability of intramolecular telomeric G-quadruplexes: sequence effects in the loops. PLoS One 8, e84113 (2013).
60. R. Tippana, W. Xiao, S. Myong, G-quadruplex conformation and dynamics are determined by loop length and sequence. Nucleic acids research 42, 8106-8114 (2014).
61. A. Guédin, A. De Cian, J. Gros, L. Lacroix, J.-L. Mergny, Sequence effects in single-base loops for quadruplexes. Biochimie 90, 686-696 (2008).
62. Y. Y. Li, D. N. Dubins, D. M. N. T. Le, K. Leung, R. B. Macgregor Jr, The role of loops and cation on the volume of unfolding of G-quadruplexes related to HTel. Biophysical Chemistry 231, 55-63 (2017).
63. Y. Y. Li, R. B. Macgregor Jr, A thermodynamic study of adenine and thymine substitutions in the loops of the oligodeoxyribonucleotide HTel. The Journal of Physical Chemistry B 120, 8830-8836 (2016).
64. A. Piazza et al., Short loop length and high thermal stability determine genomic instability induced by G-quadruplex-forming minisatellites. The EMBO journal 34, 1718-1734 (2015).
65. P. A. Rachwal, T. Brown, K. R. Fox, Sequence effects of single base loops in intramolecular quadruplex DNA. FEBS letters 581, 1657-1660 (2007).
66. P. Hazel, J. Huppert, S. Balasubramanian, S. Neidle, Loop-Length-Dependent Folding of G-Quadruplexes. Journal of the American Chemical Society 126, 16405-16415 (2004).
67. S. Lago, E. Tosoni, M. Nadai, M. Palumbo, S. N. Richter, The cellular protein nucleolin preferentially binds long-looped G-quadruplex nucleic acids. Biochimica et Biophysica Acta (BBA)-General Subjects 1861, 1371-1381 (2017).
68. K. Takahama, C. Sugimoto, S. Arai, R. Kurokawa, T. Oyoshi, Loop lengths of G-quadruplex structures affect the G-quadruplex DNA binding selectivity of the RGG motif in Ewing's sarcoma. Biochemistry 50, 5369-5378 (2011).
69. F. Bolduc, J.-M. Garant, F. Allard, J.-P. Perreault, Irregular G-quadruplexes found in the untranslated regions of human mRNAs influence translation. Journal of Biological Chemistry 291, 21751-21760 (2016).
70. K. W. Lim et al., Sequence variant (CTAGGG) n in the human telomere favors a G-quadruplex structure containing a G· C· G· C tetrad. Nucleic acids research 37, 6239-6248 (2009).
71. K. W. Lim et al., Structure of the human telomere in K+ solution: a stable basket-type G-quadruplex with only two G-tetrad layers. Journal of the American Chemical Society 131, 4301-4309 (2009).

72. V. T. Mukundan, A. T. Phan, Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. Journal of the American Chemical Society 135, 5017-5028 (2013).
73. J.-M. Garant, J.-P. Perreault, M. S. Scott, Motif independent identification of potential RNA G-quadruplexes by G4RNA screener. Bioinformatics 33, 3532-3537 (2017).
74. J. Hon, T. Martínek, J. Zendulka, M. Lexa, pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. Bioinformatics 33, 3373-3379 (2017).
75. O. Doluca, G4Catchall: A G-quadruplex prediction approach considering atypical features. Journal of Theoretical Biology 463, 92-98 (2019).
76. J.-M. Garant, M. J. Luce, M. S. Scott, J.-P. Perreault, G4RNA: an RNA G-quadruplex database. Database 2015, (2015).
77. A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neuböck, I. L. Hofacker, The vienna RNA websuite. Nucleic acids research 36, W70-W74 (2008).
78. X.-J. Lu, DSSR-enabled innovative schematics of 3D nucleic acid structures with PyMOL. Nucleic Acids Research 48, e74-e74 (2020).
79. T. Zok, M. Popenda, M. Szachniuk, ElTetrado: a tool for identification and classification of tetrads and quadruplexes. BMC bioinformatics 21, 1-7 (2020).
80. L. P. P. Patro, A. Kumar, N. Kolimi, T. Rathinavelan, 3D-NuS: a web server for automated modeling and visualization of non-canonical 3-dimensional nucleic acid structures. Journal of molecular biology 429, 2438-2448 (2017).
81. J. A. Capra, K. Paeschke, M. Singh, V. A. Zakian, G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae. PLoS computational biology 6, e1000861 (2010).
82. F. Wu et al., Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into G4 evolution. Communications biology 4, 1-11 (2021).
83. V. S. Chambers et al., High-throughput sequencing of DNA G-quadruplex structures in the human genome. Nature biotechnology 33, 877-881 (2015).
84. G. Marsico et al., Whole genome experimental maps of DNA G-quadruplexes in multiple species. Nucleic acids research 47, 3862-3874 (2019).
85. T. Seviour et al., The biofilm matrix scaffold of Pseudomonas aeruginosa contains G-quadruplex extracellular DNA structures. npj Biofilms and Microbiomes 7, 1-12 (2021).
86. X. Shao et al., RNA G-Quadruplex Structures Mediate Gene Regulation in Bacteria. mBio 11, e02926-02919 (2020).
87. K.-w. Zheng et al., Detection of genomic G-quadruplexes in living cells using a small artificial protein. Nucleic Acids Research 48, 11706-11720 (2020).
88. S. Völkel et al., Zinc finger independent genome-wide binding of Sp2 potentiates recruitment of histone-fold protein Nf-y distinguishing it from Sp1 and Sp3. PLoS genetics 11, e1005102 (2015).
89. E.-A. Raiber, R. Kranaster, E. Lam, M. Nikan, S. Balasubramanian, A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. Nucleic acids research 40, 1499-1508 (2012).

90. S. Da Ros et al., G-Quadruplex modulation of SP1 functional binding sites at the KIT proximal promoter. International journal of molecular sciences 22, 329 (2020).
91. F. Rezzoug, S. D. Thomas, E. C. Rouchka, D. M. Miller, Discovery of a family of genomic sequences which interact specifically with the c-MYC promoter to regulate c-MYC expression. PloS one 11, e0161588 (2016).
92. A. P. David et al., G-quadruplexes as novel cis-elements controlling transcription during embryonic development. Nucleic acids research 44, 4163-4173 (2016).
93. J.-D. Beaudoin, J.-P. Perreault, 5′-UTR G-quadruplex structures acting as translational repressors. Nucleic acids research 38, 7022-7036 (2010).
94. T. A. Brooks, L. H. Hurley, Targeting MYC expression through G-quadruplexes. Genes & cancer 1, 641-649 (2010).
95. A. M. Fleming, J. Zhou, S. S. Wallace, C. J. Burrows, A role for the fifth G-track in G-quadruplex forming oncogene promoter sequences during oxidative stress: Do these "spare tires" have an evolved function? ACS central science 1, 226-233 (2015).
96. S. Cogoi, L. E. Xodo, G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. Nucleic acids research 34, 2536-2549 (2006).
97. P. Agrawal, C. Lin, R. I. Mathad, M. Carver, D. Yang, The major G-quadruplex formed in the human BCL-2 proximal promoter adopts a parallel structure with a 13-nt loop in K+ solution. Journal of the American Chemical Society 136, 1750-1753 (2014).
98. P. J. Bates, D. A. Laber, D. M. Miller, S. D. Thomas, J. O. Trent, Discovery and development of the G-rich oligonucleotide AS1411 as a novel treatment for cancer. Experimental and molecular pathology 86, 151-164 (2009).
99. J. Spiegel et al., G-quadruplexes are transcription factor binding hubs in human chromatin. Genome biology 22, 1-15 (2021).
100. J. Jana, Y. M. Vianney, N. Schröder, K. Weisz, Guiding the folding of G-quadruplexes through loop residue interactions. Nucleic Acids Research 50, 7161-7175 (2022).
101. A. Marchand, V. Gabelica, Folding and misfolding pathways of G-quadruplex DNA. Nucleic acids research, gkw970 (2016).
102. R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison, Biological sequence analysis: probabilistic models of proteins and nucleic acids. (Cambridge university press, 1998).
103. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658-1659 (2006).
104. B. T. James, B. B. Luczak, H. Z. Girgis, MeShClust: an intelligent tool for clustering DNA sequences. Nucleic acids research 46, e83-e83 (2018).
105. M. Ghodsi, B. Liu, M. Pop, DNACLUST: accurate and efficient clustering of phylogenetic marker genes. BMC bioinformatics 12, 1-11 (2011).
106. E. Zorita, P. Cusco, G. J. Filion, Starcode: sequence clustering based on all-pairs search. Bioinformatics 31, 1913-1919 (2015).
107. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. Journal of molecular biology 215, 403-410 (1990).

108. E. S. Wright, DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. BMC bioinformatics 16, 1-14 (2015).
109. G. Collet. (2017).
110. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. Nucleic acids research 39, W29-W37 (2011).
111. S. P. Wilkinson, aphid: an R package for analysis with profile hidden Markov models. Bioinformatics 35, 3829-3830 (2019).
112. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842 (2010).
113. L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77, 257-286 (1989).
114. J. B. Johnson, K. S. Omland, Model selection in ecology and evolution. Trends in ecology & evolution 19, 101-108 (2004).
115. E.-J. Wagenmakers, S. Farrell, AIC model selection using Akaike weights. Psychonomic bulletin & review 11, 192-196 (2004).
116. H. Pages et al., Package 'Biostrings'. Bioconductor, 18129 (2013).
117. P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20, 53-65 (1987).
118. M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs, NbClust: an R package for determining the relevant number of clusters in a data set. Journal of statistical software 61, 1-36 (2014).
119. B. Honig, A. Nicholls, Classical Electrostatics in Biology and Chemistry. Science 268, 1144-1149 (1995).
120. R. Rohs et al., The role of DNA shape in protein–DNA recognition. Nature 461, 1248-1253 (2009).
121. M. El Hassan, C. Calladine, Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. Journal of molecular biology 259, 95-103 (1996).
122. T.-P. Chiu et al., DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics 32, 1211-1213 (2016).
123. H. Fan et al., BAHCC1 binds H3K27me3 via a conserved BAH module to mediate gene silencing and oncogenesis. Nature genetics 52, 1384-1396 (2020).
124. Y. Guo, S. Zhao, G. G. Wang, Polycomb gene silencing mechanisms: PRC2 chromatin targeting, H3K27me3'Readout', and phase separation-based compaction. Trends in Genetics 37, 547-565 (2021).
125. K. Banerjee et al., Regulation of tyrosine hydroxylase transcription by hnRNP K and DNA secondary structure. Nature communications 5, 1-13 (2014).
126. M. M. Farhath et al., G-Quadruplex-enabling sequence within the human tyrosine hydroxylase promoter differentially regulates transcription. Biochemistry 54, 5533-5545 (2015).
127. B. J. Janssen et al., Structural basis of semaphorin–plexin signalling. Nature 467, 1118-1122 (2010).
128. H. Takamatsu, A. Kumanogoh, Diverse roles for semaphorin− plexin signaling in the immune system. Trends in immunology 33, 127-135 (2012).

129. V. Kuryavyi, L. A. Cahoon, H. S. Seifert, D. J. Patel, RecA-binding pilE G4 sequence essential for pilin antigenic variation forms monomeric and 5′ end-stacked dimeric parallel G-quadruplexes. Structure 20, 2090-2102 (2012).
130. V. González, L. H. Hurley, The c-MYC NHE III1: function and regulation. Annual review of pharmacology and toxicology 50, 111-129 (2010).
131. L. H. Hurley, D. D. Von Hoff, A. Siddiqui-Jain, D. Yang, in Seminars in oncology. (Elsevier, 2006), vol. 33, pp. 498-512.
132. A. Siddiqui-Jain, C. L. Grand, D. J. Bearss, L. H. Hurley, Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. Proceedings of the National Academy of Sciences 99, 11593-11598 (2002).
133. D. Yang, L. H. Hurley, Structure of the biologically relevant G-quadruplex in the c-MYC promoter. Nucleosides, Nucleotides, and Nucleic Acids 25, 951-968 (2006).
134. A. Y. Zhang, A. Bugaut, S. Balasubramanian, A sequence-independent analysis of the loop length dependence of intramolecular RNA G-quadruplex stability and topology. Biochemistry 50, 7251-7258 (2011).
135. J. Li et al., Effects of length and loop composition on structural diversity and similarity of (G3TG3NmG3TG3) G-quadruplexes. Molecules 25, 1779 (2020).
136. E. Postel, S. Berberich, S. Flint, C. Ferrone, Human c-myc transcription factor PuF identified as nm23-H2 nucleoside diphosphate kinase, a candidate suppressor of tumor metastasis. Science 261, 478-480 (1993).
137. C. Shan et al., Chemical intervention of the NM23-H2 transcriptional programme on c-MYC via a novel small molecule. Nucleic acids research 43, 6677-6691 (2015).
138. V. González, L. H. Hurley, The C-terminus of nucleolin promotes the formation of the c-MYC G-quadruplex and inhibits c-MYC promoter activity. Biochemistry 49, 9706-9714 (2010).
139. M. J. Bywater et al., Inhibition of RNA polymerase I as a therapeutic strategy to promote cancer-specific activation of p53. Cancer cell 22, 51-65 (2012).
140. H. Xu et al., CX-5461 is a DNA G-quadruplex stabilizer with selective lethality in BRCA1/2 deficient tumours. Nature communications 8, 1-18 (2017).
141. C. Leonetti et al., G-quadruplex ligand RHPS4 potentiates the antitumor activity of camptothecins in preclinical models of solid tumors. Clinical Cancer Research 14, 7284-7291 (2008).
142. A. Local et al., APTO-253 Stabilizes G-quadruplex DNA, Inhibits MYC Expression, and Induces DNA Damage in Acute Myeloid Leukemia CellsAPTO-253 as a MYC Inhibitor and G4 Ligand for AML. Molecular cancer therapeutics 17, 1177-1186 (2018).
143. S. G. Zidanloo, A. Hosseinzadeh Colagar, H. Ayatollahi, J.-B. Raoof, Downregulation of the WT1 gene expression via TMPyP4 stabilization of promoter G-quadruplexes in leukemia cells. Tumor Biology 37, 9967-9977 (2016).
144. T. Tauchi et al., Activity of a novel G-quadruplex-interactive telomerase inhibitor, telomestatin (SOT-095), against human leukemia cells: involvement of ATM-dependent DNA damage response pathways. Oncogene 22, 5338-5347 (2003).

145. J. Liu et al., Inhibition of myc promoter and telomerase activity and induction of delayed apoptosis by SYUIQ-5, a novel G-quadruplex interactive agent in leukemia cells. Leukemia 21, 1300-1302 (2007).
146. D. Sen, W. Gilbert, Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. nature 334, 364-366 (1988).
147. D. Sen, W. Gilbert, A sodium-potassium switch in the formation of four-stranded G4-DNA. Nature 344, 410-414 (1990).
148. A. Bugaut, S. Balasubramanian, A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes. Biochemistry 47, 689-697 (2008).
149. K. B. Sutyak, P. Y. Zavalij, M. L. Robinson, J. T. Davis, Controlling molecularity and stability of hydrogen bonded G-quadruplexes by modulating the structure's periphery. Chemical Communications 52, 11112-11115 (2016).
150. I. Cheung, M. Schertzer, A. Rose, P. M. Lansdorp, Disruption of dog-1 in Caenorhabditis elegans triggers deletions upstream of guanine-rich DNA. Nature genetics 31, 405-409 (2002).
151. D. Dahan et al., Pif1 is essential for efficient replisome progression through lagging strand G-quadruplex DNA secondary structures. Nucleic acids research 46, 11847-11857 (2018).
152. K. Paeschke, J. A. Capra, V. A. Zakian, DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. Cell 145, 678-691 (2011).
153. R. Rodriguez et al., Small-molecule–induced DNA damage identifies alternative DNA structures in human genes. Nature chemical biology 8, 301-310 (2012).
154. T. B. London et al., FANCJ is a structure-specific DNA helicase associated with the maintenance of genomic G/C tracts. Journal of Biological Chemistry 283, 36132-36139 (2008).
155. C. Ribeyre et al., The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. PLoS genetics 5, e1000475 (2009).
156. A. De Magis et al., DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells. Proceedings of the National Academy of Sciences 116, 816-825 (2019).
157. A. Madireddy et al., G-quadruplex-interacting compounds alter latent DNA replication and episomal persistence of KSHV. Nucleic acids research 44, 3675-3694 (2016).
158. J. Lee et al., Dynamic interaction of BRCA2 with telomeric G-quadruplexes underlies telomere replication homeostasis. (2021).
159. Y. Mei et al., TERRA G-quadruplex RNA interaction with TRF2 GAR domain is required for telomere integrity. Scientific reports 11, 1-14 (2021).
160. J. Zimmer et al., Targeting BRCA1 and BRCA2 deficiencies with G-quadruplex-interacting compounds. Molecular cell 61, 449-460 (2016).
161. J. Gros et al., Guanines are a quartet's best friend: impact of base substitutions on the kinetics and stability of tetramolecular quadruplexes. Nucleic acids research 35, 3064-3075 (2007).

162. M.-C. Didiot et al., The G-quartet containing FMRP binding site in FMR1 mRNA is a potent exonic splicing enhancer. Nucleic acids research 36, 4902-4912 (2008).
163. S. Chaudhary, M. Kaushik, S. Ahmed, R. Kukreti, S. Kukreti, Structural Switch from Hairpin to Duplex/Antiparallel G-Quadruplex at Single-Nucleotide Polymorphism (SNP) Site of Human Apolipoprotein E (APOE) Gene Coding Region. ACS omega 3, 3173-3182 (2018).
164. S. K. Bharti et al., Specialization among iron-sulfur cluster helicases to resolve G-quadruplex DNA structures that threaten genomic stability. Journal of Biological Chemistry 288, 28217-28229 (2013).
165. A. Baral et al., Quadruplex-single nucleotide polymorphisms (Quad-SNP) influence gene expression difference among individuals. Nucleic acids research 40, 3800-3811 (2012).
166. A. A. Kuznetsova, O. S. Fedorova, N. A. Kuznetsov, Lesion recognition and cleavage of damage-containing quadruplexes and bulged structures by DNA glycosylases. Frontiers in Cell and Developmental Biology 8, 595687 (2020).
167. S. Bamford et al., The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. British journal of cancer 91, 355-358 (2004).
168. M. J. Landrum et al., ClinVar: improving access to variant interpretations and supporting evidence. Nucleic acids research 46, D1062-D1067 (2018).
169. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987-2993 (2011).
170. R. G. Cavalcante, M. A. Sartor, Annotatr: genomic regions in context. Bioinformatics 33, 2381-2383 (2017).
171. F. Hammal, P. de Langen, A. Bergon, F. Lopez, B. Ballester, ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. Nucleic Acids Research 50, D316-D325 (2022).
172. D. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4, 44-57 (2009).
173. D. Szklarczyk et al., The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research 49, D605-D612 (2021).
174. S. G. Coetzee, G. A. Coetzee, D. J. Hazelett, motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics 31, 3847-3849 (2015).
175. J. Sved, A. Bird, The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proceedings of the National Academy of Sciences 87, 4692-4696 (1990).
176. J. Youk, Y. An, S. Park, J.-K. Lee, Y. S. Ju, The genome-wide landscape of C: G> T: A polymorphism at the CpG contexts in the human population. BMC genomics 21, 1-11 (2020).
177. B. P. Belotserkovskii, J. H. Soo Shin, P. C. Hanawalt, Strong transcription blockage mediated by R-loop formation within a G-rich homopurine–

homopyrimidine sequence localized in the vicinity of the promoter. Nucleic acids research 45, 6589-6599 (2017).
178. P. Polak, P. F. Arndt, Transcription induces strand-specific mutations at the 5′ end of human genes. Genome Research 18, 1216-1223 (2008).
179. J. W. Park et al., 8-OxoG in GC-rich Sp1 binding sites enhances gene transcription in adipose tissue of juvenile mice. Scientific reports 9, 1-12 (2019).
180. J. W. Cave, D. E. Willis, G-quadruplex regulation of neural gene expression. The FEBS Journal 289, 3284-3303 (2022).
181. C. J. Westmark, J. S. Malter, FMRP mediates mGluR5-dependent translation of amyloid precursor protein. PLoS biology 5, e52 (2007).
182. J. Fürst et al., ICln159 Folds into a Pleckstrin Homology Domain-like Structure: INTERACTION WITH KINASES AND THE SPLICING FACTOR LSm4*[boxs]. Journal of Biological Chemistry 280, 31276-31282 (2005).
183. V. Gervais et al., TFIIH contains a PH domain involved in DNA nucleotide excision repair. Nature Structural & Molecular Biology 11, 616-622 (2004).
184. K. Das, M. Srivastava, S. C. Raghavan, GNG motifs can replace a GGG stretch during G-quadruplex formation in a context dependent manner. PLoS One 11, e0158794 (2016).
185. D. S. Lee, L. R. Ghanem, Y. Barash, Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. Nature communications 11, 1-12 (2020).
186. A. M. Fleming, C. J. Burrows, G-quadruplex folds of the human telomere sequence alter the site reactivity and reaction pathway of guanine oxidation compared to duplex DNA. Chemical research in toxicology 26, 593-607 (2013).
187. M. Pitié, C. Boldron, G. Pratviel, in Advances in inorganic chemistry. (Elsevier, 2006), vol. 58, pp. 77-130.
188. A. M. Fleming, J. Zhu, Y. Ding, C. J. Burrows, 8-Oxo-7, 8-dihydroguanine in the context of a gene promoter G-quadruplex is an on–off switch for transcription. ACS chemical biology 12, 2417-2426 (2017).
189. D. Sun et al., The proximal promoter region of the human vascular endothelial growth factor gene has a G-quadruplex structure that can be targeted by G-quadruplex–interactive agents. Molecular cancer therapeutics 7, 880-889 (2008).
190. I. Liguori et al., Oxidative stress, aging, and diseases. Clinical interventions in aging 13, 757 (2018).
191. S. Bielskutė, J. Plavec, P. Podbevšek, Impact of oxidative lesions on the human telomeric G-quadruplex. Journal of the American Chemical Society 141, 2594-2603 (2019).
192. D. M. Banda, N. N. Nuñez, M. A. Burnside, K. M. Bradshaw, S. S. David, Repair of 8-oxoG: A mismatches by the MUTYH glycosylase: Mechanism, metals and medicine. Free Radical Biology and Medicine 107, 202-215 (2017).
193. B. van Loon, U. Hübscher, An 8-oxo-guanine repair pathway coordinated by MUTYH glycosylase and DNA polymerase λ. Proceedings of the National Academy of Sciences 106, 18201-18206 (2009).
194. A. Viel et al., A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. EBioMedicine 20, 39-49 (2017).

195. S. Bellon, N. Shikazono, S. Cunniffe, M. Lomax, P. O'Neill, Processing of thymine glycol in a clustered DNA damage site: mutagenic or cytotoxic. Nucleic acids research 37, 4430-4440 (2009).
196. J. Zhou, A. M. Fleming, A. M. Averill, C. J. Burrows, S. S. Wallace, The NEIL glycosylases remove oxidized guanine lesions from telomeric and promoter quadruplex DNA structures. Nucleic acids research 43, 4039-4054 (2015).
197. M. Adrian, B. Heddi, A. T. Phan, NMR spectroscopy of G-quadruplexes. Methods 57, 11-24 (2012).
198. R. C. Monsen et al., The hTERT core promoter forms three parallel G-quadruplexes. Nucleic acids research 48, 5720-5734 (2020).
199. S. L. Palumbo, S. W. Ebbinghaus, L. H. Hurley, Formation of a unique end-to-end stacked pair of G-quadruplexes in the hTERT core promoter with implications for inhibition of telomerase by G-quadruplex-interactive ligands. Journal of the American Chemical Society 131, 10878-10891 (2009).
200. O. Ramon et al., Effects of 8-oxo-7, 8-dihydro-2′-deoxyguanosine on the binding of the transcription factor Sp1 to its cognate target DNA sequence (GC box). Free radical research 31, 217-229 (1999).
201. M. K. Hailer-Morrison, J. M. Kotler, B. D. Martin, K. D. Sugden, Oxidized guanine lesions as modulators of gene transcription. Altered p50 binding affinity and repair shielding by 7, 8-dihydro-8-oxo-2 '-deoxyguanosine lesions in the NF-κB promoter element. Biochemistry 42, 9761-9770 (2003).
202. S. P. Moore, K. J. Toomire, P. R. Strauss, DNA modifications repaired by base excision repair are epigenetic. DNA repair 12, 1152-1158 (2013).
203. V. Valinluck et al., Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). Nucleic acids research 32, 4100-4108 (2004).
204. N. Reynolds, A. O'Shaughnessy, B. Hendrich, Transcriptional repressors: multifaceted regulators of gene expression. Development 140, 505-512 (2013).
205. M. Vidal, K. Starowicz, Polycomb complexes PRC1 and their function in hematopoiesis. Experimental Hematology 48, 12-31 (2017).
206. M. Beltran et al., G-tract RNA removes Polycomb repressive complex 2 from genes. Nature structural & molecular biology 26, 899-909 (2019).
207. M. Subramanian et al., G–quadruplex RNA structure as a signal for neurite mRNA targeting. EMBO reports 12, 697-704 (2011).
208. R. S. Illingworth et al., Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS genetics 6, e1001134 (2010).
209. R. Illingworth et al., A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS biology 6, e22 (2008).

LIST OF APPENDIX TABLES

LIST OF APPENDIX FIGURES

Appendix A

Appendix Table A 1 Summary of Family 23.

Sequence	Location	Experimental Evidence	Gene ID	Anntoation
GGGTGGCGGGTGGGGGGAGGG	chr10:123316717-123316737	absent	9184	Intergenic
GGGTGAGGGTGCGGGTGAGGG	chr10:124610344-124610364	present	64077	Promoter
GGGTGAGGGTGCGGGTGAGGG	chr10:124610377-124610397	present	64077	Promoter
GGGTGAGGGTGCGGGTGAGGG	chr10:124610509-124610529	present	64077	Promoter
GGGTGAGGGTGCGGGTGAGGG	chr10:124610542-124610562	present	64077	Promoter
GGGTGAGGGTGCGGGTGAGGG	chr10:124610647-124610667	present	64077	Promoter
GGGTGAGGGTGCGGGTGAGGG	chr10:124610752-124610772	absent	64077	Promoter
GGGTGAGGGTGCGGGTGAGGG	chr10:124610785-124610805	absent	64077	Promoter
GGGTGAGGGTGCGGGTGAGGG	chr10:124610956-124610976	absent	64077	Promoter
GGGTGAGGGTGAGGGTGAGGG	chr10:124610410-124610430	present	64077	Promoter
GGGTGAGGGTGAGGGTGAGGG	chr10:124610857-124610877	absent	64077	Promoter
GGGTGAGGGTGAGGGTGAGGG	chr10:124610890-124610910	absent	64077	Promoter
GGGTGGAGGGGTGGGGTGGGG	chr10:132548209-132548229	present	3632	Intron
GGGTGGGGGGTGGGGAGAGGG	chr10:15725418-15725438	present	8516	Intergenic
GGGTGGGGGGTGGGGAGGGG	chr10:21717394-21717413	present	8028	Intron
GGGTGGGGGGTGGGGAGGGG	chr18:11976921-11976940	absent	3613	Promoter
GGGTTGGGGGTAGGGTGGGGG	chr10:62544042-62544062	present	22891	Intron
GGGGTGTGGGGCAGGGATGGGG	chr10:70678644-70678665	present	140766	Intron
GGGTGGGGGCTGGGGAGAGGG	chr10:78293711-78293731	absent	414243	Intron
GGGTCGGGGGCGGGGGAGGG	chr11:101129211-101129230	present	101054525	Promoter
GGGTGGGAGTGGGATGAGGG	chr11:125159894-125159913	present	103695364	Promoter
GGGTTGGGGAGTGGGGTTGGG	chr11:43926065-43926085	present	100507300	Intron
GGGTTGGGGGTGGGGTGGGG	chr11:44101102-44101121	present	2132	Promoter
GGGATGTGGGAAGGGATGGGG	chr11:69269215-69269235	present	26579	Intergenic
GGGGTGGGTGTGGGGTGGGG	chr12:113876018-113876037	present	9904	Intron
GGGGTGGGTGTGGGGTGGGG	chr16:46909101-46909120	present	84706	Intron
GGGATGGGGGTCGGGTGGGG	chr12:131917175-131917194	present	8408	Promoter
GGGGTGGGGGTGGGAGAGGG	chr12:2537890-2537909	present	775	Intron
GGGGTGAGGGTAGGGATGGGG	chr12:6640013-6640033	absent	84519	Promoter
GGGGTTGGGGAAGGGAGGGGG	chr13:112058816-112058836	present	6656	Intergenic
GGGGTGGGGAAGGGATTGGGG	chr13:53043320-53043340	absent	10562	Intron
GGGTGGGGGTGGGGGGCAGGG	chr14:100443444-100443464	present	79446	Intron
GGGTGGGGGTGGGGCAAGGG	chr14:103532880-103532899	present	115708	Promoter
GGGGTGGGTGAAGGGATGGGGG	chr14:105596969-105596990	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105597009-105597030	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105597050-105597071	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105597091-105597112	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105597132-105597153	absent	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105717880-105717901	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105717921-105717942	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105717962-105717983	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718002-105718023	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718043-105718064	absent	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718084-105718105	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718125-105718146	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718125-105718146	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718166-105718187	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718206-105718227	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718247-105718268	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718288-105718309	absent	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718329-105718350	absent	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718370-105718391	absent	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718411-105718432	absent	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718452-105718473	absent	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718493-105718514	absent	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718534-105718555	absent	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718575-105718596	absent	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718616-105718637	absent	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718657-105718678	absent	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718698-105718719	absent	102465871	Intergenic

GGGGTGGGTGAAGGGATGGGGG	chr14:105718739-105718760	absent	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718780-105718801	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718821-105718842	present	102465871	Intergenic
GGGGTGGGTGAAGGGATGGGGG	chr14:105718862-105718883	present	102465871	Intergenic
GGGGTGGGGGTGGGGTGGGGG	chr14:19037660-19037680	present	100508046	Intron
GGGGTGGGGGTGGGGTGGGGG	chr22:15760289-15760309	present	106146148	Intron
GGGTCAGGGGTGGGGTGGGG	chr14:50864482-50864501	present	145447	Intron
GGGTTGGGGGCGGGGGTGGG	chr14:65413539-65413558	present	2530	Promoter
GGGGTGGGGGTGGGGCGGGG	chr14:68908184-68908203	present	87	Intron
GGGGTGAGGGGAAGGGATGGG	chr14:72421211-72421231	absent	9628	Promoter
GGGTTGGGGGGGTGGGTGGGG	chr15:65356691-65356711	present	9543	Promoter
GGGTGGGGCTGGGGTGTGGG	chr15:88391894-88391913	absent	26589	Intergenic
GGGTTTGGGGTGGGGGTAGGG	chr15:89109459-89109479	absent	11057	Intron
GGGATTGGGGGTGGGGGAGGG	chr16:27805003-27805023	present	23247	Intron
GGGGTGGGGGCCGGGATGGGG	chr16:51096031-51096051	present	6299	Intergenic
GGGTGGGGGTGGGCTGAGGG	chr16:67172198-67172217	present	8996	Promoter
GGGAGGGGGTGGGGGGAGGG	chr16:7709746-7709765	present	54715	3' UTR
GGGGTGGGGGCAGGGTGGGG	chr16:79383220-79383239	present	51741	Intergenic
GGGTGGGGGTGGGAGTCAGGG	chr16:86705437-86705457	present	101928614	Intergenic
GGGAGTGGGGGTGGGGTGGGG	chr17:12982890-12982910	present	60528	Exon
GGGAGTGGGGGTGGGGTGGGG	chr9:76754262-76754282	present	50652	Intron
GGGTTAGGGGTGGGGTGGGGG	chr17:18127574-18127594	present	51168	Intron
GGGGTTGGGGGAGGGAGGGG	chr17:1880069-1880088	present	6117	Promoter
GGGCTGGGGGTGGGGAAGGG	chr17:50867047-50867066	present	400604	Promoter
GGGGTGGGGAGTGGGGTGGGG	chr17:65457424-65457444	present	105827617	Promoter
GGGGTGGGGGGAAGGGAGGGGG	chr17:82020660-82020681	present	201254	Promoter
GGGTTGGGGGAGGGATTGGG	chr18:24852934-24852953	absent	105372028	Intron
GGGGGTGGGGGTGGGGTGGGG	chr18:26503844-26503864	present	284252	Intron
GGGGTTGGGGGTGGGGGGGG	chr18:62610383-62610402	present	54877	Intergenic
GGGGTGGCGGGTGGGGTGGGG	chr18:77413747-77413767	present	2587	Intergenic
GGGTTGCGGGGTAGGGGAGGG	chr19:35851196-35851216	absent	4868	Promoter
GGGGTGGGGGAAGGGAGGGG	chr19:408850-408869	present	126567	Promoter
GGGTGGGGGTGGAGGGAGGG	chr19:5117318-5117337	present	23030	Promoter
GGGTTGGGGGTGGGCTGGGG	chr1:10670083-10670102	present	54897	Intron
GGGTGGGGCTGGGAGTGAGGG	chr1:1075343-1075363	present	401934	Promoter
GGGTTGGGACTGGGGTGGGG	chr1:157198319-157198338	present	2117	Intergenic
GGGGTGGGGATGGGATGGGG	chr1:202304270-202304289	present	59352	Intron
GGGGTGGGGGAAGGGGGTGGGG	chr1:211260741-211260762	present	55758	Promoter
GGGTGGAGGGTGGGGTGGGG	chr1:21865234-21865253	present	3339	Intron
GGGTGGAGGGTGGGGTGGGG	chr1:30271549-30271568	absent	101929406	Intergenic
GGGTGGGGATGGGAGTGAGGG	chr1:23911503-23911523	present	1269	Promoter
GGGGTGGGCGGTGGGGTGGGG	chr1:24584671-24584691	present	400746	Promoter
GGGTTGGGGGTGGGGTGGGGGG	chr1:36566733-36566754	present	1441	Intergenic
GGGTGGGGGTGGGGTGAGGG	chr1:36576371-36576390	present	1441	Intergenic
GGGCAGGGGTGGGGTGAGGG	chr1:44506457-44506476	present	100847089	Intron
GGGGTGGGCGAAGGGAGTGGGG	chr1:53947258-53947279	present	115353	Promoter
GGGGGGGAGTGGGGTGAGGG	chr1:87522994-87523013	absent	100505768	Intergenic
GGGGCTGGGGGTGGGGGAGGG	chr20:2820918-2820938	present	100288797	Promoter
GGGTTGCGGGGTGGGGTGGGGG	chr20:29750401-29750422	absent	245929	Intergenic
GGGTTGCGGGGTGGGGTGGGGG	chr20:30497208-30497229	absent	245929	Intergenic
GGGTTGGGGGAGGGGGTGGG	chr20:32980815-32980834	present	140732	Downstream
GGGGTGGAGGGTGGGGTGAGGG	chr20:34705516-34705537	present	58476	Promoter
GGGTGGGGATGGGGGGAGGG	chr20:43934072-43934091	present	84969	Intron
GGGTTGGGGGGTGGGGTGGG	chr20:46247013-46247032	present	64405	Promoter
GGGTTGGGGGTAGGGGGTGGG	chr21:32537338-32537358	present	59271	Intergenic
GGGTTGCGGGGTGGGGTGGG	chr22:11262057-11262076	absent	81061	Intergenic
GGGACGGGGGTGGGGTGGGG	chr22:18855039-18855058	present	8214	Intergenic
GGGGTGGGGTTGGGGTGGGG	chr22:19999303-19999322	present	421	Intron
GGGGAGGGGGGAGGGATGGGG	chr22:21451131-21451151	absent	23119	3' UTR
GGGTGGGGATGGGGTGAGGG	chr22:48336509-48336528	present	100422916	Intergenic
GGGTTGGGGGTGGGGGAGGG	chr2:10264610-10264629	present	3241	Intergenic
GGGTTGGGGGTGGGATGGGG	chr2:134334522-134334541	absent	4249	Intron
GGGTGGGGATGAGGGTGAGGG	chr2:136030238-136030258	present	101928243	Intergenic
GGGTGGGGATGGGGAGAGGG	chr2:186884280-186884299	present	151112	Intergenic
GGGGATGGGGGAGGGATGGGG	chr2:235026100-235026120	present	23677	Intron
GGGTTGGGGTTGGGGATGGGG	chr2:237474553-237474573	present	79083	Intergenic
GGGTGGGGGGAGGGGGAGGG	chr2:71452928-71452947	present	8291	Promoter
GGGTTGGGAGGTGGGGGAGGGG	chr2:80306685-80306706	absent	1496	Promoter
GGGTGTGGGTGAGGGTGAGGG	chr3:105648950-105648970	absent	214	Intergenic
GGGTGGGGGTGGGGAGAAGGG	chr3:14867703-14867723	absent	152273	Intron

GGGGTAGGGGTAAGGGATGGGG	chr3:170871296-170871317	absent	200916	Promoter
GGGTCCGGGGTGGGGTGGGG	chr3:183908290-183908309	present	100616127	Intergenic
GGGTTGGGGGAGGGATGGGG	chr3:186926653-186926672	present	6480	Promoter
GGGCTGGGGGTTGGGTGGGG	chr3:50367106-50367125	present	11068	Promoter
GGGTATTGGGGTGGGGTGGGG	chr3:73674103-73674123	present	23024	Intergenic
GGGTGGTGGGTGGGGTGGGG	chr3:9937507-9937526	present	78987	Promoter
GGGTGGGGTGGGGTGAGGGG	chr4:13540437-13540456	present	579	Promoter
GGGTTGGGGGTGGGGCAGGG	chr4:1780111-1780130	absent	2261	Intergenic
GGGTTGGGGGTGGGGCAGGG	chr9:99297662-99297681	present	100996569	Intergenic
GGGGTGGGGGTAGGGAGGGG	chr4:185101976-185101995	present	291	Intron
GGGCTGGGGCTGGGGTGGGG	chr4:3820535-3820554	present	152	Intergenic
GGGTCGGGGGTGGGGCGAGGG	chr4:94451863-94451883	absent	10611	Promoter
GGGGAGGGGGAGGGATGGGG	chr5:140141772-140141791	present	101929719	Intergenic
GGGTTGGGGGCAGGGTTGGGG	chr5:64122940-64122960	absent	285671	Intergenic
GGGGTGGGGGAAGGGAAGGG	chr6:35743073-35743092	present	221481	Intron
GGGTTGGGGGTTGGGGGTGGGG	chr6:43679847-43679868	present	55168	Intron
GGGATGGGGGTGGGGGAGGG	chr6:57178779-57178798	present	101927211	Promoter
GGGGTGGGGGAGGGACGGGG	chr7:100472520-100472539	present	81628	Promoter
GGGCAGGGGGTGGGGGAGGG	chr7:128832311-128832330	absent	2318	Promoter
GGGTTGGGGAGAGGGATGGG	chr7:149751402-149751421	absent	84626	Intergenic
GGGTTGGGGGAGTGGGAGGG	chr7:26621451-26621470	present	285941	Intergenic
GGGTTTGGGGAGGGAAGGGG	chr7:4482507-4482526	absent	221937	Intergenic
GGGTGGGGGCGGGGGAGGG	chr8:101761018-101761037	present	83988	Intron
GGGGTGGGGGCGGGGGAGGG	chr8:113439774-113439793	present	114788	Promoter
GGGTTTGGGGTGGGGTGGGGG	chr8:130519615-130519635	present	50807	Intergenic
GGGTGTGGGGGTTGGGGGAGGG	chr8:142753544-142753565	absent	137797	Promoter
GGGCCGGGGGTGGGGGAGGG	chr8:25689097-25689116	absent	64641	Intergenic
GGGTTGGGGCTTGGGGAGGG	chr8:26653400-26653419	present	1808	Intron
GGGTTGGGGGTGGGGTGGGGG	chr9:114098608-114098628	present	113220	Promoter
GGGTTGTGGGTGGGGATGGGG	chr9:116931301-116931321	present	22954	Intron
GGGATTGGGGATGGGGTGGGG	chr9:134806401-134806421	present	1289	Promoter
GGGGTTGGGGGTGGGGAGGG	chr9:91299523-91299542	present	549	Intron
GGGTGGGCGCGGGGTGAGGG	chr9:98745018-98745037	absent	203286	Intron
GGGTTAGGGGGAGGGGTGGGG	chrX:120755247-120755267	present	643311	Intergenic

Appendix Table A 2 Summary of Family 79 G4 sequences

Sequence	Location	Experimental Evidence	Gene ID	Annotation
GGGAGGGGAGGGGAGGGG	chr1:11692947-11692964	present	374946	Promoter
GGGAGGGGAGGGGAGGGG	chr1:32936592-32936609	present	127544	3' UTR
GGGAGGGGAGGGGAGGGG	chr10:12887513-12887530	present	83643	Intergenic
GGGAGGGGAGGGGAGGGG	chr11:47595913-47595930	present	114900	Promoter
GGGAGGGGAGGGGAGGGG	chr13:99003017-99003034	present	23348	Intron
GGGAGGGGAGGGGAGGGG	chr16:54930945-54930962	present	10265	Promoter
GGGAGGGGAGGGGAGGGG	chr4:151016295-151016312	present	987	Promoter
GGGCGGGGCGCGGGCGGGG	chr10:133262048-133262066	present	101	Promoter
GGGCGGGGCGCGGGCGGGG	chr10:133262092-133262110	present	101	Promoter
GGGGCGGGGAGGGGCGGGG	chr10:14604436-14604454	absent	83641	Promoter
GGGGCGGGGCGGGGCGGG	chr10:17348902-17348919	absent	33859	Intron
GGGGCGGGGCGGGGCGGG	chr11:533399-533416	present	3265	Promoter
GGGGCGGGGCGGGGCGGG	chr14:89701705-89701722	absent	1112	Intron
GGGCGGGAGGGGCGGGGG	chr10:3172964-3172980	present	10531	Promoter
GGGGAGGGGCGGGGGGG	chr11:115164500-115164517	present	23705	Intergenic
GGGGCGGGGGGGGGCGGG	chr11:134401965-134401982	present	27087	Intron
GGGAGGGGCAGGGGCAGG	chr11:2571906-2571924	absent	3784	Intron
GGGAGGGGCAGGGGCAGG	chr5:149676067-149676085	present	389337	Intergenic
GGGCGGGGCAGGGGCGGG	chr11:2902105-2902122	present	5002	Promoter
GGGCGGGGAAGGGGTGGG	chr11:2902676-2902693	present	5002	Promoter
GGGGGGGAGGGGCGGGGG	chr11:64342922-64342938	present	283234	Promoter
GGGAGGGGCCGGGGCGGG	chr11:6473947-6473965	present	10612	Promoter
GGGCGGGGCGGGGCGGGG	chr11:65558635-65558652	absent	4054	Promoter
GGGCGGGGCGGGGCGGGG	chr21:45555638-45555655	present	6573	Intergenic
GGGCGGGGCGGGGCGGGG	chr9:124777075-124777092	absent	169611	Promoter
GGGAGGGGCGGGGCGGG	chr11:7020387-7020403	present	7761	Promoter
GGGAGGGGCGGGGCGGG	chr16:371393-371409	present	10573	Promoter
GGGAGGGGCGGGGCGGG	chr19:17747734-17747750	present	23149	Promoter
GGGAGGGGCGGGGCGGG	chr19:44847716-44847732	present	5819	Promoter
GGGAGGGGCGGGGCGGG	chr9:70258881-70258897	absent	100507299	Promoter

GGGGCAGGGGCAGGGCGGG	chr11:79081753-79081771	present	260 11	Intron
GGGGAAGGGGGGGCGGGG	chr12:111597150-111597167	present	631 1	Promoter
GGGCGGGGCGGGGCGGG	chr12:48350921-48350937	absent	121 274	Promoter
GGGGCGGGGCGGGGGGG	chr14:105441449-105441465	present	911 2	Promoter
GGGGCGGGGCGGGGGGG	chr5:70126581-70126597	absent	660 6	Intergenic
GGGGCGGGGCGGGGGGG	chr5:71126892-71126908	absent	728 340	Intergenic
GGGGCGGGGCGGGGGGG	chr7:99770423-99770439	present	157 6	Intron
GGGGCGGGGCGGGGGGG	chrX:132751250-132751266	present	100 874 102	Intron
GGGCGGGGCGGGGGCGG	chr14:22871353-22871370	present	260 20	Promoter
GGGCGGGGCGGGGGCGG	chr22:37519650-37519667	absent	297 75	Promoter
GGGGCGGGGCGGGCGCGGG	chr14:55052093-55052111	present	934 87	Promoter
GGGCGGGGCGGGGCGCGGG	chr14:99645483-99645501	absent	844 39	Promoter
GGGGCGGGGCGGGGCGGG	chr1:3718211-3718229	absent	716 1	Intron
GGGGCGGGGCGGGGCGGG	chr15:101724648-101724666	present	123 283	Promoter
GGGGCGGGGCGGGGCGGG	chr17:62808389-62808407	absent	162 333	Promoter
GGGGCGGGGCGGGGCGGG	chr2:19348311-19348329	present	100 616 307	Promoter
GGGGCGGGGCGGGGCGGG	chr22:15761338-15761356	absent	106 146 148	Promoter
GGGGCGGGGCGGGGCGGG	chr22:42500547-42500565	present	940 09	Promoter
GGGCGGGGAGGGGGCGGG	chr15:39366098-39366115	present	400 360	Intron
GGGAGGGGACGGGGCGGG	chr15:69298887-69298904	absent	548 52	Promoter
GGGTGAGGGGGCGGGCGGG	chr15:88257743-88257761	absent	491 6	Promoter
GGGAGGGTGAGGGGAGGGG	chr15:92439783-92439801	present	812 8	Intron
GGGCAGGGGAGGGGCGGG	chr16:21511745-21511762	present	100 500 917	Intergenic
GGGCAGGGGAGGGGCGGG	chr16:87860829-87860846	present	814 0	Promoter
GGGTGAGGGGCGGGCGGG	chr16:22008068-22008085	present	730 094	Promoter
GGGGAGGGGGCGGGTCGGG	chr16:66604580-66604598	present	123 920	Promoter
GGGCGGGGAGGGCGCGGG	chr16:706088-706105	absent	146 330	Promoter
GGGCGGGGAGGGCGCGGG	chr16:67530169-67530186	present	795 67	Promoter
GGGAGCGGGAGGGGCGGGG	chr16:8847828-8847846	present	537 3	3' UTR
GGGCGGGACGGGGCGGGG	chr16:88686619-88686636	present	333 929	Promoter
GGGCGGGGTAGGGCGGGG	chr16:9091577-9091594	present	290 35	Promoter
GGGGAGGGGGCGGGGGGG	chr17:3635824-3635841	present	237 29	Promoter
GGGGAGGGGGCGGGGGGG	chr17:65055061-65055078	present	106 72	Promoter

GGGGAGGGGCGC GGGCGGG	chr17:43755320-43755338	absent	509 64	Promoter
GGGCGGGGGGG GCGGG	chr17:518280-518295	present	552 75	Promoter
GGGCGGGGACGG GGCGGGG	chr17:64130213-64130231	present	208 1	Promoter
GGGGCGGGGTGG GGCGGG	chr17:74380229-74380246	present	350 383	Intergenic
GGGTGAGGGCGG GGCTGGG	chr17:75740069-75740087	absent	369 1	Promoter
GGGGAAGGGGG CGGGGGGG	chr17:81067374-81067392	present	104 58	Promoter
GGGCCGGGCGGG GCGGG	chr17:81716621-81716637	present	146 8	Promoter
GGGGGGGCAGG GGCGGGGG	chr17:9825847-9825865	present	934 0	Promoter
GGGCAGGGGCAG GGCGGGG	chr18:37517402-37517420	present	568 53	Intron
GGGGCGGGGCGG GCCGGG	chr19:1418556-1418573	present	265 28	Promoter
GGGCGGGGAGG GGCGGG	chr19:1940858-1940874	absent	145 5	Promoter
GGGCGGGGCAGG GCGGGG	chr19:3557619-3557636	absent	126 321	Promoter
GGGAGGGTGAGG GGCGGGG	chr19:45497095-45497113	present	625 3	Promoter
GGGCGGGGAGG GGCGGGG	chr1:180632001-180632018	present	921 3	Promoter
GGGCGGGGAGG GGCGGGG	chr19:49388163-49388180	present	147 872	Promoter
GGGGGGGAAGG GCGGGG	chr19:49556137-49556153	present	510 70	Promoter
GGGCGGGGCCGG GGCGGG	chr19:52690611-52690628	present	557 69	Promoter
GGGGCGGGCGGG GCGGG	chr19:676544-676560	absent	102 72	Promoter
GGGGAGGGGGG CGGGGGGG	chr1:10856271-10856289	present	548 97	Intergenic
GGGGAGGGGCG GGGCAGGG	chr1:110161522-110161540	present	388 662	Intron
GGGGCGGGGCGG GGTGGG	chr1:156677373-156677390	present	107 63	Promoter
GGGGCGGGGCGG GGTGGG	chr3:141402686-141402703	present	253 461	Promoter
GGGGGGGCGGG GCGGGG	chr1:209825787-209825803	present	270 42	Promoter
GGGCGGGAAGG GCGGGG	chr1:228276035-228276051	present	840 33	Promoter
GGGCGGGGGAG GGGCGGGG	chr1:33182128-33182146	absent	552 23	Promoter
GGGGTGGGGGGG GCGGGG	chr1:42731825-42731842	present	149 461	Downstream
GGGGAGGGCAG GGCGTGGG	chr21:40155906-40155924	absent	100 616 148	Intron
GGGGCGGGGCCG GGGCGGG	chr21:42975112-42975130	present	531 6	Promoter
GGGCGGGGAGG GGAGGGG	chr22:27554681-27554698	present	433 0	Intergenic
GGGGGGGGCGG GGCGGGG	chr22:33528130-33528147	present	921 5	Intron
GGGGCGGGGCGG GGCAGGG	chr22:44891552-44891570	present	237 79	Intron
GGGGAGGGCTGG GCTGGGG	chr22:44916501-44916519	present	112 885	5' UTR
GGGAGGGGCAG GGCGGGG	chr2:11130389-11130406	present	285 150	Promoter
GGGAGGGGAGG GGCGGG	chr2:131005683-131005699	present	506 49	Intron

GGGACGGGGCGG GGCGGG	chr2:197785363-197785380	present	660 37	Promoter
GGGGTGGGGCGG GGCGGGG	chr2:231683502-231683520	present	575 7	Intergenic
GGGGAGGGGGC GGGGCGGG	chr2:44778533-44778551	present	798 23	Intergenic
GGGGAGGGCGG GTGCTGGG	chr3:127140774-127140792	present	285 311	Intergenic
GGGAGGGGAGG GGCGGGG	chr3:13018647-13018664	present	992 2	Intron
GGGAGGGGAGG GGCAGGG	chr4:1766503-1766520	absent	104 60	Intergenic
GGGCGGGGCGGG CGGGG	chr4:2418706-2418722	absent	577 32	Promoter
GGGGAGGGGCA GGGCTGGG	chr4:3779401-3779419	present	152	Intergenic
GGGCGGGAAGG GGCGGGG	chr4:41360750-41360767	present	229 98	Promoter
GGGAGGGGCAG GGGCGGGG	chr4:6782547-6782565	absent	977 8	Promoter
GGGCGGGGAGG GGGCGGGG	chr5:179795444-179795462	present	405 6	Promoter
GGGGAGGGGCA GGGCGGGG	chr5:2011684-2011702	present	101929081	Intergenic
GGGATGGGGAGG GGCGGGG	chr5:72299335-72299353	present	231 07	Intron
GGGAGGGGAAG GGGCGGGG	chr6:34223494-34223512	present	315 9	Intergenic
GGGGAGGGGAG GGCGCGGG	chr7:148884873-148884891	absent	214 6	Promoter
GGGGAGGGGGG GGCGGGG	chr7:151520014-151520031	present	600 9	Promoter
GGGCGGGGAGG GCCGGGG	chr7:1669902-1669919	present	392 617	Intergenic
GGGCAGGGCGGG GCGGG	chr7:2096544-2096560	present	837 9	Intron
GGGTAGGGGCGG GGCGGGG	chr7:75410724-75410742	absent	378 108	3' UTR
GGGCGGGGAAG GGCGGGG	chr7:7566848-7566865	absent	544 68	Promoter
GGGCGGGGCGGG GCCGGG	chr7:99375518-99375535	absent	100 95	Promoter
GGGGAGGGGTCG GGGGGG	chr8:133453251-133453268	present	648 2	Downstream
GGGAGCGGGGCG GGGCGGG	chr8:26383646-26383664	present	665	Promoter
GGGAAGGGGCA GGGCGGGG	chr9:1046289-1046307	present	102800446	Promoter
GGGGGGGCGGG GCGGG	chr9:113012640-113012655	present	169 834	Promoter
GGGGCGGGGCGG GGCCGGG	chr9:124853507-124853525	absent	401 551	Promoter
GGGCGGGGAGG GCCGGG	chr9:133557493-133557509	present	971 9	Intron
GGGGCGGGGTCG GGGCGGG	chr9:5629030-5629048	present	575 89	Promoter
GGGGCGGGCCGG GGCGGG	chr9:91423955-91423972	present	478 3	Promoter
GGGGCGGGCGCG GGGCGGG	chrX:153411372-153411390	present	139 735	Intergenic
GGGCGGGGCGGG GCGAGGG	chrX:154546963-154546981	present	253 9	Promoter
GGGCGGGCCGGG GCGGGG	chrX:155458643-155458660	absent	826 3	Promoter
GGGGAGGGGCG GGGCGGGG	chrX:76427616-76427634	present	576 92	Promoter

Appendix Table A3 Sequence repeats capable of forming multiple G4 structures.

Location	Gene ID	Gene Symbol	Sequence
chr17:81432609-81432922	57597	BAHCC1	GGGGATGGGGGGAAGGGTGGGCCTGGGGATGGGGGGGGAGGTTGGGCCTGGGGATG GGGGGAAGGGTGGGCCTGGGGATGGGGGGGGAGGGTGGGCCTGGGGATGGGGCGGA GGGTGGGCCTGGGGATGTGGGGGGAGGGTGGGCCTGGGGATGGGAGGGAGGGTGGG TCTGGGGCTCGGGGGGAGGGTGGGTCTGGGGATGGCGGGGAGGGTGGGCCTGGGGAT GCGGGGGAGGGTGGGCCCGGGGATGGGAGGGATGGGGTCTGGGGCTCAGGGGAACA GGTGGGCCCGGGAGGGTGGGTCCAGGGCTGGGG
chr20:63356114-63356297	113 7¶1 001 305 87	CHRNA4¶LOC100130587	GGGTGAGGGGTGTGGGGGAGGGCAGGGGCGGGACAGGGCCAGGGTGGGGCAGGGGC AGGGCCAGGGCAGGGCAGCAGGGTGAGGGGTGTGAGGGAGGGCAGGGGCGGGACAG GGCCAGGGTGGGGCAGCAGGGTGAGGGGTGTGAGGGAGGGCAGGGGTGGGGCAGGG CCAGGGTGGGGCAGGG
chr19:2361178-2361341			GGGGCTGGGGTGGGAGGCTGGGGTGGGAGGTGGGGCTGGGGTGGGAGGCTGGGGTG GGAGGTGGGGCTGGGGTAGGGTGCAGGGTTGGGATGGGGTGGGGTGCTGGGATGGGT TGGTGGGGTGCAGGGCTGGGATGGGTTGGTGGGGTGGGGTGCGGGGCTGGG
chr20:63836179-63836316	140685	ZBTB46	GGGAGAGCGCGGGCAGGGCAGGGGAGAGGGCAGGGGCAGGGAGAGCGTGGGCAGG GCAGGGGCAGGGGAGAGCGTGGGCAGGGCAGGGGCAGGGAGAGGGCAGGGGCAGGG AGAGCGCGGGCAGGGCAGGGGCAGGGG
chr1:1115450-1115668	54991	C1orf159	GGGGCAGGGAAGGGGGTCCCTGGGGAGGGGAGGGGGTCCCTGGGGAGGGAGGGGGG TCCCTGGGGAGGGGAAGGGTTCCCGGGGCAGGGAAGGGGGTCCCTGGGGAGGGGAG GGGGTCTCTGGGGGAGAGGGGGCACCTGGGGAGGGGAGGGGGTTCTCGGGGCAGGG AGGGGGTCCCTGGGGAGGGGATGGGACGGGCGCCCGGGGGAAGAAAGGGGG
chr1:23568081-23568175			GGGACCGCCGGGGGGTTGGGGCCGCCGGGCGCGGGGCGCGGGCGGGGAGGGCCGGGT AGGGCCAGGGTGGGGCCGCCGGGGCGCGGGCAGGGAGGG
chr1:2641112-2641244	100287898	TTC34	GGGGAGGGCTGGGAAGGGGGAGGGCTGGGAAGGGGTGTGTGGGGAGGGTGGGAAGG GGGGCTGTGGGGAGGGTTGGGAAGGGGGGTTGTGGGGAGGGTTGGGAAGGGGTGTGT GGGGAGGGCTGGGAAGGGGG
chr10:101829586-101829668	30819	KCNIP2	GGGATAGATGGGGCTGGGCGGGGCGAGGGGAGGGGCTGGGTGGGACGAGGGGAGGG TTTGGGCGGGGCAAGGCTGGGGCTGGG
chr10:132448692-132449011	170393	C10orf91	GGGGCAAGGGTGGGAGGGGCCGGGCAGGACTGGGGCAAGGGTGGGAGGGGCCGGGC AGGACTGGGGCAAGGGTGGGAGGGGCCGGGCAGGACTGGGGCAAGGGTGGGAGGGG CCGGGCAGGACTGGGGCAAGGGTGGGAGGGGCCGGGCAGGGCTGGGGCAAGGGTGG GAGGGGCCGGGCAGGACTGGGGCAAGGGTGGGAGGGGCCGGGCAGGACTGGGGCAA GGGTGGGAGGGGCCGGGCCGGACTGGGGCAAGGGTGGGAGGGGCCAGGCAGGGCTG GGGCAAGGGTGGGAGGGGCCGGGCCGGACTGGGGCAAGGG
chr18:78667905-78667977			GGGTGGCTAGGGGAGAGGTGGGAGGGTGGATGGGGGAGGGTGGATGGGGTAGGGTT GGGAGGGTGGATGGGGG
chr20:63836346-63836553			GGGCAGGGGAGAGGGCAGGGGAGAGGGCAGGGGCAGGGAGAGCGAGGGCAGGGGA GGGGGCAGGGAGAGCGCGGGCAGAGCAGGGGCAGGGAGAGCGCGGGCAGGGCAGGG GAGCGGGCAGGGGCAGGGAGAGCGCGGGCAGGGCAGGGGCAGGGAGAGTGCGGGCA GGGCAGGGAGAGCGCGGGCAGGGCAGGGGCAGGGGCAGGGG

chr21:42100704-42100958	89766	UMODL1	GGGGTTGGTGGGTAGGGAGGGTGAGGGCATGGGGTTGGTGGGTGGGGAGGGTGAGGAGGGCACGGGGTTGGTGGGGTGAGGAGGGCACGGGGTTGCTGGGGTGGGGAGGGTGAGGAGGGCACAGGGTTGGTGGGGTGGGGAGGGTGAGGAGGGCATGGGGTTGCTGGGGTGGGGAGGGTGAGGAGGGCACGGGGTTGCTGGGGTGGGGAGGG
chr4:3730542-3730742			GGGGGCTGGGGGTGCAGGGTGGGTGCCGGTGGGGGCTGGGGGTGCAGGGTGGGTGCCGGTGGGGATTGGGGGCACAGGGCGGGTGCCAGTGGGGGCTGGGGGTGCAGGGCAGGGGCCAGTGGGGGCTGGGGGTGCAGGGCAGGGGCCAGTGGGGGCTGGGGCCCCAGGGCGGGGGTGCCGGTGGGGGCAGGGGGCGCAGGG
chr8:143498684-143498908	23144	ZC3H3	GGGGGATGGGAACCGGGGGCAGAGGGGTACAGGGCGGGGCGGAGGGGCACAGGGCGGGGCGGAGGGTACAGGGCGGGGCGGAGGGGCACAGGGCGGGGCAGAGGGGCACAGGGCGGGGCAGAGGGGTACAGGGCAGGGCAGGGGGTACAGGGCGGGGCGGAGGGGCACAGGGCGGGGCGGAGGGG
chr9:137515955-137516116	375775	PNPLA7	GGGATGGGGGTGGGGCGCATGGGGGATTGGGGTAGGGGTGGAAGGGGTGGGGAGGAGGGATTGGGGTGGGGAGGGGGATTTCGGGGAATGGGGGGTTGGGGTAGGTAGGGAGGATTTGGGGGAGGGGGGTTGGGGTAGGGAGGGAGGATTTGGGGGAGGGGGG
chr10:744186-744301		RP11-164C1.2	GGGCTGGGTGGGGGCCCTGGGCTGGGTGGGGACAGTGGGCTGGGTGGGGGCCCTGGGCTGGGTAGGGACACTGGGCTGGGTGGGGGCACTGGGCTGGGTGGGGGCACTGGGTTGGG
chr11:1234711-1234787	727897	MUC5B	GGGGAGGGCGGGGGCGGGGAGGGCAGCGGGTGGGGAGGCAGCGGGCAGGGAGGGCAGGGGGCGGGGAGGGCAGGGGG
chr15:80152528-80152585	2184	FAH	GGGCCTGGGACTCTCGGGTACCCGGGCCAGGGGAGGGGCGGGGCTCAGGGAGGGAGGG
chr15:80152616-80152668	2184	FAH	GGGGCGAGGGGAGGGGCGGGGCGAGGGGAGGGGCGGGGCTCAGGGAGGGAGGG
chr16:2905425-2905477			GGGACAGCGGGAAGGGCGGGGCCCGTGGGCAGGGGAGGGACCAGTGGGCGGGG
chr19:35141265-35141367	5348¶1631 75¶538 22	FXYD1¶CTD-2527I21.4¶LGI4¶FXYD7	GGGAGGGGTTGGGGCGGGGCGGGGCCAGGGAGGGGCGGGACCAGGGAGAAGCGGGGCCAGGGAGAGGGGGGCTAGGGAGAGGCGGGGCCAGGGAGAGGCGGGG
chr19:41383630-41383688	593¶64164 9	BCKDHA¶TMEM91	GGGTATGTTGGGTGGGGGAGGGACCAGGGGAAGGGTCTGGGACTGAGGGGATGCCTGGG
chr2:218980736-218980797	151300	LINC00608	GGGATGGGGTGGGGTGAGGGTTGGGTGGAAGGGTGGGTGAGATGGGGAGGGTGGGGTGCGGG
chr20:62498310-62498366			GGGATGGGTGGGTGTTGGGGATGGGTGGGTGTGGGGGTGGGTGGGTGTTGGGATGGG

chr20:62498625-62498714			GGGATGGGTGGGTGGTGGGGATGGGTGGGTGTTGGGGTGCATGGGTGTTGGGGTGGGTGGGTGTTGGGATGGGTGGGTGTTGGGGATGGG
chr22:38077913-38077956			GGGGCTGGGCGGGCACAGGGCTCAGGGGCACAGGGCTGGGTGGG
chr22:47343631-47343717			GGGCAGGGATTGGCAGGGAGCGGGTAGGGCAGGGACTGGGTTGGGTAGGGTAGGGATTGGGTAGGGCAGGGAGCGGGTAGGGCAGGG
chr4:49566547-49566621			GGGGGGCGGGGAGGGTTGGGGGATTGGGGGGCGAGGAGGGTTGGAGGGGATTGGGGGTTGGGAGGGTTGGAGGGG
chr5:1214849-1214907	340024	SLC6A19	GGGGGCCTGGGTAGGGAGGGTGGGCCCTGGGTGTGAGGGGCGGGGCTGGGCAGGGAGGG
chr5:176030647-176030835	84321	THOC3	GGGTGGGGGTGTGTGGGTGAGGGTGTGGGTGGGGGTGTGTGGGTGAGGGTGTGGGTGGGGGTGTGTGGGTGAGGGTGTGGGTGGGGGTGTGTGGGGGTGTGGGGGTGTGGGCGAGGGTGTGGGTGGGGGTGTGGGGGGGAGGGTTTGGATGGGGG
chr5:176030847-176031041	84321	THOC3	GGGTGTGGGTGAGGGTGTGGGTGGGGGTGTGGGCGAGGGTGTGGGTGGGGATGTGAGGGTGGGGGTGTGTGGGTGAGGGTGTGGGTGGGGGTGTGTGAGGGTGGGGGTGTGTGGGTGAGGGTGTGGGTGGGGGTGTGTGGGTGAGGGTGTGGGTGGGGGTGTGTGGGTGAGGG
chr7:150995720-150996053	4846	NOS3	GGGGTAGGGGCAGAGGGAGGAGGGGTGCAGGGGTGGGATGGGAGCAGAGGGAGGAGGGGTGCAGGGATGGGACGGGGAGAGGGAGGAAGGGTGCAGGGATGGGACAGGGGAGAGGGAGGAGGGGTTTAGGGGTGGGATGGGGGAGAGGGAGGAGGGGTGTAGGGGTGGGACGGGGCAGAGGGAGGAGGGATGCAGGGGTGGGACGGGGGAGAGGGAGGAGGGGTGCAGGGGTGGGTCGGGGCAGAGGGAGGAGGGGGTGCAGGGGTGGAATGGGGCAGAGGGAGGAGGGGTGTAGGGGTGGG
chr8:22733773-22733858	157310	PEBP4¶RP11-459E5.1	GGGGTGGGGATGGGGGGTTTGGGGAGGGTGGGGATGGGGGAATTGGGGAGGGTGGGGATGGGGGAATTGGGGAGGGTGGGGATGGG
chr1:214136289-214136320			GGGGGTGGTTGGGGGTTGGGGAGGGTTGAGGG
chr10:117501567-117501602	196047	EMX2OS	GGGGGCAGGGCCAGGGGCCTGGGCGGGCAGGCTGGG
chr10:132448613-132448675	170393	C10orf91	GGGTTGGGACGAGTGGGAGGGGCCAAGCAGGGGCTGGGCAAGGGTGGGAGGGGCCGGGCAGGG
chr10:132605375-132605400	3632	INPP5A	GGGGGATGGGGAAGGGACAGGGAGGG
chr10:133252389-133252567	574448¶10192767 1	MIR202¶MIR202HG	GGGATGGGGTACAGGGCAGGACGGGGTGCAGGGCAGGACGGGGTGCAGGGAGAGCTGGGGTGCAGGGAGGGATAGGGTGCAGGGCGGGCCGGGGTGCAGGGTGGGCTGGGGTGCAGGGTGGGGCGGGGTGCAGGGCGGGCTGGGGTGCAGGGAGGGATGGGGTGCAGGGCGGGGCCGGGG
chr10:42884606-42884649			GGGGGCAGGGAGAGGGATGGGTTAGGGCTGGGTGGGGACCAGGG
chr11:1224526-1224599	727897	MUC5B	GGGGCGGGGGAGGGGCTGTAGGGCCAGGGAGGGGCTGCCTGGGGCTGGGGAGGGGCTGCTGGGGTGGGGAGGGG

chr11:33700500-33700555	100131378¶966	C11orf91¶CD59	GGGTGGTGGCGGGGGCGGGGGCGCCGGGGCGGGGAGCGCCTGGGACAGGGACCGGG
chr12:108850623-108850742	54434	SSH1	GGGGACTTCAGGGAGGGGAGATGGGAGAGGGGATTGGGGAGGGATGGTTAGGGATGGGATCTAGGGAGGGGATTTGGGGGAGGGAAGCTTGGGGAGGGGATCTGGGGGAGGGGAGTGGGG
chr12:113358512-113358546	196463¶80024	PLBD2¶SLC8B1	GGGCGGGGCCGAGGGCGGGCGGGGCCGGGCTTGGG
chr12:131894875-131894930	8408	ULK1	GGGCCGGGGCGGGCGGGCCGGGGGCGCGGGGCCGGGGGCGCGGGGCCGGGCGAGGG
chr15:73685123-73685163	80381	CD276	GGGGGTGGGGAGGGAATTGGGATGGGCAGTTTGGGCTTGGG
chr16:89828297-89828329	84501	SPIRE2	GGGGGACGGGTGAGGGGCAGGGCGGGGCGCGGG
chr17:82709075-82709123			GGGGGAGGGCATGGGGGCAGTGGGTTGGGTGGGTGGTAAGGGGAGAGGG
chr18:77117772-77118250	4155	MBP	GGGATGGGGGTCAGTGGGTTGGGGTGGGACGGGGGTCAGTGGGTTGGAGTGGGATGGGGGTCAGTGGGTTGGAGTGGGATGGGGTTGGAGTGGGACAGGGGACAGTGGGTTGGAGTGGGAGGGGGGACATTGGGTTGGGGTGGGACGGGGGTCAGTGGGTTGGGGCGGGATGGGGTCAGTGGGTTGGGGTGGGACGGGGACAGGGGGTTGGGGCGGGATGGGGGACAGTGGGTTGGGGTGGGATGGGGACAGTGGGTTGGGGCGGGATGGGGACAGTGGGTTGGGGCGGGATGGGGACAGTGGGTTGGGGTGGGATGGGGACAGTGGGTTGGGGTGGGATGGGGACAGTGGGTTGGGGTGGGATGGGGACAGTGGGTTGGGGTGGGATGGGGTCAGTGGGTTCGGG
chr18:79617375-79617447			GGGGGGCGGGGTCCCGGGGGGGGCGGGGTCCCGGGGAGGGCGGGGTCCCGGGGGGGGCGGGGTCCCGGGGAGGG
chr18:80247568-80247654	84552	PARD6G	GGGGCGGGGCTGTGGGCGGGGATATGGGCGGGTCCTCCGGGGCTCGGGGGCGGGGCGGGGCGCGGGCGCGGGCGCCGGGAGGTGGGG
chr19:43479561-43479598	653583	PHLDB3	GGGCCTGGAGGGTGGGGTGGGGTGGGTGGGGCCTGGGG
chr19:45782822-45782871	17600¶1762	DMPK¶DMWD	GGGGGCTAGGGGGTGAGGGCTGGGGGTTGGGGCTGGGTGGGAGAAAGGGG
chr19:47480224-47480277	11133¶100505681	KPTN¶NAPA-AS1	GGGGTAAGGGGTGGGGTTGAGGGCTAGAGGGCGGGGCCAGGGTGGGGCTGAGGG
chr2:218980655-218980727	151300	LINC00608	GGGGTGGGGTGAGGGGCGGGTGGAAGGGTGGGTGAGATGGGGAGGGTTGGGTGGGGTGGGAGTGGGGGTTGGG
chr2:2236336-2236438	23040	MYT1L	GGGTTGGGTTGTACTGGGCTGGGTTGGGATGTACTGGGTTAGGTTGGGATGTACTGGGGTGGGTTGGGATGTACTGGGGTGGGTTGGGATGTACTGGGTTGGG

chr2:241896728-241896872	101927289	AC131097.4¶LINC01237	GGGGTGTGGGTGAAGGGGTGTGGGTGTTGGGGTGTGGGTGTTGGGGTGTGGGTGAAGGGGGTGTGGGTGTTGGGGTGTGGGTGTGTGGGTGTGGGGGTGTGGGTGTTGGGGTGTGGGTGAAGGGTGTGGGTGTGGGGGTGGGGG
chr2:64876270-64876313			GGGTGGGTGGGTTTGGAGGGAGGGGGTTGGGAATAGGGGTGGGG
chr20:61950814-61950870			GGGCTGGGGAAGGGGAGAGGGTGGGAGGGAGGGCCCAGGGCTGGGGGAGGGAAGGGG
chr20:62497558-62497620			GGGGAAGGGGGGGTGGGTGTGGGGGATGGGTGGGTGTTGGGATGGGTGGGTGTTGGGGATGGG
chr20:62497759-62497831			GGGGTGGGTGGGTGTTGGGGTGGGTGGGTGTTGGGGATGGGTGGGTGTTGGGGTGGGTGGGTGTTGGGATGGG
chr20:62498375-62498431			GGGGATGGGTGGGTGTTGGGATGGGTGGGTGTTGGGGTGGGTGGGTGTTGGGATGGG
chr20:62498483-62498530			GGGTGTTGGGGATGGGTGGGTGTTGGGGTGGGTGGGTGTTGGGATGGG
chr20:62498556-62498616			GGGGTGGGTGGGTGTTGGGGTGGGTGGGTGTTGGGATGGGTGGGTGCTGGGGATGGGCGGG
chr21:13368441-13368521			GGGGGTGTGTGGGTGGGGGTGTGTGGGTGTGGGTGAGGGTGTGTGGGTAGGGGTGTGTGGGTAGGGGTGGGTGGGTTAGGG
chr22:47342809-47342880			GGGTAGGACAGGGATTGGGTAGGACAGGGATTGGGTAGGGAGGGGATTGGGTGGGACAGGGATTGGGTAGGG
chr3:195909590-195909656	100128262¶10188	TNK2-AS1¶TNK2	GGGCGGGGAGGCGGGCGCGGGACTCGGGGGCGGCCCGGGGCGGGAGGGGAGGGCCGGGGCGGGCGGG
chr4:15937283-15937321	9982	FGFBP1	GGGGTTTGGGTGGGGTGGGTGGGGTTGTGGGGGTGGGGG
chr5:1028761-1028825	85409	NKD2	GGGACAGGGCCAGAGGGATTGAGGGGGAGGGTTGGGCCCTCGGGAGGGAGTGAGGGTAGGGTGGG
chr5:1448491-1448526	6531	SLC6A3	GGGGGTAGGGGGCTTGGGGAGGGCAAGGGTGAGGGG
chr5:170278314-170278414	3937	LCP2	GGGGGGCTGGGGAGGGGGATGGGAGTGCAGGGGGGGGCTGGGGGGGGATGGGAGTGCAGGGGTGGGGGCTGGCGAGGGGGATGGGAGTGCAGGGGTTGGGGG
chr6:163414542-163414582	9444¶1005268	QKI¶CAHM	GGGGAGGGCCGGGGGCGGGCGGCGGGCGGGGCCGCGCAGGG
chr7:100601135-100601174	10012984551¶51	PCOLCE-AS1¶PCOLCE	GGGGGCTGGGAGGGACAGGGGCAGGGATGGGGTGGGTGGG

chr7:1607861-1607906			GGGCTGGGGCTGGGCAAGGGTTGGGCTGGGCAGGGGTTGGGCAGGG
chr8:102411698-102411730	51366	UBR5	GGGGCCAGGGGCCCCGGGGTGGGCTGGGCTGGG
chr8:105318225-105318271	23414	ZFPM2	GGGGGATGGGGGCAGGGGTGGGTGGGAGTTGGAGGGGAGGCGTGGGG
chr8:132076387-132076451	100 86¶ 729 330	HHLA1¶ OC90	GGGGGTTTGGGGGTGGGGGATGGGAGGGGTGGGAAGCTTGGGGACAAGGGCGTGGT GGGATGGGG
chr8:133653707-133653756			GGGTGGGGTCTGGGGTGGGGTGTGGGTGCAGGGCACAGGGAGAAGAAGGG
chr9:111803461-111803494			GGGGGAAAGGGGTGGGTTGGGAGGGGTTGTAGGG
chr9:76756827-76756858	158471	PRUNE2	GGGGGGGCAGGGCACGGGCAGGGTGGGTGGGG
chr1:1202461-1202565			GGGTGCTGGGCTCGGGGCTGGGTACTGGGCTCGGGGCTGGGTACTGGGCTCGGGGCT GGGTACTGGGCTGGGGGCTGGGTACTGGGCTCGGGGCTGGGTACTGGG
chr1:165369760-165369792			GGGGTGGGAGTTGGGGCAGGGATGGGCAAGGGG
chr1:180631413-180631438	9213	XPR1	GGGGTTGGGGGTGAGGGTTGGGAGGG
chr1:2175339-2175394	55900 728 116	PRKCZ ZBTB8B	GGGGTTGGGGTTGGGATGGGGGTTGGGTAGATTTGGGGTTGGGATGGGGATGGGGG
chr1:32490207-32490237	¶65 312 1	¶ZBTB8A	GGGGTAGGGAGGGGCAAAGGGTTGGGATGGG
chr1:3454567-3454635	27237	ARHGEF16	GGGCGCTGGGCGGGCGGGGCTGCTCCGGGTCCGAGGGCCCGGGCGGGGAGGGCCGGG GAGGGGGCGGGG
chr1:53501990-53502054			GGGTTTGGGAGGGGTGGGAGGGACAAGGGGCCTGGGGTGGGAGGGACAAAGGGGCT TGGAAGGGG
chr10:71548130-71548176	64072 102	CDH23	GGGGCAGCGGGAGGGCTGGGATTGGGGAGGGAGCCCAGGGGACAGGG
chr11:1256504-1256551	466 725 ¶72 789 7	MIR6744 ¶MUC5B	GGGGCTCAGGGTGGGCGGGGCTGGGGGCGGGGACGGGTGGGGCTGGGG
chr12:122986600-122986642	576 05¶ 234 57	PITPNM2 ¶ABCB9	GGGGCCCGGGAGGATGGGGCAGGGCAGGGGTGGGGGCAGAGGG
chr13:114156015-114156095			GGGGAAGGGTGGGTGCTGGGCGTGGGGGTGGAGGGATGGGTGGGGTGGGTGCTGGGC CTGGGTTTGGAGGGGTGGGTGGGG

chr13:114156176-114156330			GGGTAGAGGGGTGGGTGCTGGGTGTGGGGGTGGAGGGGTGGGGAAGGGTGGGTGCTGGGCGTGGGGGGGGAGGGGTGGGGAAGGGTGGGTGCTGGGCGTGGATGGGGTGGGTGCTGGGCGTGGGGGTGGAGGGGTGGGGAAGGGTGGGTGCTGGG
chr13:18796421-18796470			GGGAGGGTGGGGAGGGTTGGGGGAATTGGGGGGTGGGGAGGGTTGGGGGG
chr13:50411581-50411619	10301	DLEU1	GGGGCTGTGGGGCTGGGCCAGGGAAGCGGGAGGGAAGGG
chr13:99988955-99988988			GGGCAGGGATAGGGGTGGGGTTGGGGTGGGGGGG
chr14:105154412-105154445	3714¶1024654 58	JAG2¶RP11-44N21.4¶MIR6765	GGGGCTGGGGCCCAGGGTCTGGGGTGGGCACGGG
chr15:80152687-80152733	2184	FAH	GGGGCTCGCGGGGTCGGGGCAAGGGGAGGGGCGGGGCTCAGGGAGGG
chr16:1372485-1372516	64718	UNKL	GGGGTTGGGGGTGGGGACTTGGGAGGGAAGGG
chr16:27805155-27805281	146395	GSG1L	GGGCACTGGGATTGGGGGTGGGAGGAGGGCTGTGGGCACTGGGATTGGGGGTGGGGGGAGGGCTGTGGGCACTGGGATTGGGGTGGGGGGAGGGCTGTGGGCACTGGGATTGGGGGTGGGGGGAGGG
chr16:57802324-57802362	3801¶3882 82	CTD-2600O9.1¶KIFC3¶LOC388282	GGGGGGCCCGGGCGGGCTGGGTGCGAGCGGGGCGCTGGG
chr17:20518605-20518666			GGGCCTGGCGGGACCCTGGGACCCAGGGCGGGGCTTGGGGTGGTGGGCAGGGCAGGGTGGGG
chr17:41625606-41625663	3872	KRT17	GGGACCCAGGGCCGGGCTTGGGGTGGTGGGCAGGGCAGGGTGGGGCTGTGGGGTGGGG
chr17:79495489-79495646	146713	RBFOX3	GGGGGAGGGGTGCAGAGGGTGGGGAGGTGGGGGAGGGGTGCAGAGGGTGGGGAGGTGGGGGGCAGGGGTACTGGGGGCAGGGATGGGGATGTGGGAAGGTGGGGGAGGGGTACAGAGGGTGGGGAGGTGGGGGAAAGGGTACTGGGGCAGGGATGGGGG
chr18:48161154-48161184	201501	ZBTB7C	GGGGGCTGTGGGGGAGGGAGGGCCTCCTGGG
chr19:16904176-16904217	27151	CPAMD8	GGGCAGGGCCAGAGGGAGGGGCTCAGGGCTGGGTGGGTGGGG
chr19:3276183-3276222	60680	CELF5	GGGGCGGGGCCTGGGGAGGGGATGGGGCTGCAGGGTGGGG
chr19:47480030-47480127	11133¶100505 681	KPTN¶NAPA-AS1	GGGTGAGGTAGGGGAAGTGGGGGAAGGGATGAGGGTGGGGATTGGGGGCGGGGCTAGGGTGGGGTTGAGGGCTTCAGAGGGGCGGGGCTAGGGTGGGG

chr19:56108248-56108430	126208	ZNF787	GGGCAGGGGCAGGGTGGAGGGAGCTGGGGAGCTCTGGGCAGGGGTTGGGTGGAGGGAGCCAGGGAGCTGGGCAGGGGCGGGGTGGAGGGAGCCGGGGAGCTCTGGGCAGGGGCTGGGTGGTGGGAGCTGGGAGTTCTGGGCAGGGGCGGGGTGGAGGGAGCCAGGGAGTGCTGGGCAGGGG
chr2:118875603-118875645			GGGGACTGGGAAGGGGTGGGACAGGGCTGGGGTGTGGGGAGGG
chr2:217821675-217821731	7145	TNS1	GGGAAGGGGCAGTGGGTGGGGGATGGGAATCCGGGCCTGGGACTGGGACGGGATGGG
chr2:218992604-218992681	1412	CRYBA2	GGGGCAGGGGTAGGGGGCGGCAGGGTGGGAAAAGCTGGGCTCTGGGAGACCAGGGTGGGGCCAGGGAAAGGGATTGGG
chr2:240216936-240216979			GGGTGAGGGGTGGGTGGGGTGGGGTGGAGGGACAGGGGTGGGGG
chr20:63356306-63356351	100130587¶1137	LOC100130587¶CHRNA4	GGGCAGCAGGGTGAGGGGTGTGGGGGAGGGCAGGGGTGGGGCAGGG
chr21:45999522-45999555	1291	COL6A1	GGGGCTGGGTAGGGAGGGACCGGGCAGGGGTGGG
chr22:19760592-19760651	6899	TBX1	GGGCGGGCTGGGGGCCGGGGAGGGGGAAGGGGCGGGGGAGGGGCGAGGGCCGGGGGAGGG
chr22:22558009-22558124	648691¶23532	LL22NC03-63E9.3¶PRAME	GGGTGGGGGAGGGGTGGGCATGGGGAAAGGGAGACAGGGTGGGGGAGGGGTGGGCATGGGGGAAGGGAGACAGGGTGGGGGAGGGGTGGGCATGCGGGAAGGGAGACAGGGTGGGG
chr22:38468717-38468749	11015	KDELR3	GGGGGGCGGGGTGGGGTCTGGGAAGGGGCAGGG
chr4:1318076-1318118	10296	MAEA	GGGGGAAGCCGGGCACACGGGGCCAGGGAGGCGGGTGGGTGGG
chr4:2756374-2756472	79155	TNIP2	GGGCGGGGCTGCGCGGGGAAGGGCGGGGCTGCGCGGGGGGCGGGGCTGCGTGGGGGAGGGCGGGGCCGGGCTGTTGTGGGTGGGCGGGGCGCACCGGGG
chr4:49572959-49573006			GGGGGGTGGGGAGGGTTGGGGGGATTAAGGGGTGGGGAGGGTTGGGGG
chr5:1295088-1295155	7015	TERT	GGGGAGGGGCTGGGAGGGCCCGGAGGGGGCTGGGCCGGGGACCCGGGAGGGGTCGGGACGGGGCGGGG
chr5:2204989-2205021			GGGAAGTGGGGGTGGGGCTGGGGCTTGGGCGGG
chr7:101017720-101018062	102724094¶14045 3¶1007 1	LOC1027240944¶MUC17¶MUC12	GGGGGAAGGGAAGGGGTCCCAGGGGGAGGGAGGGAGTCCCAGGGGGAAGGGAAGGGGTCCCAGGGGGAGGGGTCCCAGGGGGAGGGAGGGAGTCCCAGGGGGAAGGGAAGGGGTCCCAGGGGGAGGGAGGGAGTCCCAGGGGGAAAGGAAGGGGTCCCAGGGGGAAGGGAAGGGAGTCCCAGGGGGAGAAGGGAGTCCCAGGGGGAGAAGGGAGTCCCGGGGGAAGAGGG

chr7:129978853-129978883		RP11-306G20.1	GGGGGTGGGGGTGGGTGTGGGAGGGCAGGGG
chr7:149018792-149018849	9601	PDIA4	GGGGTGGGGGTGGGGGGGTAGGAGGGGGGAGTAGTGGGGTGGGCAGGGTGGGTTGGGG
chr7:26864641-26864689	8935	SKAP2	GGGCCGGGGCGGGGAGATGGGTGGGAAGGGACACGAAGGGCCTGAGGGG
chr7:44224942-44224975	816	CAMK2B	GGGAGGGGCTGGGCAGGGCTGGGAAAGGGGTGGG
chr7:74218035-74218071	7462	LAT2	GGGGGCTGGGGGTGGGCAGGGCCTGAGGGGAGAGGGG
chr8:142348488-142348542	203062	TSNARE1	GGGGCAGCCTGGGGCGGGAGCGGGGGCCAGGGGAGGGTGGGCATGGGGTGCCGGG
chr8:142348556-142348610	203062	TSNARE1	GGGGCAGCCTGGGGCGGGAGCGGGGGCCAGGGGAGGGTGGGCATGGGGTGCCGGG
chr8:144465691-144465773	90990¶50626	KIFC2¶CYHR1	GGGGAGAAGGGCCGGGGCGGGGCTGCGAGGGGCGGGGTCTGGGCGGGGCTGCGAGGGGCGGGGGGTCTGGGCGGGGCTGAGGG
chr9:32956186-32956230			GGGAGGGTCCGGGAAGGGTCCCTGGGTGGGGGAGGGGGAAAGGGG
chr9:39145609-39145636	79937	CNTNAP3	GGGGGCTGGGGGCATGGGGAGGGTAGGG
chr9:5840502-5840539			GGGGGCCGGGGAGGGGCAGTGGGCATGGGTAGGGAGGG
chr9:93210282-93210324	65268	WNK2	GGGGTGAGGGATGGGCAGGGTGGGCAGGGATGGGGGACTGGGG
chrX:107610224-107610250			GGGGGGCAGGGGCAGGGAAGGGGAGGG
chrX:12711746-12711782	9758	FRMPD4	GGGGACTTGGGGCGGGGGGCAGGGTTGGGGGGAAGGG
chr1:11971187-11971260	5351	PLOD1	GGGGCAGGGGGATGGGGTGGGAGGGGTAGGGTGGAGTGGGGGGCTTGGGTGGAAGGGCCAGGGGTGGGTGGGGG
chr1:38005556-38005593	2275¶51118	FHL3¶UTP11	GGGCTGGGGGGCCGGGGGCGGGGTCCGGGCGGGGCGGG
chr10:131738939-131738977			GGGGTGGGTGGGGAGGGAGCATGGGGTGGGCAGGGTGGG
chr10:44647571-44647613			GGGTGAGGGGGATGGGTGGGGGATGGGCAGGGTAGGGCAGGGG
chr10:86347049-86347092	2894	GRID1	GGGTGGGTGGGGCAGGGCAGGAGGGTGGGGCTGGGCAGTTAGGG

chr11:119695264-119695328	5818	NECTIN1	GGGGACTGGTGGGGAGGGTGGGGACCTGGGAGGGGTGGGAGAGGGAGATGGGAATATGGGCAGGG
chr11:63562894-63562932	54979	HRASLS2	GGGAGGGATTAGCTGGGGAGGGAGGGTCCAGGGAAAGGG
chr12:123034471-123034509	57605	PITPNM2	GGGTCGTGGGGCAAGGGTGGGTCATGGGGTGAGGGTGGG
chr12:47939348-47939389	7421	VDR	GGGTGGGGCTTGGGGGAGGTGGGTCCTGGGGTGGGGATGGGG
chr12:53499343-53499404	6895¶7786	TARBP2¶MAP3K12	GGGGGGGTGGGGGGCAGGGATGGGTCTGGGTCCTGGGATCCGGGCGTGGAGGGAGGGTCGGG
chr14:23386616-23386678	4624	MYH6	GGGAGGCCTGGGAAGGGGTGGGGCGAGGGCGGGCAGACAGGGCACAGGGCAGGGTTGAGAGGG
chr16:3957697-3957730	115	ADCY9	GGGATGGGGGTCCTGGGAGGGCAGGGCTAGGGGG
chr17:50604073-50604120	8913	CACNA1G	GGGGGTGGGGAGCAGGGTCAAGGGACAAGGGAGGGTCTGGGCTGGGGG
chr17:82698358-82698455	10966	RAB40B	GGGTGAGCGCGGGCGGAGGGCGTGCCGGGGTGCGGGCGCGGGGCCGGGGAGGGGCGCGGGGCTGGGGAGGGGGTGCGGGTGGGGGTCCGGGTCCGGGG
chr18:79395460-79395704	4772	NFATC1	GGGGGGGCGCACGGGGAGGGGGGGGCGCACGGGGAGGGGGGCGCCCGGGGAGGGGGGCGCCCGGGGAGGGGGGGCGCACGGGGAGGGGGGCGCACGGGGAGGGGCGCACGGGGAGGGGATGGGGGCGTAGGGGCGGGAACGGGGAATCCGGGGGCCGGGCAGGGGGGCGCTGGGGCTGGCGGGGGAAACGGGGGCGAACGGGCCAGACGGG
chr18:79617457-79617512			GGGGGGCGGGGTCCTGGGGGGGCAGGGTCCCGGGGAGGGCGGGGTCCCGGGGAGGG
chr19:6373235-6373294	84266	ALKBH7	GGGGTTGTGGGGCCAGGGGGGTCGGGCGCAGGGATGGGGCGGGGCCACGCTGGGGCGGGG
chr19:9851039-9851068			GGGGGTGAGGGGGTGGGTGGGAGGTGAGGG
chr20:1225680-1225711	642636	RAD21L1	GGGACCGGGGGCAGGGGGCGGGGAAGGGCGGG
chr21:40006234-40006276			GGGCCTGGGGAGAGGGAGGGCCTGGGGAGAGGGAGGGACTGGG
chr21:45287489-45287569	23275¶642852	POFUT2¶LOC642852	GGGGCAGGGGCCAGGGGGATGGGATGGAGCGGGGTCAGGGGGCAGGGGTCAGGGGAATGGGATGGGGTCAGGGTTCTGGGG
chr22:43280423-43280510	101927447¶80274	Z99756.1¶LOC101927447	GGGTGAGGGGAAGGGACGGGGATGGGTGAGGGGAAGGGACGGGAGGATGGGTGAGGGGAAGGGACAGGGGGATGGGTGAGGGGAAGGG

		UBE1	
chr22:49037526-49037614			GGGAGCGGGAGGGGCCAGGGGGTGGGACGGGGCGGGGAGAGGGAGAAGAGGGGTTCTGGGTGGGAAGGATTGGGGAGCGGGAGGAGGGG
chr3:129606851-129606910	23129	PLXND1	GGGCGGCCAGGGGCAGGCGGGGGTCCCGGGGGCGGGCGGGGCCGGGGCGGGGAGTGAGGG
chr4:88699286-88699364	891 6¶2 66812	HER C3¶ NAP1L5	GGGCAGGGTGGGGAAGAAGAGGGTGGGGCTGCGTGGGTGGGTGGGGAGGAAGAGGGTGGGGCTGGGTGGGTGGGGG
chr5:524556-524624	6550	SLC9A3	GGGCTCCGGGGGAGGGTGGGCACCAGGGAGCGCGGGGTGGGCGTGCGCGGGCGGGGCGGGCGTGCCGGG
chr6:34486149-34486233	29993	PAC SIN1	GGGGGTGAGGGGTGGAGGGACAGGGGGCCTGGGAACCCAGGGAGAGGGAGGCAGGGCCTAGGGGTGGGGTGAGGTGGGTTTGGGG
chr6:44222239-44222286	2030	SLC29A1	GGGCTGGCGGGGATGTGGGGGATGGGGGTGGGGTGGGGGAGGGTTGGG
chr7:157138531-157138583	9690	UBE3C	GGGCCGGGATGGGGTGCAGGGCAGGGTGCGGGGTGCAGGGTGGGGTGCAGGGG
chr8:142348624-142348668	203062	TSNARE1	GGGGCGGGAGCGGGGGCCAGGGGAGGGTGGGCATGGGGTGCCGGG
chr8:142464060-142464098	575	ADGRB1	GGGGGCAGGGAGGCGGGGCAAGGGTGGGATGGGAGAGGG
chr8:26383526-26383612	665	BNIP3L	GGGCGGGGCGGGGCGGGGGCGGGCCTGGGGGGCGGGGAGGCCGGGTGGGCGGAGCGGGCCGCGGAGGGGACGTGGGCCGGGATGGGG
chr9:132589622-132589662	56751	BARHL1	GGGGGGCACTGGGCTGGGGCGCCAGGGAGGGCCGGGCAGGG
		RP11-216L13.19¶C	
chr9:136800509-136800550	84960¶55684	CDC183¶ RABL6¶	GGGAAGGGCGGGGGTCAGGGGCTGGGATCTGGGAGGGGCGGG
		RP11-216L13.18	
chrX:120157742-120157823	727940	RHOXF2B	GGGGTGGGGGGAGTAGGGCGGGGGAGGGAGTAGGGCGGGGGGGCGTAGGGTGGAGGGGGGAGTAGGGCGGGGGGGCCGGGG
chr1:2640997-2641101	100287898	TTC34	GGGGCAGGGAAGCGGGGTGTGGGGAGGGCAGGGAAGGGGGGTGTGGGGAGGGCTGGGAAGGGAGGTATGGGGAGGGCTGGGAAGGGAGGTATGGGGAGGGCTGGG
chr1:29727790-29727836			GGGTGGGCTTGGGGAGGGGTGGAGGGAAGGGTGGGCTGAGGGAGGGG

chr14:103109168-103109254	91828	EXOC3L4	GGGAGGGAGACACGGGGACAGGGTGGAGAGGGAGGGAGGGAGGGAGACATGGGGACAGGGTGGCGAGGGAGAGAGGGAGACCGGGGG
chr14:103109565-103109722	91828	EXOC3L4	GGGAGACCCAGGGACAGGGTGGAGAGGGAGGGAGGGAGGGAGACATGGGGGCAGGGTGGAGAGGGAGGGAGGAGGGAGACATGGGGGCAGGGTGGAGAGGGAGGGAGGAGGGAGACATGGGG
chr16:47887333-47887374	101927132¶10050753 4	RP11-523L20.2¶LINC02133¶LINC02192	GGGTCACAGGGTCAGGGAGGGGGCTGCGGGTGGGAGGCAGGG
chr19:38390491-38390536	399473¶19972 0	SPRED3¶GGN	GGGGCATGCGGGGAGGGTAGGGACCTGGGGAGGGAGGGGAGAGGGG
chr2:129877629-129877660			GGGGAGAGGGGCGGGGGCGGGGCCGGGCCGGG
chr21:43075944-43076011	875	CBS	GGGGTGGGGAAGGGGTGGGGGGGAGGGGCCCGGGCTGGGTGGGGTGGAGGAGGGGCTGGGGGGGCGGG
chr22:44939685-44939728	112885	PHF21B	GGGGATGGGTGGGAGCAGGGCTAGGGAGGGGCGAAGGGATAGGG
chr3:50204812-50204840	10991	SLC38A3	GGGGTGGGGTGGGGGCGAGGGTGGGAGGG
chr8:10852146-10852235			GGGAGGGGAAGGGGAGGGCAGGGAAGAGAAGGGTCAGCACGGGGAAGAGAAGGGGAGCGCGGGGAAGGGAAGGGTCAGCGCGGGGAAGGG
chr9:127809879-127809969	2356	FPGS	GGGGCGGGATCTTGGGGAAGGGCGGGGCGGGGTCGTGGGGAAGGGCGGGGGCGGGCCCATGGGGAGGGGCGGGGTCGTGGGCGGGGACGGG
chr9:135757071-135757140	57582	KCNT1	GGGGTATCAGCGGGGGCGATGGAGGGTGGGGGTGGGGCCCAGCAGGGGAGGGGCAGGGTGGGGAAGAGGG
chrX:154065081-154065110	4204	MECP2	GGGGTGGGTGGGGTGGGGGCCGGGGAAGGG
chr5:181098279-181098403			GGGGGAAGGGACTGGAGGGGAGGGGAGGGGAGGGGACGGGTGGGGAGGGGTGGGGCGGGGAGGGAAAGGGTGGGGAGCAGAGGGGTGGGGAGGGGAGGGGTGGGGAGAGGTAGGGAGGGGAGGGG
chr7:76282546-76282601	222183	SRRM3	GGGGGCTGGGGGCGGGGAGGGTTCCCTGGGGCGGGGCTTAGGGCGGAGGGGCGGGG

Appendix Table A 4 Summary of Family 2 G4 sequences.

Sequence	Location	Experimental Evidence	Gene ID	Anntation
GGGGAGGGCCTGGGACAGGG	chr11:1648838-1648857	absent	387742	Intergenic
GGGAGGGCCTTGGGACAGGG	chr1:226482764-226482783	absent	375057	Intergenic
GGGAGGGGCCTGGGACAGGG	chr9:135803176-135803195	present	57582	Intergenic
GGGGGAATGGGCTGGGACAGGG	chr1:7715171-7715192	absent	23261	Intron
GGGAAGGGGGCTGGGAAAGGG	chr17:74059333-74059353	absent	6169	Intron
GGGCTGGGCATGGGACAGGG	chr14:74084190-74084209	present	4329	Promoter
GGGGAGTGGGCTGGGACAGGG	chr1:120652107-120652127	absent	101954277	Intergenic
GGGGAGTGGGCTGGGACAGGG	chr1:149272171-149272191	absent	400818	Intergenic

Appendix Table A 5 Summary of Family 3 G4 sequences.

Sequence	Location	Experimental Evidence	Gene ID	Annotation
GGGAGGGGGCTGCAGGGAGCTGGG	chr19:41700278-41700301	ppresent	1087	Iintron
GGGAGGGGGCTGCAGGGAGCTGGG	chr19:41700314-41700337	ppresent	1087	Iintron
GGGAGGGGGCTGCAGGGAGCTGGG	chr19:41700350-41700373	ppresent	1087	Iintron
GGGAGGGGGCTGCAGGGAGCTGGG	chr19:41700386-41700409	ppresent	1087	Iintron
GGGAGGGGGCTGCAGGGATGGGGG	chr12:124531250-124531273	ppresent	9612	Ppromoter
GGGAGGGGGCTGCAGGGATGGGGG	chr3:53193909-53193932	aabsent	5580	Iintergenic
GGGAGGGGCTGCAGGGATGGGGG	chr22:43426947-43426969	aabsent	758	Iintron
GGGAGGGGGAGGCAGGGTTGGGG	chr1:206041760-206041782	aabsent	440712	Iintron
GGGAGGGTGCTCCTGGGATGGGG	chr17:1312485-1312507	ppresent	286753	Iintergenic
GGGAGGGGGCTTCTGGGGTGGGG	chr3:13852428-13852450	ppresent	7476	Iintron

Appendix Table A 6 Summary of Family 4 G4 sequences

Sequence	Location	Experimental Evidence	Gene ID	Annotation
GGGCCTGGGAGGGAAGGAGAGGG	chr4:3513681-3513703	aabsent	4043	Intron
GGGCTAGGGTCGGGAGTAGAGGG	chr2:88972574-88972596	aabsent	100616399	Intron
GGGGCTGTGGAGGGAGGGAGAGGG	chr15:41893960-41893983	aabsent	51332	Promoter
GGGCTGGGGCGGGAAGGAGAGGG	chr1:121185193-121185215	aabsent	653464	Promoter
GGGCTGGGGCGGGAAGGAGAGGG	chr1:143972835-143972857	aabsent	554282	Promoter
GGGGCTGGGGCGGGAAGGAGAGGG	chr1:206203718-206203741	ppresent	729533	Promoter
GGGCAGGGCGAGGGATGGAGAGGG	chr17:39144386-39144409	aabsent	57125	Intron
GGGCATGGGCGGGGTGGAGAGGG	chr3:143313988-143314010	aabsent	100885796	Intron
GGGGTGGGGAGGGAATGTGAGGG	chr10:70886991-70887013	aabsent	5092	Promoter

Appendix Table A 7 Summary of Family 32 G4 sequences.

Sequence	Location	Experimental Evidence	Gene ID	Annotation
GGGAAGGGGAAGGGACAGGG	chr1:1136393-1136412	present	254099	Promoter
GGGGTGGGGTGGGGAGAGGG	chr1:161197724-161197743	present	4720	Promoter
GGGCTGGGGTTGGGGCTGGG	chr1:201275493-201275512	present	5317	Intergenic
GGGCTGGGGCTGGGGCAGGG	chr1:229251426-229251445	absent	5867	3' UTR
GGGGTGGGGTGGGGGATGGG	chr1:3110907-3110926	present	63976	Intron
GGGCTGTGGGCGGGGCTAGGG	chr1:37735516-37735536	present	284656	Promoter
GGGCTGGGAGAGGGCCTGGG	chr1:54138165-54138184	absent	200008	Intron
GGGATGGGCATGGGGGAGGG	chr10:124644752-124644771	present	64077	Intron
GGGGTGGGGGTGGGGTTGGG	chr10:130945693-130945712	present	100422867	Intergenic
GGGATGGGCTGGGGGCTGGG	chr10:27715853-27715872	present	283078	Intron
GGGGTGGGGTGGGGGCAGGG	chr10:69897023-69897042	present	1305	Intron
GGGATGGGGCATGGGGAGGG	chr10:79375641-79375660	present	10105	Intergenic
GGGCAGGGGGTGGGGCAGGG	chr11:404765-404784	present	11187	Promoter
GGGCAGGGGGAGGGGGAGGG	chr11:64934344-64934363	absent	170589	Promoter
GGGGTGAGGGTGGGGGTGGG	chr12:129184860-129184879	present	101927735	Intron
GGGGTGGGGGATGGGGAGGG	chr12:51905879-51905898	present	94	Promoter
GGGGTGGGGGAGGGGCAGGG	chr13:111296247-111296266	present	8874	Intron
GGGCAGGGCTGGGTGGAGGG	chr13:31009892-31009911	absent	122046	Intergenic
GGGCAGGGGTTGGGTGAGGG	chr14:24147068-24147087	present	5721	Promoter
GGGTTGGGGGCGGGGGTGGG	chr14:65413539-65413558	present	2530	Promoter
GGGGTGTGGGTGGGGCAGGG	chr14:95332851-95332870	present	101929080	Promoter
GGGGAGGGGTGGGGACCGGG	chr14:96219408-96219427	present	623	Intron
GGGATGGGGAGGGGACAGGG	chr15:77583296-77583315	absent	84894	Intron
GGGCTGGGGAGGGGACAGGG	chr15:85258124-85258143	absent	11214	Intergenic
GGGGAGGGCTGGGACCAGGG	chr15:90884045-90884064	present	2242	Promoter
GGGCTGGGACAGGGCCAGGG	chr16:32301164-32301183	absent	729264	Intergenic
GGGCAGGGCTGGGTCCAGGG	chr16:32302937-32302956	absent	729264	Intergenic
GGGCTGGGACAGGGCCAGGG	chr16:33507434-33507453	absent	24150	Intergenic
GGGCAGGGCTGGGTCCAGGG	chr16:33509201-33509220	absent	24150	Intergenic
GGGGAGGGCATGGGGCAGGG	chr16:46623763-46623782	present	79801	Promoter
GGGGTGGGGTTGGGGGAGGG	chr16:54970795-54970814	present	10265	Intergenic
GGGCAGGGCTGGGAGAAGGG	chr17:10232070-10232089	absent	8522	Intergenic
GGGAAGGGGAGGGGTCTGGG	chr17:58116701-58116720	absent	140735	Intergenic
GGGCTGGGCCTGGGCCTGGG	chr17:61402103-61402122	absent	6909	Promoter
GGGCGGGGCTGGGTCCTGGG	chr17:76079074-76079093	present	353174	Promoter
GGGGTGGGGGTGGGGATGGG	chr17:77413847-77413866	present	10801	Intron
GGGCTGGGGGAGGGGGTGGG	chr17:82435883-82435902	present	284004	Promoter
GGGGTGGGGTGGGTGGAGGG	chr17:8743246-8743265	absent	146849	Promoter
GGGCAGGGGCCTGGGGAGGG	chr18:10168387-10168406	present	9218	Intergenic
GGGGGGGGGGTGGGGGTGGG	chr18:22548091-22548110	present	64693	Intergenic
GGGCCCAGGGTGGGGCAGGG	chr18:37415961-37415980	present	56853	Intron
GGGCCGGGGCTCTGGGCGGG	chr18:62596252-62596271	absent	54877	Intergenic
GGGGTGGGATGGGGGCTGGG	chr19:17101093-17101112	present	4650	Promoter
GGGCAGGTGGGTGGGCAGGG	chr19:42351769-42351788	absent	102465875	Promoter
GGGAGGGGCGAGGGCCAGGG	chr19:51421895-51421914	present	89790	Promoter
GGGTTGGGGGTGGGGGAGGG	chr2:10264610-10264629	present	3241	Intergenic
GGGAAGGGGGTGGGAGAGGG	chr2:205734541-205734560	absent	8828	Intron
GGGCAGGGACATGGGGTGGG	chr2:233976450-233976469	present	79054	Promoter
GGGGTGAGGGTGGGGGTGGG	chr20:32612791-32612810	present	149950	Intergenic
GGGTTGGGGGAGGGGGTGGG	chr20:32980815-32980834	present	140732	Downstream
GGGGTGGGACAGTGGGAGGG	chr21:40286094-40286113	present	1826	Intron
GGGTGGGGCTAGGGCCAGGG	chr22:21613160-21613179	absent	150223	Intron
GGGGTGGGAGTGAGGGTGGG	chr3:11222231-11222250	present	3269	Promoter
GGGCTGGGGCTGGGCCAGGG	chr3:125190497-125190516	absent	84561	Promoter
GGGCTGGGGCAGGGGCCGGG	chr3:127823053-127823072	absent	11343	Promoter
GGGGAGGGCATGGGGCAGGG	chr3:129088461-129088480	absent	2815	3' UTR
GGGCTGGGGGAGAGGGTGGG	chr3:129348117-129348136	present	339942	Intergenic
GGGGGGGGGTGGGGGCAGGG	chr3:131466690-131466709	present	11222	Promoter
GGGAGGGGCTGGGGCCTGGG	chr3:13628839-13628858	absent	2199	Promoter
GGGCAGGGCTCGGGACAGGG	chr3:42732393-42732412	absent	100874114	Promoter
GGGTAGGGAAAGGGAAAGGG	chr3:45979832-45979851	absent	79443	Intron
GGGGTGGGGTAGGGGGAGGG	chr3:49489904-49489923	present	1605	Promoter
GGGGTGGGGTGGGGGATGGG	chr3:49642027-49642046	present	8927	Promoter
GGGGTGGGTGCGGGGCAGGG	chr4:102826891-102826910	present	7323	Promoter
GGGGTGGGGAGGGGGCTGGG	chr4:141929124-141929143	present	3600	Intergenic

GGGCAGGGGAATTGGGTGGG	chr4:152304991-152305010	absent	55294	Intergenic
GGGCAGGGGTGGTGGGTGGG	chr5:113656482-113656501	absent	64848	Intergenic
GGGGTGGGGCTTGGGGAGGG	chr5:134151255-134151274	present	6932	3' UTR
GGGGTGTGGGCGGGGCAGGG	chr5:151772043-151772062	present	10146	Promoter
GGGCTGGGGCTAGGGGCGGG	chr5:168410001-168410020	present	23286	Promoter
GGGCAGGGGCAGGGTGAGGG	chr6:157921891-157921910	absent	51429	Intron
GGGCTGGGGCAGGGGGAGGG	chr6:166558569-166558588	absent	6196	Intron
GGGCAGGGGGTGGGGGAGGG	chr7:128832311-128832330	absent	2318	Promoter
GGGAGGGGCTGGGGGCTGGG	chr7:30915933-30915952	absent	358	Promoter
GGGAGGGGCCGGGAGCTGGG	chr7:74454399-74454418	absent	9569	Promoter
GGGGTGGGGGCGGGGGAGGG	chr8:113439774-113439793	present	114788	Promoter
GGGCAGGGGTGGGGGGAGGG	chr8:127255311-127255330	present	100507056	Intergenic
GGGAGGGGTTGGGGGCTGGG	chr8:144496661-144496680	present	84988	Promoter
GGGCTGTGGGCGGGCCAGGG	chr8:144502891-144502910	absent	2875	Promoter
GGGGTGGGGGAGGGGTTGGG	chr8:25517201-25517220	present	157313	Intergenic
GGGGTGGGAGTAGGGGAGGG	chr8:38451245-38451264	present	2260	Intron
GGGGTTGGGGTGGGGGAGGG	chr8:54466113-54466132	present	64321	Intergenic
GGGCTGGGGTTGGGGAGGGG	chr8:99974271-99974290	present	26166	Promoter
GGGATGGGCTTGGGCCTGGG	chr9:113536716-113536735	absent	5998	Promoter
GGGCAGAGGGTGGGGCAGGG	chr9:121763670-121763689	absent	153090	Promoter
GGGGTTGGGTGGGGGCTGGG	chr9:124506593-124506612	present	2516	Promoter
GGGCAGGGGACGGGGGTGGG	chr9:27312142-27312161	absent	54586	Intergenic
GGGCTGGGGTCGGGGTGGGG	chr9:84670271-84670290	present	4915	Promoter
GGGATGGGGAGGGAACAGGG	chrX:18891734-18891753	present	100132163	Promoter
GGGCTGGGGCAGGGATAGGG	chrX:19081377-19081396	absent	10149	Promoter

Appendix Table A 8 Summary of Family 75 G4 sequences

Sequence	Location	Experimental Evidence	Gene ID	Annotation
GGGGGAGGGAGGGCCTGGG	chr11:19817544-19817562	present	89797	Intron
GGGGGTGGGAGGGGCAGGG	chr11:65967601-65967619	present	9092	Intron
GGGTGGAGGGAGGGCTGGG	chr12:130663161-130663179	absent	23504	Intron
GGGGGTGGGGGGGCCTGGG	chr12:56096956-56096974	present	2065	Promoter
GGGGGTGGGAGGGCAGGG	chr16:78000971-78000988	present	10143	Intergenic
GGGGTGGGAGGGGCATGGG	chr17:41888505-41888523	present	47	Intron
GGGGGTGGGAGGGCATGGG	chr19:52643765-52643783	present	55769	Intron
GGGGGTGGGAGGGCATGGG	chr19:52700738-52700756	present	55769	Downstream
GGGGGTGGGAGGGCATGGG	chr19:52762022-52762040	present	162966	Downstream
GGGGGTGGGAGGGCATGGG	chr19:52818986-52819004	present	7576	Promoter
GGGGGTGGGAGGGCATGGG	chr19:52855275-52855293	present	7576	Promoter
GGGGGTGGGAGGGCATGGG	chr19:52938557-52938575	present	388559	Intron
GGGGGTGGGAGGGCACGGG	chr19:53129043-53129061	present	55786	Promoter
GGGTGGTGGGAGGGATGGG	chr1:18332823-18332841	absent	84966	Intron
GGGGGTGGGAGGGCCTGGG	chr20:63293908-63293926	present	57642	Promoter
GGGGGTGGGTAGGGCCGGG	chr2:219060082-219060100	present	3549	Promoter
GGGGAGAGGGAGGGCCGGG	chr4:40244011-40244029	absent	399	3' UTR
GGGGGAGGGAGGGCTTGGG	chr8:124855332-124855350	present	157381	Intron

Appendix Table A 9 Summary of Family 80 G4 sequences

Sequence	Location	Experimental Evidence	Gene ID	Distance To TSS	annot
GGGGCGGGCTGGGGGCGGGG	chr11:67630369-67630388	present	254552	-439	Promoter
GGGTCGGGGCCGGGGGAGGG	chr11:968778-968797	present	161	-16587	Intron
GGGGCGGGCCTCGGGGGCGGGG	chr14:93184925-93184946	present	64112	0	Promoter
GGGTGCGGGGCCGGGGGGAGGGG	chr17:39927095-39927117	absent	94103	112	Promoter
GGGGCTGGGGGCGGGGGCGGG	chr17:79836273-79836293	present	8535	1588	Promoter
GGGGCGGGGCCGGGGGCGGG	chr19:18097748-18097767	present	23031	-26	Promoter
GGGGCGGGTCGTGGGCGGGG	chr19:2096722-2096741	present	126308	-49	Promoter
GGGGCTGGGTCGGGGGCGGGG	chr19:55081615-55081635	present	54869	-57	Promoter
GGGGCGGGCCCAGGGGCGGG	chr1:16700286-16700305	present	100500876	19031	Intergenic
GGGGCGGGCCGGGGGGAGGGG	chr1:185411921-185411941	present	100288079	-76882	Intergenic
GGGGCGGGCCGGGGGCGGGG	chr20:5001518-5001537	present	9962	-19	Promoter
GGGGGCGGGCTCGGGGGCGGGG	chr21:5022922-5022943	present	23308	389	Promoter
GGGGCGGGGCACGGGGGAGGG	chr4:1011532-1011552	present	53834	-270	Promoter
GGGGCGGGACCGGGGAGAGGGG	chr6:11043967-11043988	present	100506409	207	Promoter
GGGGGGGGTAGTGGGCGGGG	chr7:156169292-156169311	present	389602	206660	Intergenic
GGGGCGGGCCGTGGGCCGGG	chr7:158829641-158829660	absent	57488	-13	Promoter
GGGGCAGGCCGGGCGGGAGGGG	chr7:20798511-20798532	absent	221833	-11625	Intergenic
GGGGCGGGGGCCGCGGGCGGG	chr7:6374902-6374922	absent	5879	339	Promoter
GGGGCCGGGGCCGGGGCCGGG	chr8:22049068-22049088	absent	2039	-61	Promoter
GGGGCGGGCTCGGGGGCGGGG	chr8:66962618-66962638	present	100129654	-28	Promoter
GGGGCCGGGCCGAGGGGCGGG	chr9:132241390-132241410	present	84628	-31	Promoter

Appendix Table A 10 Enriched GO:BP categories for Family 4.

GO Term	GO Term Name	P Value	ADJ P Value
GO:0021815	modulation of microtubule cytoskeleton involved in cerebral cortex radial glia guided migration	2.13E-05	2.13E-05
GO:0021816	extension of a leading process involved in cell motility in cerebral cortex radial glia guided migration	2.13E-05	2.13E-05
GO:0021814	cell motility involved in cerebral cortex radial glia guided migration	8.51E-05	8.51E-05
GO:0022030	telencephalon glial cell migration	0.000382	0.000382
GO:0021801	cerebral cortex radial glia-guided migration	0.000382	0.000382
GO:0021799	cerebral cortex radially oriented cell migration	0.000389	0.000389
GO:0031269	pseudopodium assembly	0.000552	0.000552
GO:0031268	pseudopodium organization	0.000557	0.000557
GO:0021795	cerebral cortex cell migration	0.00059	0.00059
GO:1904861	excitatory synapse assembly	0.00059	0.00059
GO:1904862	inhibitory synapse assembly	0.00059	0.00059
GO:0022029	telencephalon cell migration	0.000893	0.000893
GO:0021885	forebrain cell migration	0.000899	0.000899
GO:0008347	glial cell migration	0.000983	0.000983
GO:2001222	regulation of neuron migration	0.001145	0.001145
GO:0021987	cerebral cortex development	0.003356	0.003356
GO:0046847	filopodium assembly	0.004232	0.004232
GO:0060996	dendritic spine development	0.004705	0.004705
GO:0021543	pallium development	0.00474	0.00474
GO:0001764	neuron migration	0.007448	0.007448
GO:0021537	telencephalon development	0.009569	0.009569
GO:0007416	synapse assembly	0.0142	0.0142
GO:0030900	forebrain development	0.017572	0.017572
GO:0016358	dendrite development	0.017572	0.017572
GO:0042063	gliogenesis	0.017572	0.017572
GO:0030336	negative regulation of cell migration	0.031913	0.031913
GO:2000146	negative regulation of cell motility	0.034153	0.034153
GO:0050808	synapse organization	0.035988	0.035988
GO:0040013	negative regulation of locomotion	0.035988	0.035988
GO:0034329	cell junction assembly	0.046932	0.046932

Appendix Table A 11 Enriched GO: BP categories for Family 32.

GO ID	GO Name	P Value	ADJ P Value
GO:0031346	positive regulation of cell projection organization	0.017646	0.017646
GO:0035378	carbon dioxide transmembrane transport	0.017646	0.017646
GO:0032989	cellular component morphogenesis	0.023881	0.023881
GO:0048842	positive regulation of axon extension involved in axon guidance	0.023881	0.023881
GO:0048858	cell projection morphogenesis	0.023881	0.023881
GO:0003097	renal water transport	0.023881	0.023881
GO:0051130	positive regulation of cellular component organization	0.023881	0.023881
GO:0051239	regulation of multicellular organismal process	0.023881	0.023881
GO:0048846	axon extension involved in axon guidance	0.023881	0.023881
GO:0120039	plasma membrane bounded cell projection morphogenesis	0.023881	0.023881
GO:1902284	neuron projection extension involved in neuron projection guidance	0.023881	0.023881
GO:1903955	positive regulation of protein targeting to mitochondrion	0.023881	0.023881
GO:0032990	cell part morphogenesis	0.024503	0.024503
GO:1903749	positive regulation of establishment of protein localization to mitochondrion	0.025365	0.025365
GO:0097485	neuron projection guidance	0.029584	0.029584
GO:0007411	axon guidance	0.029584	0.029584
GO:1903214	regulation of protein targeting to mitochondrion	0.032099	0.032099
GO:0007167	enzyme-linked receptor protein signaling pathway	0.032099	0.032099
GO:0048518	positive regulation of biological process	0.032578	0.032578
GO:0051094	positive regulation of developmental process	0.033021	0.033021
GO:0050772	positive regulation of axonogenesis	0.03945	0.03945
GO:0048468	cell development	0.03945	0.03945
GO:1903747	regulation of establishment of protein localization to mitochondrion	0.03945	0.03945
GO:0007409	axonogenesis	0.03945	0.03945
GO:0022603	regulation of anatomical structure morphogenesis	0.039943	0.039943
GO:0048812	neuron projection morphogenesis	0.045216	0.045216
GO:0000902	cell morphogenesis	0.045216	0.045216
GO:0008361	regulation of cell size	0.045216	0.045216
GO:0051347	positive regulation of transferase activity	0.045887	0.045887
GO:0061564	axon development	0.045887	0.045887
GO:0090066	regulation of anatomical structure size	0.046587	0.046587
GO:0120035	regulation of plasma membrane bounded cell projection organization	0.046587	0.046587

Appendix Table A 12 Enriched GO:BP categories for Family 75.

GO ID	GO Name	P Value	ADJ P Value
GO:0045582	positive regulation of T cell differentiation	0.015898	0.015898
GO:0045621	positive regulation of lymphocyte differentiation	0.015898	0.015898
GO:1903708	positive regulation of hemopoiesis	0.016146	0.016146
GO:1902107	positive regulation of leukocyte differentiation	0.016146	0.016146
GO:0045580	regulation of T cell differentiation	0.016146	0.016146
GO:0045619	regulation of lymphocyte differentiation	0.019728	0.019728
GO:0050870	positive regulation of T cell activation	0.03222	0.03222
GO:0030217	T cell differentiation	0.03222	0.03222
GO:1903039	positive regulation of leukocyte cell-cell adhesion	0.03674	0.03674
GO:1902105	regulation of leukocyte differentiation	0.03674	0.03674
GO:0022409	positive regulation of cell-cell adhesion	0.047638	0.047638
GO:1903037	regulation of leukocyte cell-cell adhesion	0.04765	0.04765
GO:0030155	regulation of cell adhesion	0.04765	0.04765
GO:0030098	lymphocyte differentiation	0.04765	0.04765
GO:0050863	regulation of T cell activation	0.04765	0.04765
GO:1903706	regulation of hemopoiesis	0.049277	0.049277

Appendix Table A 13 Enriched GO:BP categories for Family 80

GO ID	GO Name	P Value	ADJ P Value
GO:0098562	cytoplasmic side of membrane	1.57E-02	1.57E-02
GO:0005886	plasma membrane	1.57E-02	1.57E-02
GO:0009898	cytoplasmic side of plasma membrane	1.57E-02	1.57E-02
GO:0098590	plasma membrane region	1.57E-02	1.57E-02
GO:0031253	cell projection membrane	1.81E-02	1.81E-02
GO:0101003	ficolin-1-rich granule membrane	1.86E-02	1.86E-02
GO:0098552	side of membrane	1.86E-02	1.86E-02
GO:0071944	cell periphery	1.86E-02	1.86E-02
GO:0032587	ruffle membrane	1.86E-02	1.86E-02
GO:0030667	secretory granule membrane	1.93E-02	1.93E-02
GO:0031256	leading edge membrane	3.64E-02	3.64E-02
GO:0005884	actin filament	3.64E-02	3.64E-02
GO:0031234	extrinsic component of cytoplasmic side of plasma membrane	3.64E-02	3.64E-02
GO:0001726	ruffle	3.85E-02	3.85E-02
GO:0016020	membrane	3.85E-02	3.85E-02
GO:0019897	extrinsic component of plasma membrane	4.29E-02	4.29E-02
GO:0031224	intrinsic component of membrane	4.29E-02	4.29E-02
GO:0031227	intrinsic component of endoplasmic reticulum membrane	4.29E-02	4.29E-02
GO:0098797	plasma membrane protein complex	4.29E-02	4.29E-02
GO:0101002	ficolin-1-rich granule	4.29E-02	4.29E-02
GO:0070820	tertiary granule	4.29E-02	4.29E-02

Appendix Table A 14 Enriched GO:BP categories for experimentally validated G4s overlapping enhancers, group 1.

GO ID	GO Name	P Value	adjustedPValue
GO:0035556	intracellular signal transduction	2.28E-06	2.28E-06
GO:0010033	response to organic substance	2.36E-06	2.36E-06
GO:0007165	signal transduction	3.86E-06	3.86E-06
GO:1902531	regulation of intracellular signal transduction	3.86E-06	3.86E-06
GO:0009966	regulation of signal transduction	3.86E-06	3.86E-06
GO:0050896	response to stimulus	6.92E-06	6.92E-06
GO:0048584	positive regulation of response to stimulus	6.92E-06	6.92E-06
GO:0007166	cell surface receptor signaling pathway	2.89E-05	2.89E-05
GO:0007154	cell communication	2.89E-05	2.89E-05
GO:0010646	regulation of cell communication	6.71E-05	6.71E-05
GO:1902533	positive regulation of intracellular signal transduction	6.71E-05	6.71E-05
GO:0023051	regulation of signaling	6.71E-05	6.71E-05
GO:0023052	signaling	6.71E-05	6.71E-05
GO:0034097	response to cytokine	6.96E-05	6.96E-05
GO:0051716	cellular response to stimulus	7.39E-05	7.39E-05
GO:0009967	positive regulation of signal transduction	7.42E-05	7.42E-05
GO:0009615	response to virus	1.18E-04	1.18E-04
GO:0010647	positive regulation of cell communication	2.19E-04	2.19E-04
GO:0032101	regulation of response to external stimulus	2.29E-04	2.29E-04
GO:0023056	positive regulation of signaling	2.29E-04	2.29E-04
GO:0071310	cellular response to organic substance	2.35E-04	2.35E-04
GO:0019221	cytokine-mediated signaling pathway	2.36E-04	2.36E-04
GO:0042221	response to chemical	2.36E-04	2.36E-04
GO:0012501	programmed cell death	2.36E-04	2.36E-04
GO:0006915	apoptotic process	2.37E-04	2.37E-04
GO:0009605	response to external stimulus	2.38E-04	2.38E-04
GO:0044419	biological process involved in interspecies interaction between organisms	2.71E-04	2.71E-04
GO:0097190	apoptotic signaling pathway	3.78E-04	3.78E-04
GO:0070887	cellular response to chemical stimulus	3.98E-04	3.98E-04
GO:0051607	defense response to virus	3.98E-04	3.98E-04
GO:0140546	defense response to symbiont	3.98E-04	3.98E-04
GO:0007249	I-kappaB kinase/NF-kappaB signaling	3.98E-04	3.98E-04
GO:1904747	positive regulation of apoptotic process involved in development	6.93E-04	6.93E-04
GO:1902339	positive regulation of apoptotic process involved in morphogenesis	6.93E-04	6.93E-04
GO:0009753	response to jasmonic acid	6.93E-04	6.93E-04
GO:0071395	cellular response to jasmonic acid stimulus	6.93E-04	6.93E-04
GO:0031347	regulation of defense response	6.93E-04	6.93E-04
GO:0071345	cellular response to cytokine stimulus	6.94E-04	6.94E-04
GO:0048518	positive regulation of biological process	1.03E-03	1.03E-03
GO:0051055	negative regulation of lipid biosynthetic process	1.30E-03	1.30E-03
GO:0008219	cell death	1.31E-03	1.31E-03
GO:0080134	regulation of response to stress	1.31E-03	1.31E-03
GO:0070542	response to fatty acid	1.35E-03	1.35E-03
GO:1901798	positive regulation of signal transduction by p53 class mediator	1.35E-03	1.35E-03
GO:0016032	viral process	1.42E-03	1.42E-03
GO:0002376	immune system process	1.55E-03	1.55E-03
GO:0043122	regulation of I-kappaB kinase/NF-kappaB signaling	1.99E-03	1.99E-03
GO:0043207	response to external biotic stimulus	2.34E-03	2.34E-03
GO:0051707	response to other organism	2.34E-03	2.34E-03
GO:0038061	NIK/NF-kappaB signaling	2.34E-03	2.34E-03
GO:0043067	regulation of programmed cell death	3.17E-03	3.17E-03
GO:1903829	positive regulation of protein localization	3.17E-03	3.17E-03
GO:0008630	intrinsic apoptotic signaling pathway in response to DNA damage	3.50E-03	3.50E-03
GO:1901222	regulation of NIK/NF-kappaB signaling	3.50E-03	3.50E-03
GO:0045071	negative regulation of viral genome replication	3.50E-03	3.50E-03
GO:0033209	tumor necrosis factor-mediated signaling pathway	3.65E-03	3.65E-03
GO:0071398	cellular response to fatty acid	3.65E-03	3.65E-03
GO:1902337	regulation of apoptotic process involved in morphogenesis	3.68E-03	3.68E-03

GO:0009607	response to biotic stimulus	3.68E-03	3.68E-03
GO:1904748	regulation of apoptotic process involved in development	3.68E-03	3.68E-03
GO:0045833	negative regulation of lipid metabolic process	3.70E-03	3.70E-03
GO:0016601	Rac protein signal transduction	4.14E-03	4.14E-03
GO:0072331	signal transduction by p53 class mediator	4.42E-03	4.42E-03
GO:0042981	regulation of apoptotic process	4.42E-03	4.42E-03
GO:0006952	defense response	4.49E-03	4.49E-03
GO:0048522	positive regulation of cellular process	4.54E-03	4.54E-03
GO:0071677	positive regulation of mononuclear cell migration	4.73E-03	4.73E-03
GO:0050789	regulation of biological process	5.14E-03	5.14E-03
GO:0048525	negative regulation of viral process	5.14E-03	5.14E-03
GO:0032502	developmental process	5.42E-03	5.42E-03
GO:0050793	regulation of developmental process	5.89E-03	5.89E-03
GO:0008625	extrinsic apoptotic signaling pathway via death domain receptors	6.34E-03	6.34E-03
GO:1902644	tertiary alcohol metabolic process	6.59E-03	6.59E-03
GO:0030647	aminoglycoside antibiotic metabolic process	6.61E-03	6.61E-03
GO:0070383	DNA cytosine deamination	6.61E-03	6.61E-03
GO:0044597	daunorubicin metabolic process	6.61E-03	6.61E-03
GO:0097193	intrinsic apoptotic signaling pathway	6.61E-03	6.61E-03
GO:0060561	apoptotic process involved in morphogenesis	6.61E-03	6.61E-03
GO:0070278	extracellular matrix constituent secretion	6.61E-03	6.61E-03
GO:0032103	positive regulation of response to external stimulus	6.71E-03	6.71E-03
GO:0071356	cellular response to tumor necrosis factor	7.37E-03	7.37E-03
GO:0043065	positive regulation of apoptotic process	7.88E-03	7.88E-03
GO:0046890	regulation of lipid biosynthetic process	7.88E-03	7.88E-03
GO:0031327	negative regulation of cellular biosynthetic process	8.12E-03	8.12E-03
GO:0030638	polyketide metabolic process	8.23E-03	8.23E-03
GO:0010648	negative regulation of cell communication	8.23E-03	8.23E-03
GO:0044598	doxorubicin metabolic process	8.23E-03	8.23E-03
GO:0016554	cytidine to uridine editing	8.23E-03	8.23E-03
GO:0006950	response to stress	8.23E-03	8.23E-03
GO:0032880	regulation of protein localization	8.23E-03	8.23E-03
GO:0048856	anatomical structure development	8.36E-03	8.36E-03
GO:0097191	extrinsic apoptotic signaling pathway	8.54E-03	8.54E-03
GO:0044249	cellular biosynthetic process	8.55E-03	8.55E-03
GO:0023057	negative regulation of signaling	8.55E-03	8.55E-03
GO:0006954	inflammatory response	9.07E-03	9.07E-03
GO:0009890	negative regulation of biosynthetic process	9.62E-03	9.62E-03
GO:0043068	positive regulation of programmed cell death	9.62E-03	9.62E-03
GO:0002831	regulation of response to biotic stimulus	9.90E-03	9.90E-03
GO:0048523	negative regulation of cellular process	1.02E-02	1.02E-02
GO:0009893	positive regulation of metabolic process	1.02E-02	1.02E-02
GO:0030865	cortical cytoskeleton organization	1.03E-02	1.03E-02
GO:0002682	regulation of immune system process	1.07E-02	1.07E-02
GO:0010941	regulation of cell death	1.09E-02	1.09E-02
GO:1901576	organic substance biosynthetic process	1.09E-02	1.09E-02
GO:0051896	regulation of protein kinase B signaling	1.09E-02	1.09E-02
GO:0034612	response to tumor necrosis factor	1.09E-02	1.09E-02
GO:0032102	negative regulation of response to external stimulus	1.12E-02	1.12E-02
GO:0002468	dendritic cell antigen processing and presentation	1.24E-02	1.24E-02
GO:0019216	regulation of lipid metabolic process	1.31E-02	1.31E-02
GO:0048585	negative regulation of response to stimulus	1.31E-02	1.31E-02
GO:0045892	negative regulation of DNA-templated transcription	1.36E-02	1.36E-02
GO:0002484	antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway	1.39E-02	1.39E-02
GO:0002486	antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway, TAP-independent	1.39E-02	1.39E-02
GO:0071798	response to prostaglandin D	1.39E-02	1.39E-02
GO:1902679	negative regulation of RNA biosynthetic process	1.39E-02	1.39E-02
GO:1903507	negative regulation of nucleic acid-templated transcription	1.39E-02	1.39E-02
GO:0048869	cellular developmental process	1.39E-02	1.39E-02
GO:0071799	cellular response to prostaglandin D stimulus	1.39E-02	1.39E-02
GO:0045006	DNA deamination	1.40E-02	1.40E-02
GO:2000010	positive regulation of protein localization to cell surface	1.40E-02	1.40E-02
GO:1900025	negative regulation of substrate adhesion-dependent cell spreading	1.40E-02	1.40E-02
GO:0042448	progesterone metabolic process	1.40E-02	1.40E-02

GO:0010771	negative regulation of cell morphogenesis involved in differentiation	1.40E-02	1.40E-02
GO:0032879	regulation of localization	1.41E-02	1.41E-02
GO:0009058	biosynthetic process	1.41E-02	1.41E-02
GO:0045069	regulation of viral genome replication	1.42E-02	1.42E-02
GO:1901700	response to oxygen-containing compound	1.43E-02	1.43E-02
GO:0008207	C21-steroid hormone metabolic process	1.47E-02	1.47E-02
GO:0009649	entrainment of circadian clock	1.47E-02	1.47E-02
GO:0031341	regulation of cell killing	1.47E-02	1.47E-02
GO:0045732	positive regulation of protein catabolic process	1.48E-02	1.48E-02
GO:0046649	lymphocyte activation	1.48E-02	1.48E-02
GO:0019079	viral genome replication	1.49E-02	1.49E-02
GO:0050792	regulation of viral process	1.56E-02	1.56E-02
GO:0065007	biological regulation	1.59E-02	1.59E-02
GO:0002687	positive regulation of leukocyte migration	1.61E-02	1.61E-02
GO:0043542	endothelial cell migration	1.66E-02	1.66E-02
GO:0048519	negative regulation of biological process	1.76E-02	1.76E-02
GO:0031349	positive regulation of defense response	1.76E-02	1.76E-02
GO:0019058	viral life cycle	1.76E-02	1.76E-02
GO:0030154	cell differentiation	1.78E-02	1.78E-02
GO:0043491	protein kinase B signaling	1.82E-02	1.82E-02
GO:0045869	negative regulation of single stranded viral RNA replication via double stranded DNA intermediate	1.87E-02	1.87E-02
GO:0034694	response to prostaglandin	1.87E-02	1.87E-02
GO:0030036	actin cytoskeleton organization	1.94E-02	1.94E-02
GO:0051253	negative regulation of RNA metabolic process	1.95E-02	1.95E-02
GO:1904377	positive regulation of protein localization to cell periphery	1.99E-02	1.99E-02
GO:0001775	cell activation	1.99E-02	1.99E-02
GO:0050688	regulation of defense response to virus	1.99E-02	1.99E-02
GO:0019882	antigen processing and presentation	2.08E-02	2.08E-02
GO:0031326	regulation of cellular biosynthetic process	2.09E-02	2.09E-02
GO:0061044	negative regulation of vascular wound healing	2.10E-02	2.10E-02
GO:0071072	negative regulation of phospholipid biosynthetic process	2.10E-02	2.10E-02
GO:0003332	negative regulation of extracellular matrix constituent secretion	2.10E-02	2.10E-02
GO:0010604	positive regulation of macromolecule metabolic process	2.10E-02	2.10E-02
GO:0010942	positive regulation of cell death	2.16E-02	2.16E-02
GO:1900744	regulation of p38MAPK cascade	2.42E-02	2.42E-02
GO:0060341	regulation of cellular localization	2.42E-02	2.42E-02
GO:0016137	glycoside metabolic process	2.46E-02	2.46E-02
GO:0006955	immune response	2.46E-02	2.46E-02
GO:0051173	positive regulation of nitrogen compound metabolic process	2.46E-02	2.46E-02
GO:0051701	biological process involved in interaction with host	2.51E-02	2.51E-02
GO:0050691	regulation of defense response to virus by host	2.58E-02	2.58E-02
GO:0030029	actin filament-based process	2.59E-02	2.59E-02
GO:0009889	regulation of biosynthetic process	2.59E-02	2.59E-02
GO:0010558	negative regulation of macromolecule biosynthetic process	2.59E-02	2.59E-02
GO:0051897	positive regulation of protein kinase B signaling	2.60E-02	2.60E-02
GO:0045595	regulation of cell differentiation	2.60E-02	2.60E-02
GO:0009968	negative regulation of signal transduction	2.61E-02	2.61E-02
GO:1903900	regulation of viral life cycle	2.62E-02	2.62E-02
GO:0030155	regulation of cell adhesion	2.70E-02	2.70E-02
GO:0044403	biological process involved in symbiotic interaction	2.70E-02	2.70E-02
GO:0043652	engulfment of apoptotic cell	2.70E-02	2.70E-02
GO:0016553	base conversion or substitution editing	2.70E-02	2.70E-02
GO:0033036	macromolecule localization	2.70E-02	2.70E-02
GO:0002685	regulation of leukocyte migration	2.70E-02	2.70E-02
GO:0002684	positive regulation of immune system process	2.70E-02	2.70E-02
GO:0002483	antigen processing and presentation of endogenous peptide antigen	2.70E-02	2.70E-02
GO:0033993	response to lipid	2.70E-02	2.70E-02
GO:0043123	positive regulation of I-kappaB kinase/NF-kappaB signaling	2.70E-02	2.70E-02
GO:0050794	regulation of cellular process	2.70E-02	2.70E-02
GO:0009719	response to endogenous stimulus	2.70E-02	2.70E-02
GO:0045091	regulation of single stranded viral RNA replication via double stranded DNA intermediate	2.70E-02	2.70E-02
GO:0071222	cellular response to lipopolysaccharide	2.73E-02	2.73E-02

GO:0019222	regulation of metabolic process	2.73E-02	2.73E-02
GO:0043277	apoptotic cell clearance	2.73E-02	2.73E-02
GO:0051247	positive regulation of protein metabolic process	2.73E-02	2.73E-02
GO:0008610	lipid biosynthetic process	2.73E-02	2.73E-02
GO:0071675	regulation of mononuclear cell migration	2.81E-02	2.81E-02
GO:0051093	negative regulation of developmental process	2.81E-02	2.81E-02
GO:1903726	negative regulation of phospholipid metabolic process	2.81E-02	2.81E-02
GO:1904238	pericyte cell differentiation	2.81E-02	2.81E-02
GO:0001910	regulation of leukocyte mediated cytotoxicity	2.86E-02	2.86E-02
GO:0019218	regulation of steroid metabolic process	2.86E-02	2.86E-02
GO:0045934	negative regulation of nucleobase-containing compound metabolic process	2.86E-02	2.86E-02
GO:0039692	single stranded viral RNA replication via double stranded DNA intermediate	2.86E-02	2.86E-02
GO:0031324	negative regulation of cellular metabolic process	2.92E-02	2.92E-02
GO:0002697	regulation of immune effector process	3.07E-02	3.07E-02
GO:0045862	positive regulation of proteolysis	3.07E-02	3.07E-02
GO:0051234	establishment of localization	3.18E-02	3.18E-02
GO:0019748	secondary metabolic process	3.31E-02	3.31E-02
GO:2000059	negative regulation of ubiquitin-dependent protein catabolic process	3.31E-02	3.31E-02
GO:2000630	positive regulation of miRNA metabolic process	3.31E-02	3.31E-02
GO:0002819	regulation of adaptive immune response	3.31E-02	3.31E-02
GO:0045321	leukocyte activation	3.32E-02	3.32E-02
GO:1904375	regulation of protein localization to cell periphery	3.35E-02	3.35E-02
GO:0071496	cellular response to external stimulus	3.39E-02	3.39E-02
GO:0045444	fat cell differentiation	3.39E-02	3.39E-02
GO:1905475	regulation of protein localization to membrane	3.48E-02	3.48E-02
GO:0071219	cellular response to molecule of bacterial origin	3.48E-02	3.48E-02
GO:1902742	apoptotic process involved in development	3.52E-02	3.52E-02
GO:1901701	cellular response to oxygen-containing compound	3.52E-02	3.52E-02
GO:0098542	defense response to other organism	3.52E-02	3.52E-02
GO:2001141	regulation of RNA biosynthetic process	3.54E-02	3.54E-02
GO:0030178	negative regulation of Wnt signaling pathway	3.54E-02	3.54E-02
GO:0048260	positive regulation of receptor-mediated endocytosis	3.69E-02	3.69E-02
GO:1905897	regulation of response to endoplasmic reticulum stress	3.74E-02	3.74E-02
GO:0006629	lipid metabolic process	3.78E-02	3.78E-02
GO:0031328	positive regulation of cellular biosynthetic process	3.88E-02	3.88E-02
GO:0038066	p38MAPK cascade	3.89E-02	3.89E-02
GO:0045807	positive regulation of endocytosis	3.89E-02	3.89E-02
GO:0051179	localization	3.89E-02	3.89E-02
GO:0048513	animal organ development	3.97E-02	3.97E-02
GO:0062014	negative regulation of small molecule metabolic process	4.07E-02	4.07E-02
GO:0009059	macromolecule biosynthetic process	4.08E-02	4.08E-02
GO:0097435	supramolecular fiber organization	4.15E-02	4.15E-02
GO:0001667	ameboidal-type cell migration	4.15E-02	4.15E-02
GO:0001892	embryonic placenta development	4.30E-02	4.30E-02
GO:0022604	regulation of cell morphogenesis	4.36E-02	4.36E-02
GO:0030199	collagen fibril organization	4.38E-02	4.38E-02
GO:0001912	positive regulation of leukocyte mediated cytotoxicity	4.38E-02	4.38E-02
GO:0030162	regulation of proteolysis	4.38E-02	4.38E-02
GO:0050729	positive regulation of inflammatory response	4.47E-02	4.47E-02
GO:0051785	positive regulation of nuclear division	4.64E-02	4.64E-02
GO:0009891	positive regulation of biosynthetic process	4.64E-02	4.64E-02
GO:0043433	negative regulation of DNA-binding transcription factor activity	4.68E-02	4.68E-02
GO:0031652	positive regulation of heat generation	4.68E-02	4.68E-02
GO:0019883	antigen processing and presentation of endogenous antigen	4.68E-02	4.68E-02
GO:0010876	lipid localization	4.73E-02	4.73E-02
GO:0007586	digestion	4.74E-02	4.74E-02
GO:0016477	cell migration	4.74E-02	4.74E-02
GO:0051090	regulation of DNA-binding transcription factor activity	4.85E-02	4.85E-02
GO:0051246	regulation of protein metabolic process	4.85E-02	4.85E-02
GO:0090090	negative regulation of canonical Wnt signaling pathway	4.85E-02	4.85E-02
GO:1903506	regulation of nucleic acid-templated transcription	4.94E-02	4.94E-02

Appendix Table A 15 Enriched GO:BP categories for experimentally validated G4s overlapping enhancers, group 2

GO ID	GO Name	P Value	ADJ P Value
GO:0002376	immune system process	4.15E-23	4.15E-23
GO:0006955	immune response	5.99E-23	5.99E-23
GO:0002682	regulation of immune system process	2.97E-21	2.97E-21
GO:0002684	positive regulation of immune system process	7.09E-18	7.09E-18
GO:0050776	regulation of immune response	1.08E-15	1.08E-15
GO:0002764	immune response-regulating signaling pathway	8.25E-15	8.25E-15
GO:0050778	positive regulation of immune response	1.77E-13	1.77E-13
GO:0002429	immune response-activating cell surface receptor signaling pathway	1.83E-13	1.83E-13
GO:0002757	immune response-activating signal transduction	1.83E-13	1.83E-13
GO:0002768	immune response-regulating cell surface receptor signaling pathway	8.77E-13	8.77E-13
GO:0002253	activation of immune response	1.39E-12	1.39E-12
GO:0046649	lymphocyte activation	1.75E-11	1.75E-11
GO:0045321	leukocyte activation	4.23E-11	4.23E-11
GO:0048584	positive regulation of response to stimulus	6.13E-11	6.13E-11
GO:0007165	signal transduction	6.13E-11	6.13E-11
GO:0002252	immune effector process	5.25E-10	5.25E-10
GO:0001819	positive regulation of cytokine production	1.28E-09	1.28E-09
GO:0001775	cell activation	1.28E-09	1.28E-09
GO:0002250	adaptive immune response	1.30E-09	1.30E-09
GO:0006952	defense response	3.61E-09	3.61E-09
GO:0050851	antigen receptor-mediated signaling pathway	4.06E-09	4.06E-09
GO:0023052	signaling	4.37E-09	4.37E-09
GO:0007154	cell communication	5.81E-09	5.81E-09
GO:0002697	regulation of immune effector process	6.28E-09	6.28E-09
GO:0050852	T cell receptor signaling pathway	9.32E-09	9.32E-09
GO:0007166	cell surface receptor signaling pathway	9.32E-09	9.32E-09
GO:0001817	regulation of cytokine production	9.32E-09	9.32E-09
GO:0001816	cytokine production	1.03E-08	1.03E-08
GO:0002700	regulation of production of molecular mediator of immune response	2.40E-08	2.40E-08
GO:0048583	regulation of response to stimulus	2.48E-08	2.48E-08
GO:0042110	T cell activation	2.73E-08	2.73E-08
GO:0032103	positive regulation of response to external stimulus	7.15E-08	7.15E-08
GO:0050896	response to stimulus	1.27E-07	1.27E-07
GO:0002440	production of molecular mediator of immune response	2.12E-07	2.12E-07
GO:0051716	cellular response to stimulus	2.31E-07	2.31E-07
GO:0002702	positive regulation of production of molecular mediator of immune response	2.55E-07	2.55E-07
GO:0002699	positive regulation of immune effector process	3.32E-07	3.32E-07
GO:1903131	mononuclear cell differentiation	4.15E-07	4.15E-07
GO:0032101	regulation of response to external stimulus	4.74E-07	4.74E-07
GO:0030098	lymphocyte differentiation	4.95E-07	4.95E-07
GO:0031347	regulation of defense response	7.94E-07	7.94E-07
GO:0046631	alpha-beta T cell activation	1.63E-06	1.63E-06
GO:0002521	leukocyte differentiation	1.65E-06	1.65E-06
GO:0070663	regulation of leukocyte proliferation	2.73E-06	2.73E-06
GO:0006954	inflammatory response	2.82E-06	2.82E-06
GO:0031663	lipopolysaccharide-mediated signaling pathway	3.23E-06	3.23E-06
GO:0046629	gamma-delta T cell activation	3.28E-06	3.28E-06
GO:0031349	positive regulation of defense response	3.75E-06	3.75E-06
GO:0010628	positive regulation of gene expression	4.15E-06	4.15E-06
GO:1903037	regulation of leukocyte cell-cell adhesion	4.15E-06	4.15E-06
GO:0043207	response to external biotic stimulus	4.62E-06	4.62E-06
GO:0051707	response to other organism	4.62E-06	4.62E-06
GO:0051240	positive regulation of multicellular organismal process	4.87E-06	4.87E-06
GO:0098542	defense response to other organism	5.73E-06	5.73E-06
GO:0002831	regulation of response to biotic stimulus	5.76E-06	5.76E-06
GO:0050670	regulation of lymphocyte proliferation	5.76E-06	5.76E-06
GO:0019221	cytokine-mediated signaling pathway	5.76E-06	5.76E-06
GO:0032944	regulation of mononuclear cell proliferation	6.38E-06	6.38E-06
GO:1903039	positive regulation of leukocyte cell-cell adhesion	7.12E-06	7.12E-06
GO:0002718	regulation of cytokine production involved in immune response	7.48E-06	7.48E-06

GO:0002367	cytokine production involved in immune response	7.48E-06	7.48E-06
GO:0009607	response to biotic stimulus	7.59E-06	7.59E-06
GO:0032755	positive regulation of interleukin-6 production	8.88E-06	8.88E-06
GO:0032735	positive regulation of interleukin-12 production	1.12E-05	1.12E-05
GO:0045785	positive regulation of cell adhesion	1.26E-05	1.26E-05
GO:0007159	leukocyte cell-cell adhesion	1.40E-05	1.40E-05
GO:0000165	MAPK cascade	1.41E-05	1.41E-05
GO:0032675	regulation of interleukin-6 production	1.41E-05	1.41E-05
GO:0032637	interleukin-8 production	1.41E-05	1.41E-05
GO:0032635	interleukin-6 production	1.41E-05	1.41E-05
GO:0032677	regulation of interleukin-8 production	1.41E-05	1.41E-05
GO:0070661	leukocyte proliferation	1.68E-05	1.68E-05
GO:0051249	regulation of lymphocyte activation	1.91E-05	1.91E-05
GO:0002833	positive regulation of response to biotic stimulus	2.18E-05	2.18E-05
GO:0032757	positive regulation of interleukin-8 production	2.52E-05	2.52E-05
GO:0044419	biological process involved in interspecies interaction between organisms	2.79E-05	2.79E-05
GO:0046651	lymphocyte proliferation	3.25E-05	3.25E-05
GO:0097530	granulocyte migration	3.40E-05	3.40E-05
GO:0022409	positive regulation of cell-cell adhesion	3.49E-05	3.49E-05
GO:0002221	pattern recognition receptor signaling pathway	3.55E-05	3.55E-05
GO:0032943	mononuclear cell proliferation	3.56E-05	3.56E-05
GO:0070371	ERK1 and ERK2 cascade	3.65E-05	3.65E-05
GO:0050900	leukocyte migration	4.13E-05	4.13E-05
GO:0071345	cellular response to cytokine stimulus	4.32E-05	4.32E-05
GO:0070374	positive regulation of ERK1 and ERK2 cascade	4.36E-05	4.36E-05
GO:0002520	immune system development	4.94E-05	4.94E-05
GO:0030097	hemopoiesis	5.42E-05	5.42E-05
GO:0009605	response to external stimulus	6.53E-05	6.53E-05
GO:0043410	positive regulation of MAPK cascade	6.53E-05	6.53E-05
GO:0002443	leukocyte mediated immunity	6.61E-05	6.61E-05
GO:0035556	intracellular signal transduction	6.61E-05	6.61E-05
GO:0050865	regulation of cell activation	6.67E-05	6.67E-05
GO:0048534	hematopoietic or lymphoid organ development	7.00E-05	7.00E-05
GO:1990266	neutrophil migration	7.20E-05	7.20E-05
GO:0022407	regulation of cell-cell adhesion	7.20E-05	7.20E-05
GO:0002220	innate immune response activating cell surface receptor signaling pathway	7.20E-05	7.20E-05
GO:0032615	interleukin-12 production	7.75E-05	7.75E-05
GO:0032655	regulation of interleukin-12 production	7.75E-05	7.75E-05
GO:0050863	regulation of T cell activation	8.28E-05	8.28E-05
GO:0002758	innate immune response-activating signal transduction	8.30E-05	8.30E-05
GO:0032760	positive regulation of tumor necrosis factor production	8.79E-05	8.79E-05
GO:0002694	regulation of leukocyte activation	9.15E-05	9.15E-05
GO:0050870	positive regulation of T cell activation	9.63E-05	9.63E-05
GO:0070372	regulation of ERK1 and ERK2 cascade	0.000101363	0.000101363
GO:0002720	positive regulation of cytokine production involved in immune response	0.000114437	0.000114437
GO:1903557	positive regulation of tumor necrosis factor superfamily cytokine production	0.000116819	0.000116819
GO:0050764	regulation of phagocytosis	0.000125399	0.000125399
GO:0070665	positive regulation of leukocyte proliferation	0.000139526	0.000139526
GO:0043408	regulation of MAPK cascade	0.000161131	0.000161131
GO:0034097	response to cytokine	0.000169097	0.000169097
GO:0002460	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	0.000175603	0.000175603
GO:0009966	regulation of signal transduction	0.000177725	0.000177725
GO:0002683	negative regulation of immune system process	0.000178272	0.000178272
GO:0032680	regulation of tumor necrosis factor production	0.000187589	0.000187589
GO:0032640	tumor necrosis factor production	0.000187589	0.000187589
GO:0007249	I-kappaB kinase/NF-kappaB signaling	0.000187589	0.000187589
GO:0042129	regulation of T cell proliferation	0.000195896	0.000195896
GO:1902531	regulation of intracellular signal transduction	0.000201588	0.000201588
GO:0050867	positive regulation of cell activation	0.000201588	0.000201588
GO:0050766	positive regulation of phagocytosis	0.000203772	0.000203772
GO:0002703	regulation of leukocyte mediated immunity	0.000203772	0.000203772
GO:0030155	regulation of cell adhesion	0.00020608	0.00020608
GO:0001818	negative regulation of cytokine production	0.000225163	0.000225163
GO:0051251	positive regulation of lymphocyte activation	0.000225163	0.000225163

GO:0051209	release of sequestered calcium ion into cytosol	0.000225163	0.000225163
GO:0050727	regulation of inflammatory response	0.000225163	0.000225163
GO:1903555	regulation of tumor necrosis factor superfamily cytokine production	0.000225163	0.000225163
GO:0002639	positive regulation of immunoglobulin production	0.000225163	0.000225163
GO:0071706	tumor necrosis factor superfamily cytokine production	0.000225163	0.000225163
GO:0051283	negative regulation of sequestering of calcium ion	0.000233805	0.000233805
GO:0051282	regulation of sequestering of calcium ion	0.000248588	0.000248588
GO:0051235	maintenance of location	0.000289196	0.000289196
GO:0051651	maintenance of location in cell	0.000293015	0.000293015
GO:0051208	sequestering of calcium ion	0.00029746	0.00029746
GO:0042098	T cell proliferation	0.000326578	0.000326578
GO:0045059	positive thymic T cell selection	0.000326578	0.000326578
GO:0002675	positive regulation of acute inflammatory response	0.000330722	0.000330722
GO:0043122	regulation of I-kappaB kinase/NF-kappaB signaling	0.000347721	0.000347721
GO:0097529	myeloid leukocyte migration	0.000414075	0.000414075
GO:0002224	toll-like receptor signaling pathway	0.000415086	0.000415086
GO:0050671	positive regulation of lymphocyte proliferation	0.000415086	0.000415086
GO:0032602	chemokine production	0.000416113	0.000416113
GO:0032642	regulation of chemokine production	0.000416113	0.000416113
GO:0032946	positive regulation of mononuclear cell proliferation	0.000432255	0.000432255
GO:0097553	calcium ion transmembrane import into cytosol	0.000435396	0.000435396
GO:0048518	positive regulation of biological process	0.000435396	0.000435396
GO:0030217	T cell differentiation	0.000462365	0.000462365
GO:0043405	regulation of MAP kinase activity	0.000495183	0.000495183
GO:0050794	regulation of cellular process	0.000499699	0.000499699
GO:0002637	regulation of immunoglobulin production	0.00051273	0.00051273
GO:0002696	positive regulation of leukocyte activation	0.000546307	0.000546307
GO:0009617	response to bacterium	0.000588485	0.000588485
GO:0032613	interleukin-10 production	0.000614909	0.000614909
GO:0032653	regulation of interleukin-10 production	0.000614909	0.000614909
GO:0010646	regulation of cell communication	0.000614909	0.000614909
GO:0043368	positive T cell selection	0.000617145	0.000617145
GO:0031664	regulation of lipopolysaccharide-mediated signaling pathway	0.000617145	0.000617145
GO:0023051	regulation of signaling	0.000633952	0.000633952
GO:0042102	positive regulation of T cell proliferation	0.000640656	0.000640656
GO:0050729	positive regulation of inflammatory response	0.000648482	0.000648482
GO:0032663	regulation of interleukin-2 production	0.000648482	0.000648482
GO:0070383	DNA cytosine deamination	0.000648482	0.000648482
GO:0032623	interleukin-2 production	0.000648482	0.000648482
GO:0071310	cellular response to organic substance	0.00072903	0.00072903
GO:0038093	Fc receptor signaling pathway	0.000782428	0.000782428
GO:0002449	lymphocyte mediated immunity	0.000790591	0.000790591
GO:0006909	phagocytosis	0.000838027	0.000838027
GO:0045061	thymic T cell selection	0.00088504	0.00088504
GO:0016554	cytidine to uridine editing	0.00088504	0.00088504
GO:0071216	cellular response to biotic stimulus	0.000891177	0.000891177
GO:0045089	positive regulation of innate immune response	0.000898357	0.000898357
GO:0002879	positive regulation of acute inflammatory response to non-antigenic stimulus	0.000907555	0.000907555
GO:0002426	immunoglobulin production in mucosal tissue	0.000907555	0.000907555
GO:2000557	regulation of immunoglobulin production in mucosal tissue	0.000907555	0.000907555
GO:0045087	innate immune response	0.000907555	0.000907555
GO:2000558	positive regulation of immunoglobulin production in mucosal tissue	0.000907555	0.000907555
GO:0002525	acute inflammatory response to non-antigenic stimulus	0.000907555	0.000907555
GO:0033993	response to lipid	0.000907555	0.000907555
GO:0002877	regulation of acute inflammatory response to non-antigenic stimulus	0.000907555	0.000907555
GO:0071674	mononuclear cell migration	0.001081991	0.001081991
GO:0009615	response to virus	0.001111069	0.001111069
GO:0032722	positive regulation of chemokine production	0.00112736	0.00112736
GO:0002685	regulation of leukocyte migration	0.001201974	0.001201974
GO:0140546	defense response to symbiont	0.001215436	0.001215436
GO:0051607	defense response to virus	0.001215436	0.001215436
GO:0071396	cellular response to lipid	0.001236267	0.001236267
GO:0006950	response to stress	0.001259325	0.001259325
GO:0038094	Fc-gamma receptor signaling pathway	0.001262965	0.001262965
GO:0030593	neutrophil chemotaxis	0.001262965	0.001262965
GO:0045058	T cell selection	0.001262965	0.001262965

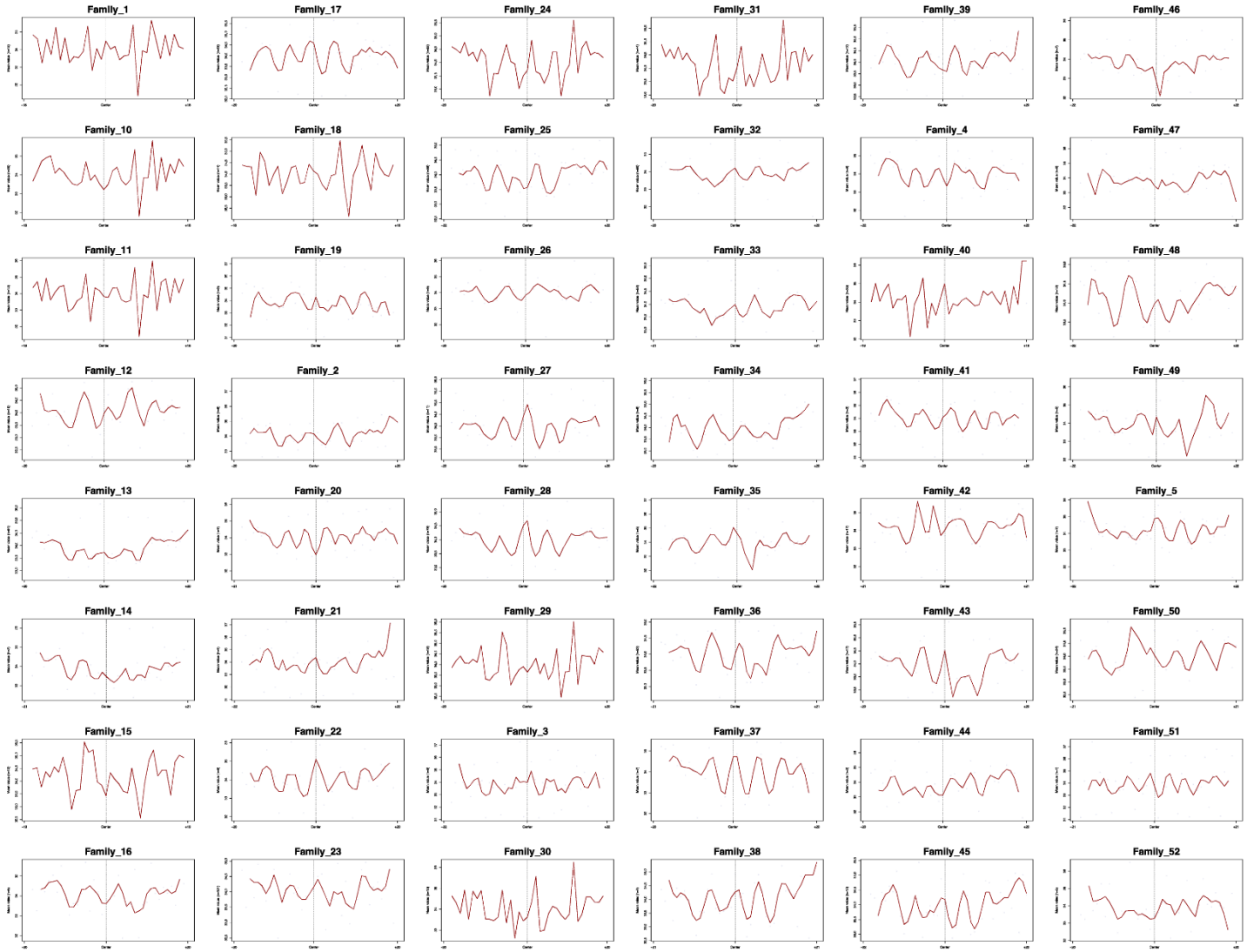
GO:0032609	interferon-gamma production	0.001262965	0.001262965
GO:0032649	regulation of interferon-gamma production	0.001262965	0.001262965
GO:0002218	activation of innate immune response	0.001366144	0.001366144
GO:0002532	production of molecular mediator involved in inflammatory response	0.001366144	0.001366144
GO:0080134	regulation of response to stress	0.001406934	0.001406934
GO:0045123	cellular extravasation	0.001467829	0.001467829
GO:0071222	cellular response to lipopolysaccharide	0.001546387	0.001546387
GO:1902533	positive regulation of intracellular signal transduction	0.001547171	0.001547171
GO:0071677	positive regulation of mononuclear cell migration	0.001551313	0.001551313
GO:0032733	positive regulation of interleukin-10 production	0.001551313	0.001551313
GO:0002526	acute inflammatory response	0.001551313	0.001551313
GO:0070588	calcium ion transmembrane transport	0.001602101	0.001602101
GO:0045006	DNA deamination	0.001724872	0.001724872
GO:0002673	regulation of acute inflammatory response	0.001728104	0.001728104
GO:0051239	regulation of multicellular organismal process	0.001746288	0.001746288
GO:0050789	regulation of biological process	0.001862778	0.001862778
GO:0002819	regulation of adaptive immune response	0.002010754	0.002010754
GO:0002705	positive regulation of leukocyte mediated immunity	0.002074331	0.002074331
GO:0071219	cellular response to molecule of bacterial origin	0.002170104	0.002170104
GO:0009967	positive regulation of signal transduction	0.002226784	0.002226784
GO:0002238	response to molecule of fungal origin	0.00225885	0.00225885
GO:0072676	lymphocyte migration	0.00225885	0.00225885
GO:0030183	B cell differentiation	0.00225885	0.00225885
GO:0071226	cellular response to molecule of fungal origin	0.00225885	0.00225885
GO:0002237	response to molecule of bacterial origin	0.002327333	0.002327333
GO:0045869	negative regulation of single stranded viral RNA replication via double stranded DNA intermediate	0.002533743	0.002533743
GO:0045088	regulation of innate immune response	0.002533743	0.002533743
GO:0002385	mucosal immune response	0.002533743	0.002533743
GO:0033077	T cell differentiation in thymus	0.002533743	0.002533743
GO:0045859	regulation of protein kinase activity	0.002698387	0.002698387
GO:0006816	calcium ion transport	0.002720968	0.002720968
GO:0031295	T cell costimulation	0.002775759	0.002775759
GO:0002366	leukocyte activation involved in immune response	0.002847835	0.002847835
GO:0002251	organ or tissue specific immune response	0.003046143	0.003046143
GO:0071621	granulocyte chemotaxis	0.003053229	0.003053229
GO:0002377	immunoglobulin production	0.003062258	0.003062258
GO:0010033	response to organic substance	0.003271259	0.003271259
GO:0002263	cell activation involved in immune response	0.003282432	0.003282432
GO:0031294	lymphocyte costimulation	0.00328994	0.00328994
GO:0007252	I-kappaB phosphorylation	0.00353432	0.00353432
GO:0098609	cell-cell adhesion	0.003796477	0.003796477
GO:0060326	cell chemotaxis	0.003910682	0.003910682
GO:0002437	inflammatory response to antigenic stimulus	0.003910682	0.003910682
GO:0043549	regulation of kinase activity	0.003974909	0.003974909
GO:0016553	base conversion or substitution editing	0.004051782	0.004051782
GO:0045091	regulation of single stranded viral RNA replication via double stranded DNA intermediate	0.004051782	0.004051782
GO:0007186	G protein-coupled receptor signaling pathway	0.004051782	0.004051782
GO:0002710	negative regulation of T cell mediated immunity	0.004051782	0.004051782
GO:0033630	positive regulation of cell adhesion mediated by integrin	0.004051782	0.004051782
GO:0071398	cellular response to fatty acid	0.004051782	0.004051782
GO:0001954	positive regulation of cell-matrix adhesion	0.004142522	0.004142522
GO:0002534	cytokine production involved in inflammatory response	0.004490953	0.004490953
GO:1900015	regulation of cytokine production involved in inflammatory response	0.004490953	0.004490953
GO:0039692	single stranded viral RNA replication via double stranded DNA intermediate	0.004640311	0.004640311
GO:0002698	negative regulation of immune effector process	0.004640311	0.004640311
GO:0050853	B cell receptor signaling pathway	0.004640311	0.004640311
GO:0002719	negative regulation of cytokine production involved in immune response	0.004640311	0.004640311
GO:1901222	regulation of NIK/NF-kappaB signaling	0.00486036	0.00486036
GO:0014065	phosphatidylinositol 3-kinase signaling	0.005022521	0.005022521
GO:0032651	regulation of interleukin-1 beta production	0.005085136	0.005085136
GO:0032611	interleukin-1 beta production	0.005085136	0.005085136
GO:0019722	calcium-mediated signaling	0.005258622	0.005258622
GO:0006935	chemotaxis	0.005390484	0.005390484
GO:0042330	taxis	0.005390484	0.005390484

GO:0070887	cellular response to chemical stimulus	0.005911985	0.005911985
GO:0030888	regulation of B cell proliferation	0.005958639	0.005958639
GO:0002706	regulation of lymphocyte mediated immunity	0.006008889	0.006008889
GO:0045577	regulation of B cell differentiation	0.00601741	0.00601741
GO:2000523	regulation of T cell costimulation	0.006154972	0.006154972
GO:0032496	response to lipopolysaccharide	0.006435286	0.006435286
GO:0010811	positive regulation of cell-substrate adhesion	0.006435286	0.006435286
GO:0010647	positive regulation of cell communication	0.006465847	0.006465847
GO:0032703	negative regulation of interleukin-2 production	0.006743232	0.006743232
GO:0002891	positive regulation of immunoglobulin mediated immune response	0.006743232	0.006743232
GO:0002714	positive regulation of B cell mediated immunity	0.006743232	0.006743232
GO:0023056	positive regulation of signaling	0.006761765	0.006761765
GO:0050777	negative regulation of immune response	0.006800008	0.006800008
GO:0048525	negative regulation of viral process	0.006959196	0.006959196
GO:1903169	regulation of calcium ion transmembrane transport	0.007021788	0.007021788
GO:0050854	regulation of antigen receptor-mediated signaling pathway	0.007120548	0.007120548
GO:0032102	negative regulation of response to external stimulus	0.007488582	0.007488582
GO:0006801	superoxide metabolic process	0.007609318	0.007609318
GO:0051924	regulation of calcium ion transport	0.007691448	0.007691448
GO:0014066	regulation of phosphatidylinositol 3-kinase signaling	0.00792478	0.00792478
GO:0032652	regulation of interleukin-1 production	0.008265842	0.008265842
GO:0032612	interleukin-1 production	0.008265842	0.008265842
GO:0061099	negative regulation of protein tyrosine kinase activity	0.008378341	0.008378341
GO:0050790	regulation of catalytic activity	0.008464202	0.008464202
GO:0002692	negative regulation of cellular extravasation	0.008464202	0.008464202
GO:1903721	positive regulation of I-kappaB phosphorylation	0.008464202	0.008464202
GO:0033634	positive regulation of cell-cell adhesion mediated by integrin	0.008464202	0.008464202
GO:0065007	biological regulation	0.008720037	0.008720037
GO:0016064	immunoglobulin mediated immune response	0.008720037	0.008720037
GO:0002832	negative regulation of response to biotic stimulus	0.008851095	0.008851095
GO:0045071	negative regulation of viral genome replication	0.008978332	0.008978332
GO:0002701	negative regulation of production of molecular mediator of immune response	0.009166584	0.009166584
GO:0019724	B cell mediated immunity	0.009947069	0.009947069
GO:0038096	Fc-gamma receptor signaling pathway involved in phagocytosis	0.010123844	0.010123844
GO:0032743	positive regulation of interleukin-2 production	0.010123844	0.010123844
GO:0050672	negative regulation of lymphocyte proliferation	0.010123844	0.010123844
GO:0002433	immune response-regulating cell surface receptor signaling pathway involved in phagocytosis	0.010123844	0.010123844
GO:0032945	negative regulation of mononuclear cell proliferation	0.010661701	0.010661701
GO:0072678	T cell migration	0.010661701	0.010661701
GO:0031666	positive regulation of lipopolysaccharide-mediated signaling pathway	0.010859184	0.010859184
GO:0035701	hematopoietic stem cell migration	0.010859184	0.010859184
GO:0002752	cell surface pattern recognition receptor signaling pathway	0.010859184	0.010859184
GO:0045619	regulation of lymphocyte differentiation	0.010859184	0.010859184
GO:0002732	positive regulation of dendritic cell cytokine production	0.010859184	0.010859184
GO:2000272	negative regulation of signaling receptor activity	0.010859184	0.010859184
GO:0010529	negative regulation of transposition	0.010859184	0.010859184
GO:0010528	regulation of transposition	0.010859184	0.010859184
GO:0070542	response to fatty acid	0.010859184	0.010859184
GO:1903719	regulation of I-kappaB phosphorylation	0.010859184	0.010859184
GO:0030595	leukocyte chemotaxis	0.010859184	0.010859184
GO:0008284	positive regulation of cell population proliferation	0.010859184	0.010859184
GO:0071675	regulation of mononuclear cell migration	0.011265926	0.011265926
GO:0038061	NIK/NF-kappaB signaling	0.011265926	0.011265926
GO:0042113	B cell activation	0.011265926	0.011265926
GO:0039694	viral RNA genome replication	0.011832838	0.011832838
GO:0050901	leukocyte tethering or rolling	0.011832838	0.011832838
GO:0048015	phosphatidylinositol-mediated signaling	0.012112633	0.012112633
GO:0043254	regulation of protein-containing complex assembly	0.012577554	0.012577554
GO:0048017	inositol lipid-mediated signaling	0.012863661	0.012863661
GO:2000406	positive regulation of T cell migration	0.012911804	0.012911804
GO:0080111	DNA demethylation	0.012911804	0.012911804
GO:0070664	negative regulation of leukocyte proliferation	0.013451563	0.013451563
GO:0042221	response to chemical	0.013977709	0.013977709
GO:0007204	positive regulation of cytosolic calcium ion concentration	0.013977709	0.013977709
GO:0051279	regulation of release of sequestered calcium ion into cytosol	0.013993161	0.013993161

GO:0150077	regulation of neuroinflammatory response	0.013993161	0.013993161
GO:0032196	transposition	0.013993161	0.013993161
GO:0032689	negative regulation of interferon-gamma production	0.013993161	0.013993161
GO:0016477	cell migration	0.014137848	0.014137848
GO:0072507	divalent inorganic cation homeostasis	0.014505758	0.014505758
GO:0030890	positive regulation of B cell proliferation	0.015166606	0.015166606
GO:0043507	positive regulation of JUN kinase activity	0.015166606	0.015166606
GO:0061097	regulation of protein tyrosine kinase activity	0.015435761	0.015435761
GO:1902105	regulation of leukocyte differentiation	0.015477851	0.015477851
GO:0050864	regulation of B cell activation	0.015762353	0.015762353
GO:0071356	cellular response to tumor necrosis factor	0.015762353	0.015762353
GO:0051338	regulation of transferase activity	0.016043375	0.016043375
GO:0045730	respiratory burst	0.016189681	0.016189681
GO:0002431	Fc receptor mediated stimulatory signaling pathway	0.016189681	0.016189681
GO:0042554	superoxide anion generation	0.016189681	0.016189681
GO:0050850	positive regulation of calcium-mediated signaling	0.016189681	0.016189681
GO:0006812	cation transport	0.016473414	0.016473414
GO:0002725	negative regulation of T cell cytokine production	0.016473414	0.016473414
GO:0002371	dendritic cell cytokine production	0.016473414	0.016473414
GO:0001771	immunological synapse formation	0.016473414	0.016473414
GO:0002730	regulation of dendritic cell cytokine production	0.016473414	0.016473414
GO:2001187	positive regulation of CD8-positive, alpha-beta T cell activation	0.016473414	0.016473414
GO:0002274	myeloid leukocyte activation	0.01667619	0.01667619
GO:0051345	positive regulation of hydrolase activity	0.017031812	0.017031812
GO:0006874	cellular calcium ion homeostasis	0.017190273	0.017190273
GO:0035510	DNA dealkylation	0.017190273	0.017190273
GO:0071900	regulation of protein serine/threonine kinase activity	0.018455118	0.018455118
GO:0032715	negative regulation of interleukin-6 production	0.018634454	0.018634454
GO:0046632	alpha-beta T cell differentiation	0.01969355	0.01969355
GO:0002687	positive regulation of leukocyte migration	0.01969355	0.01969355
GO:1905155	positive regulation of membrane invagination	0.01969355	0.01969355
GO:0050856	regulation of T cell receptor signaling pathway	0.01969355	0.01969355
GO:0034142	toll-like receptor 4 signaling pathway	0.01969355	0.01969355
GO:0033632	regulation of cell-cell adhesion mediated by integrin	0.01969355	0.01969355
GO:0060100	positive regulation of phagocytosis, engulfment	0.01969355	0.01969355
GO:1902622	regulation of neutrophil migration	0.01969355	0.01969355
GO:2000403	positive regulation of lymphocyte migration	0.01969355	0.01969355
GO:0002381	immunoglobulin production involved in immunoglobulin-mediated immune response	0.01969355	0.01969355
GO:0031348	negative regulation of defense response	0.020163552	0.020163552
GO:1903706	regulation of hemopoiesis	0.020163552	0.020163552
GO:0002823	negative regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	0.021179088	0.021179088
GO:0010469	regulation of signaling receptor activity	0.02188832	0.02188832
GO:0055074	calcium ion homeostasis	0.02188832	0.02188832
GO:0034612	response to tumor necrosis factor	0.022143164	0.022143164
GO:0048870	cell motility	0.022143164	0.022143164
GO:0030101	natural killer cell activation	0.022143164	0.022143164
GO:0002889	regulation of immunoglobulin mediated immune response	0.022536238	0.022536238
GO:0042100	B cell proliferation	0.023000802	0.023000802
GO:0034154	toll-like receptor 7 signaling pathway	0.023073529	0.023073529
GO:0042116	macrophage activation	0.023979626	0.023979626
GO:0002712	regulation of B cell mediated immunity	0.023995791	0.023995791
GO:0002707	negative regulation of lymphocyte mediated immunity	0.023995791	0.023995791
GO:0050871	positive regulation of B cell activation	0.024457447	0.024457447
GO:2000404	regulation of T cell migration	0.025703611	0.025703611
GO:0002822	regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	0.026016257	0.026016257
GO:0018108	peptidyl-tyrosine phosphorylation	0.026467035	0.026467035
GO:0034695	response to prostaglandin E	0.026639653	0.026639653
GO:0060099	regulation of phagocytosis, engulfment	0.026639653	0.026639653
GO:0071346	cellular response to interferon-gamma	0.026639653	0.026639653
GO:1905153	regulation of membrane invagination	0.026639653	0.026639653
GO:0001867	complement activation, lectin pathway	0.026639653	0.026639653
GO:0001779	natural killer cell differentiation	0.026639653	0.026639653
GO:0043277	apoptotic cell clearance	0.026973093	0.026973093
GO:0018212	peptidyl-tyrosine modification	0.027081775	0.027081775

GO:0043269	regulation of ion transport	0.0274162	0.0274162
GO:0002695	negative regulation of leukocyte activation	0.027885725	0.027885725
GO:0007155	cell adhesion	0.028279444	0.028279444
GO:0046634	regulation of alpha-beta T cell activation	0.028279444	0.028279444
GO:0009620	response to fungus	0.028279444	0.028279444
GO:0002820	negative regulation of adaptive immune response	0.028279444	0.028279444
GO:0033628	regulation of cell adhesion mediated by integrin	0.028279444	0.028279444
GO:0045069	regulation of viral genome replication	0.028279444	0.028279444
GO:0043124	negative regulation of I-kappaB kinase/NF-kappaB signaling	0.028279444	0.028279444
GO:0019932	second-messenger-mediated signaling	0.028784641	0.028784641
GO:1903900	regulation of viral life cycle	0.029052059	0.029052059
GO:0002821	positive regulation of adaptive immune response	0.029228647	0.029228647
GO:0031341	regulation of cell killing	0.029228647	0.029228647
GO:2000010	positive regulation of protein localization to cell surface	0.029563665	0.029563665
GO:1903428	positive regulation of reactive oxygen species biosynthetic process	0.029563665	0.029563665
GO:0002281	macrophage activation involved in immune response	0.029563665	0.029563665
GO:0002438	acute inflammatory response to antigenic stimulus	0.029563665	0.029563665
GO:0150078	positive regulation of neuroinflammatory response	0.029563665	0.029563665
GO:0031665	negative regulation of lipopolysaccharide-mediated signaling pathway	0.029563665	0.029563665
GO:0061756	leukocyte adhesion to vascular endothelial cell	0.029563665	0.029563665
GO:0002708	positive regulation of lymphocyte mediated immunity	0.029821865	0.029821865
GO:0002456	T cell mediated immunity	0.029821865	0.029821865
GO:0051090	regulation of DNA-binding transcription factor activity	0.030156343	0.030156343
GO:0072503	cellular divalent inorganic cation homeostasis	0.030900223	0.030900223
GO:0001932	regulation of protein phosphorylation	0.032852444	0.032852444
GO:0062208	positive regulation of pattern recognition receptor signaling pathway	0.032901349	0.032901349
GO:0043506	regulation of JUN kinase activity	0.032901349	0.032901349
GO:0050730	regulation of peptidyl-tyrosine phosphorylation	0.033073105	0.033073105
GO:0010820	positive regulation of T cell chemotaxis	0.033396178	0.033396178
GO:0032725	positive regulation of granulocyte macrophage colony-stimulating factor production	0.033396178	0.033396178
GO:0002283	neutrophil activation involved in immune response	0.033396178	0.033396178
GO:0051403	stress-activated MAPK cascade	0.033557327	0.033557327
GO:0002704	negative regulation of leukocyte mediated immunity	0.034455812	0.034455812
GO:0050732	negative regulation of peptidyl-tyrosine phosphorylation	0.034455812	0.034455812
GO:0043406	positive regulation of MAP kinase activity	0.035443875	0.035443875
GO:0065009	regulation of molecular function	0.036370684	0.036370684
GO:0042130	negative regulation of T cell proliferation	0.036381229	0.036381229
GO:0031098	stress-activated protein kinase signaling cascade	0.036387884	0.036387884
GO:1901701	cellular response to oxygen-containing compound	0.036631881	0.036631881
GO:0034694	response to prostaglandin	0.037075732	0.037075732
GO:0034116	positive regulation of heterotypic cell-cell adhesion	0.037075732	0.037075732
GO:0010819	regulation of T cell chemotaxis	0.037075732	0.037075732
GO:0034134	toll-like receptor 2 signaling pathway	0.037075732	0.037075732
GO:0033631	cell-cell adhesion mediated by integrin	0.037075732	0.037075732
GO:0040011	locomotion	0.037075732	0.037075732
GO:0062207	regulation of pattern recognition receptor signaling pathway	0.038791444	0.038791444
GO:0002285	lymphocyte activation involved in immune response	0.039305427	0.039305427
GO:0032645	regulation of granulocyte macrophage colony-stimulating factor production	0.041185212	0.041185212
GO:0061154	endothelial tube morphogenesis	0.041185212	0.041185212
GO:0032604	granulocyte macrophage colony-stimulating factor production	0.041185212	0.041185212
GO:0051092	positive regulation of NF-kappaB transcription factor activity	0.041185212	0.041185212
GO:1900221	regulation of amyloid-beta clearance	0.041185212	0.041185212
GO:0098655	cation transmembrane transport	0.041185212	0.041185212
GO:0003159	morphogenesis of an endothelium	0.041185212	0.041185212
GO:0050866	negative regulation of cell activation	0.041185212	0.041185212
GO:0035746	granzyme A production	0.04227481	0.04227481
GO:0002669	positive regulation of T cell anergy	0.04227481	0.04227481
GO:0002325	natural killer cell differentiation involved in immune response	0.04227481	0.04227481
GO:2000334	positive regulation of blood microparticle formation	0.04227481	0.04227481
GO:0045584	negative regulation of cytotoxic T cell differentiation	0.04227481	0.04227481
GO:0150129	positive regulation of interleukin-33 production	0.04227481	0.04227481
GO:0072682	eosinophil extravasation	0.04227481	0.04227481
GO:0035782	mature natural killer cell chemotaxis	0.04227481	0.04227481
GO:2000332	regulation of blood microparticle formation	0.04227481	0.04227481
GO:0097534	lymphoid lineage cell migration	0.04227481	0.04227481

GO:0032826	regulation of natural killer cell differentiation involved in immune response	0.04227481	0.04227481
GO:0002442	serotonin secretion involved in inflammatory response	0.04227481	0.04227481
GO:0097535	lymphoid lineage cell migration into thymus	0.04227481	0.04227481
GO:0042325	regulation of phosphorylation	0.04227481	0.04227481
GO:0090721	primary adaptive immune response involving T cells and B cells	0.04227481	0.04227481
GO:2000526	positive regulation of glycoprotein biosynthetic process involved in immunological synapse formation	0.04227481	0.04227481
GO:0090720	primary adaptive immune response	0.04227481	0.04227481
GO:0002554	serotonin secretion by platelet	0.04227481	0.04227481
GO:2000513	positive regulation of granzyme A production	0.04227481	0.04227481
GO:2000511	regulation of granzyme A production	0.04227481	0.04227481
GO:2000517	regulation of T-helper 1 cell activation	0.04227481	0.04227481
GO:0042543	protein N-linked glycosylation via arginine	0.04227481	0.04227481
GO:0038123	toll-like receptor TLR1:TLR2 signaling pathway	0.04227481	0.04227481
GO:0002913	positive regulation of lymphocyte anergy	0.04227481	0.04227481
GO:2000420	negative regulation of eosinophil extravasation	0.04227481	0.04227481
GO:2000419	regulation of eosinophil extravasation	0.04227481	0.04227481
GO:0002439	chronic inflammatory response to antigenic stimulus	0.04227481	0.04227481
GO:2000518	negative regulation of T-helper 1 cell activation	0.04227481	0.04227481
GO:0014895	smooth muscle hypertrophy	0.04227481	0.04227481
GO:0072564	blood microparticle formation	0.04227481	0.04227481
GO:0048298	positive regulation of isotype switching to IgA isotypes	0.04227481	0.04227481
GO:0002351	serotonin production involved in inflammatory response	0.04227481	0.04227481
GO:0150127	regulation of interleukin-33 production	0.04227481	0.04227481
GO:0072639	interleukin-33 production	0.04227481	0.04227481
GO:0061048	negative regulation of branching involved in lung morphogenesis	0.04227481	0.04227481
GO:1901700	response to oxygen-containing compound	0.04227481	0.04227481
GO:0010604	positive regulation of macromolecule metabolic process	0.042332873	0.042332873
GO:1900017	positive regulation of cytokine production involved in inflammatory response	0.042404035	0.042404035
GO:0033623	regulation of integrin activation	0.042404035	0.042404035
GO:0034162	toll-like receptor 9 signaling pathway	0.042404035	0.042404035
GO:0046635	positive regulation of alpha-beta T cell activation	0.042404035	0.042404035
GO:0032753	positive regulation of interleukin-4 production	0.042404035	0.042404035
GO:0050792	regulation of viral process	0.042597594	0.042597594
GO:0001952	regulation of cell-matrix adhesion	0.044537257	0.044537257
GO:2001185	regulation of CD8-positive, alpha-beta T cell activation	0.046537297	0.046537297
GO:2000379	positive regulation of reactive oxygen species metabolic process	0.046537297	0.046537297
GO:0150076	neuroinflammatory response	0.046537297	0.046537297
GO:1900225	regulation of NLRP3 inflammasome complex assembly	0.046537297	0.046537297
GO:0046641	positive regulation of alpha-beta T cell proliferation	0.046537297	0.046537297
GO:0043652	engulfment of apoptotic cell	0.046537297	0.046537297
GO:0030050	vesicle transport along actin filament	0.046537297	0.046537297
GO:0034341	response to interferon-gamma	0.046932004	0.046932004
GO:0030001	metal ion transport	0.04704543	0.04704543
GO:0051336	regulation of hydrolase activity	0.047943669	0.047943669
GO:1903038	negative regulation of leukocyte cell-cell adhesion	0.048216447	0.048216447
GO:0031343	positive regulation of cell killing	0.048216447	0.048216447
GO:1904062	regulation of cation transmembrane transport	0.048851314	0.048851314
GO:0042127	regulation of cell population proliferation	0.04917162	0.04917162


Family_1
Family_17
Family_24
Family_31
Family_39
Family_46
Family_10
Family_18
Family_25
Family_32
Family_4
Family_47
Family_11
Family_19
Family_26
Family_33
Family_40
Family_48
Family_12
Family_2
Family_27
Family_34
Family_41
Family_49
Family_13
Family_20
Family_28
Family_35
Family_42
Family_5
Family_14
Family_21
Family_29
Family_36
Family_43
Family_50
Family_15
Family_22
Family_3
Family_37
Family_44
Family_51
Family_16
Family_23
Family_30
Family_38
Family_45
Family_52

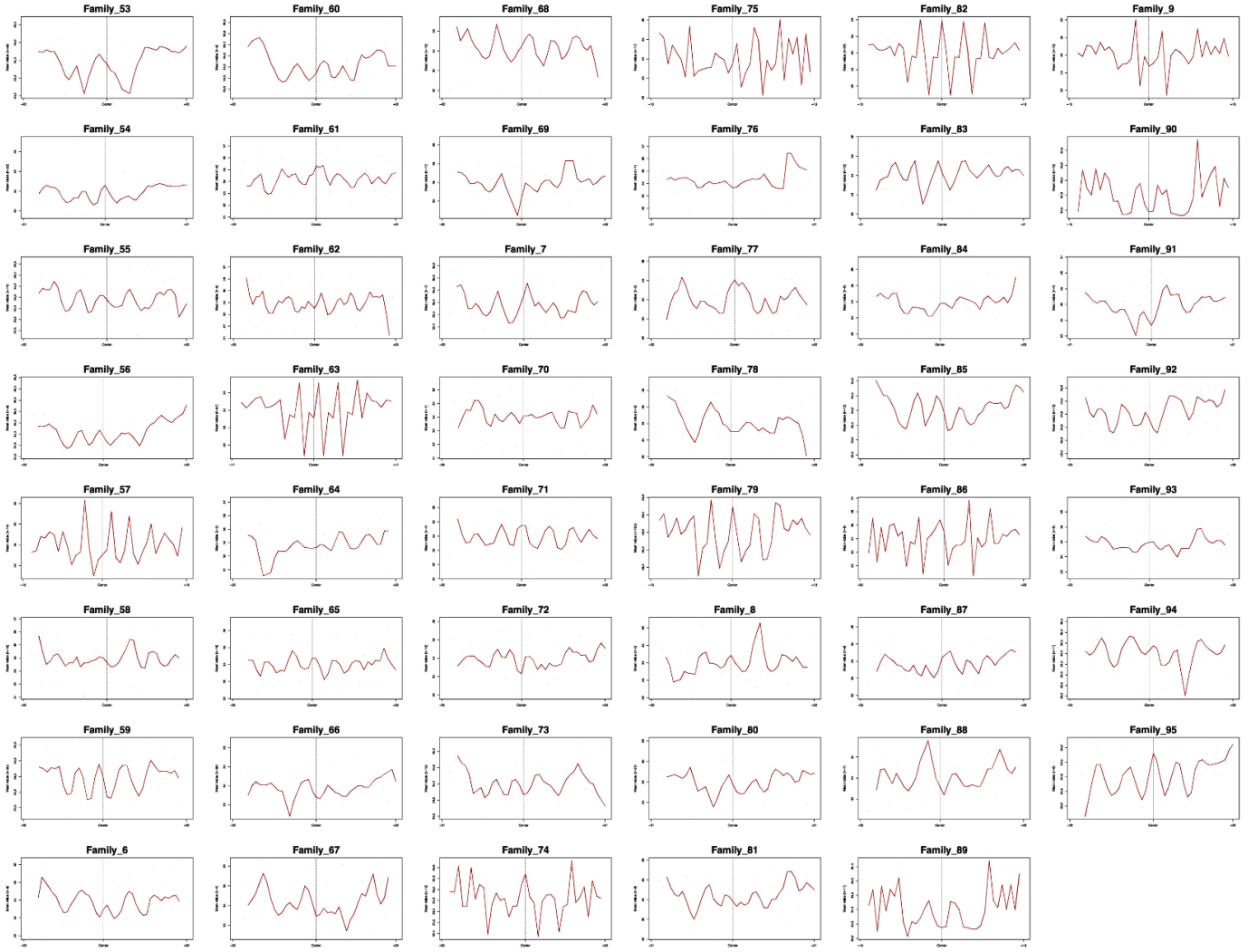
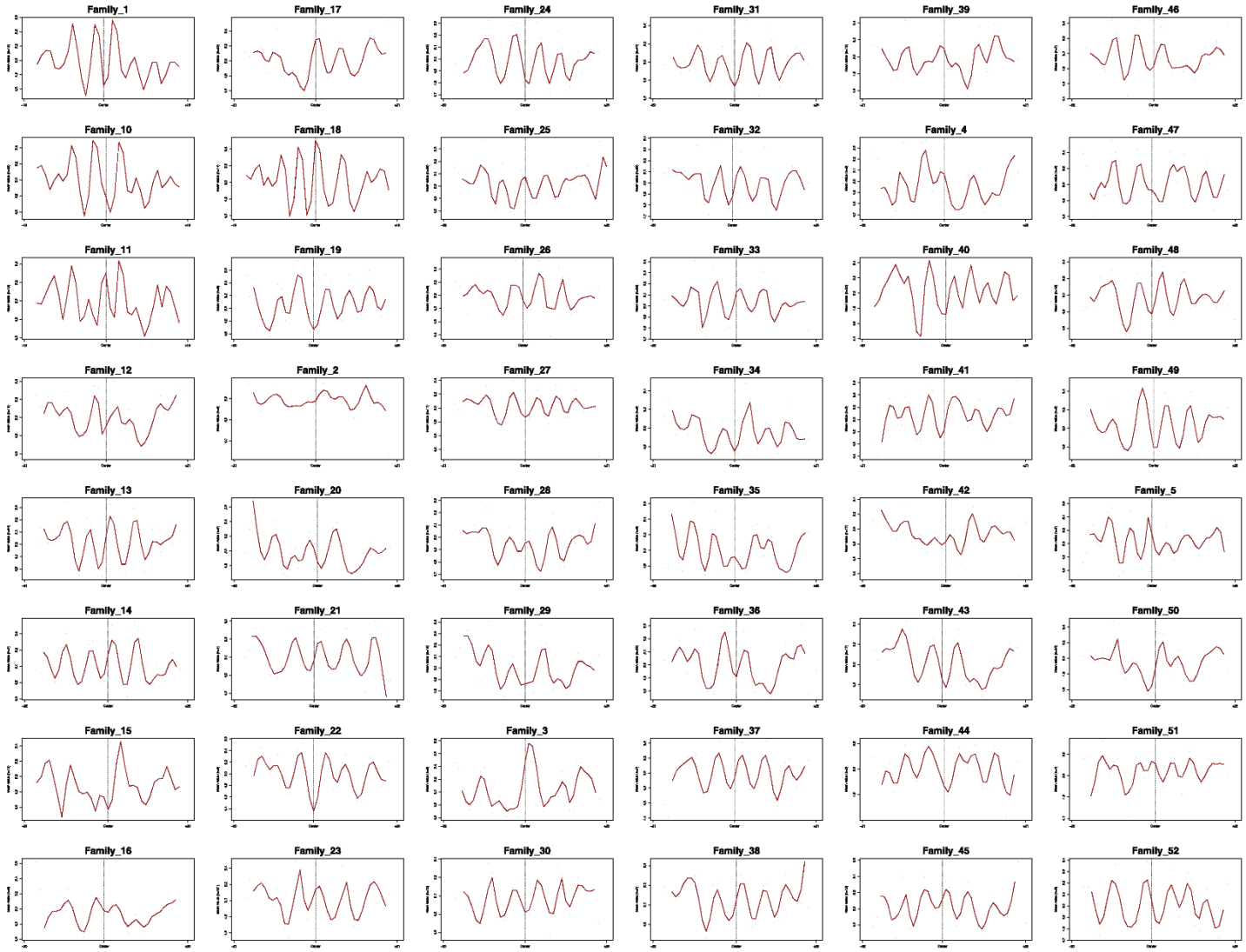
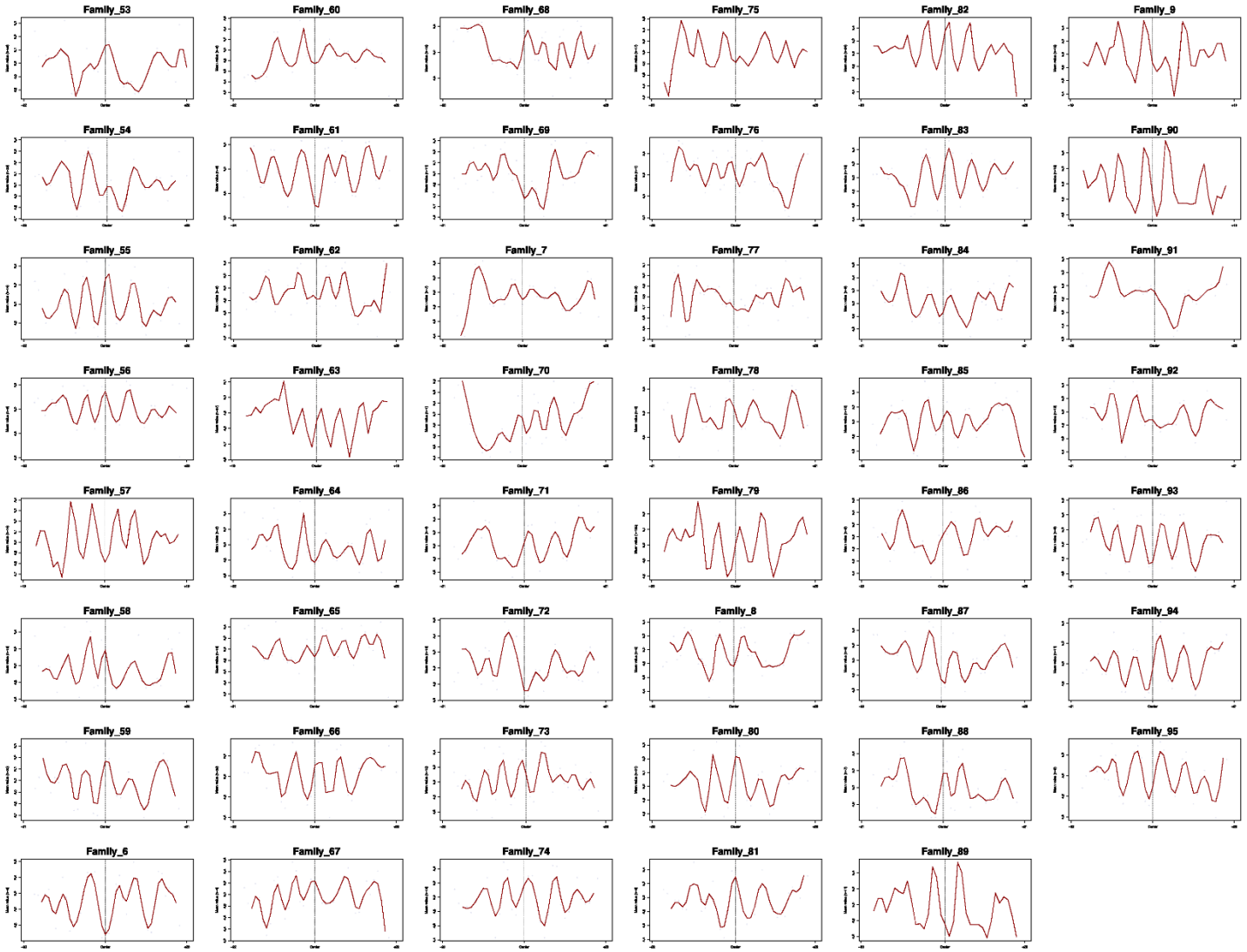
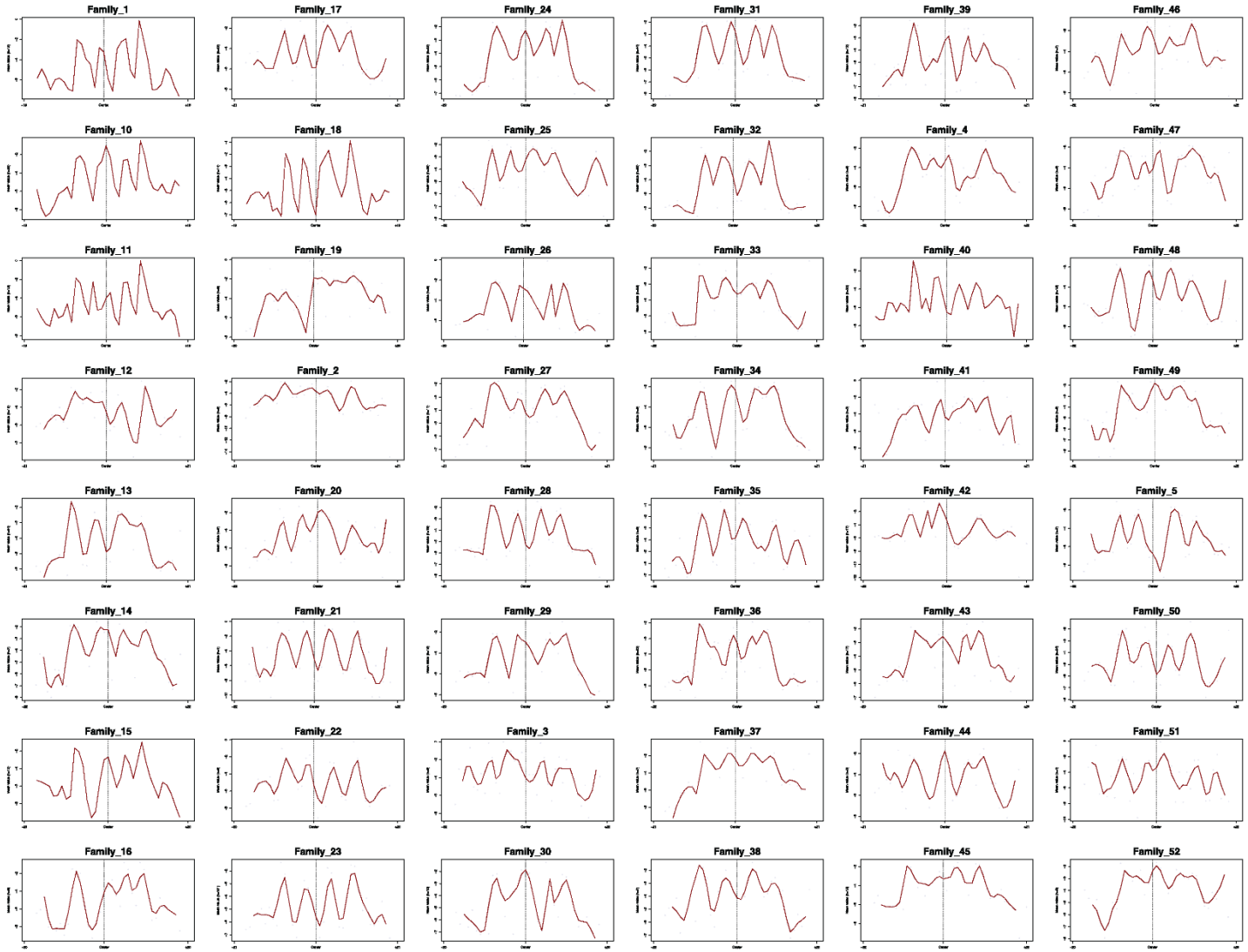


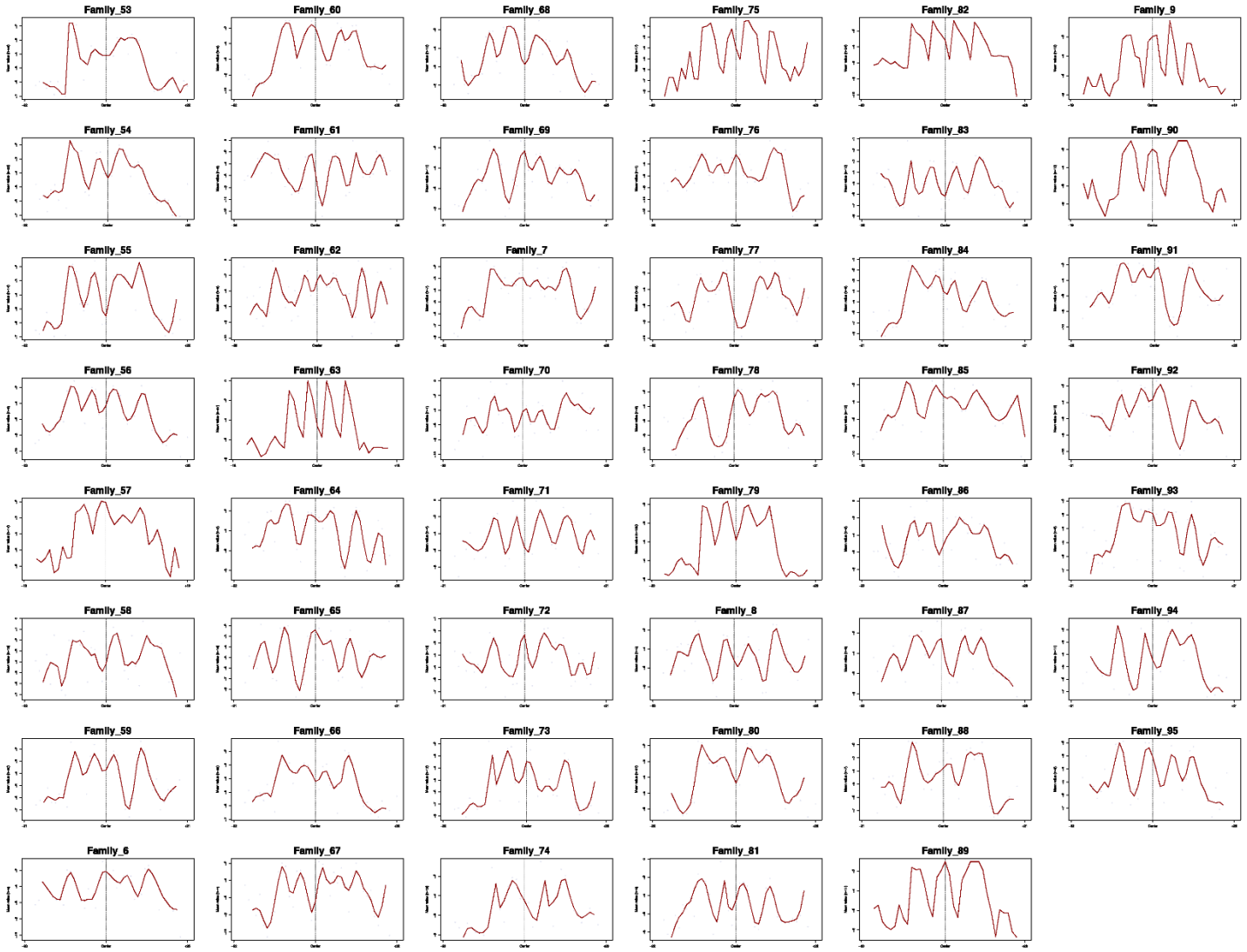
Figure A 1 Helical Twist across all Families


Family_1
Family_17
Family_24
Family_31
Family_39
Family_46
Family_10
Family_18
Family_25
Family_32
Family_4
Family_47
Family_11
Family_19
Family_26
Family_33
Family_40
Family_48
Family_12
Family_2
Family_27
Family_34
Family_41
Family_49
Family_13
Family_20
Family_28
Family_35
Family_42
Family_5
Family_14
Family_21
Family_29
Family_36
Family_43
Family_50
Family_15
Family_22
Family_3
Family_37
Family_44
Family_51
Family_16
Family_23
Family_30
Family_38
Family_45
Family_52

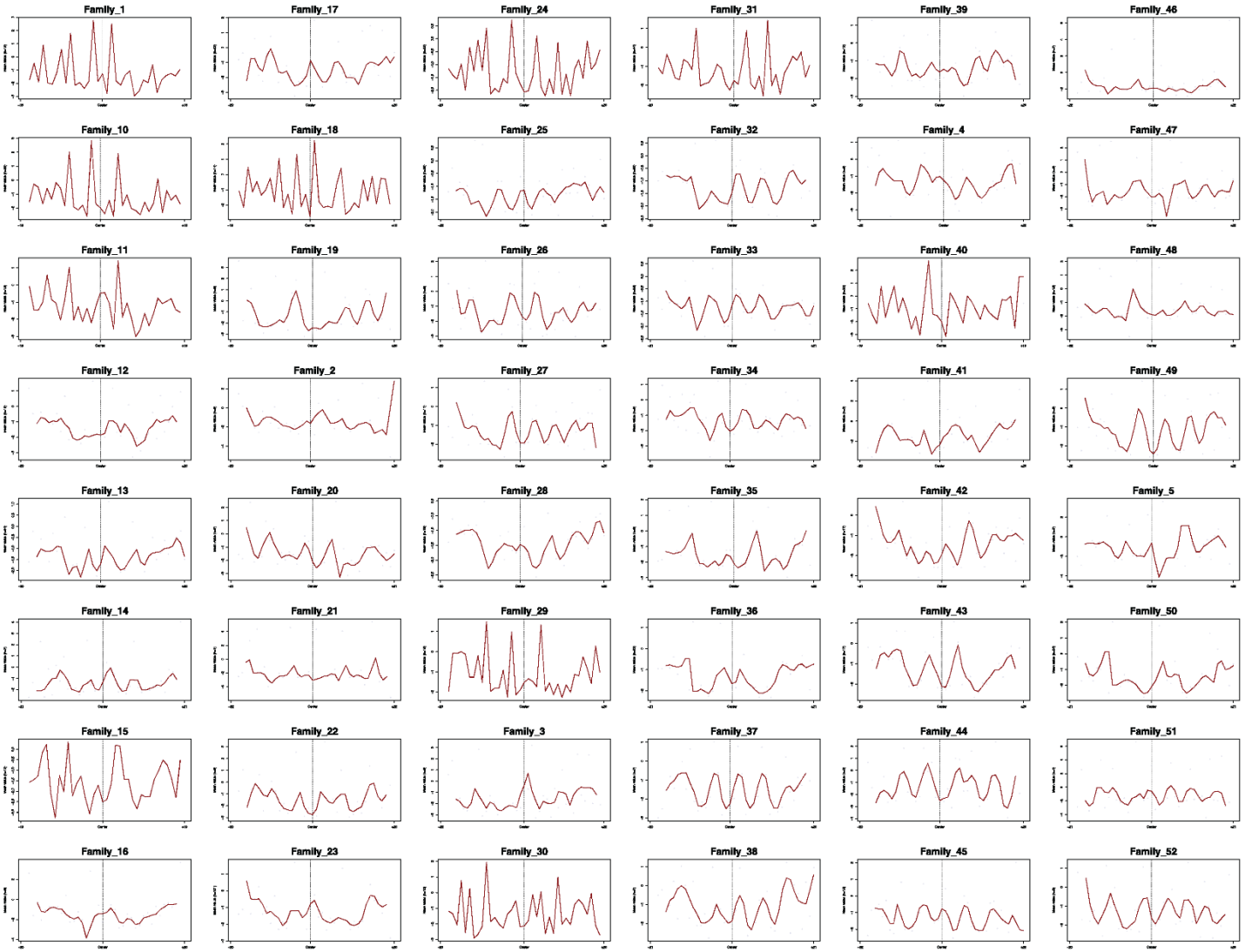


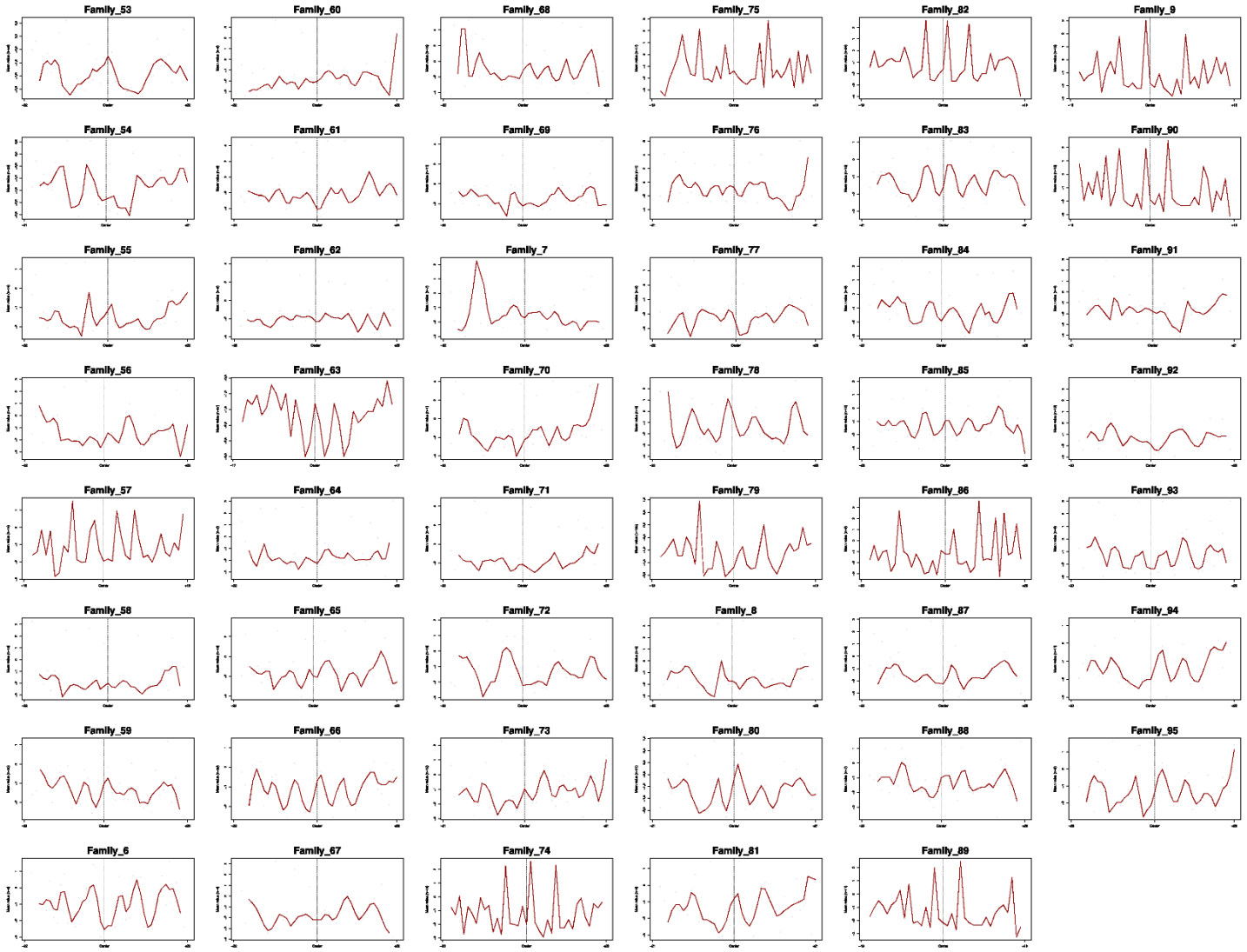
Appendix Figure A 2 Minor Groove width across all families


Family_1
Family_17
Family_24
Family_31
Family_39
Family_46
Family_10
Family_18
Family_25
Family_32
Family_4
Family_47
Family_11
Family_19
Family_26
Family_33
Family_40
Family_48
Family_12
Family_2
Family_27
Family_34
Family_41
Family_49
Family_13
Family_20
Family_28
Family_35
Family_42
Family_5
Family_14
Family_21
Family_29
Family_36
Family_43
Family_50
Family_15
Family_22
Family_3
Family_37
Family_44
Family_51
Family_16
Family_23
Family_30
Family_38
Family_45
Family_52



Appendix Figure A 3 Propeller twist across families


Family_1
Family_17
Family_24
Family_31
Family_39
Family_46
Family_10
Family_18
Family_25
Family_32
Family_4
Family_47
Family_11
Family_19
Family_26
Family_33
Family_40
Family_48
Family_12
Family_2
Family_27
Family_34
Family_41
Family_49
Family_13
Family_20
Family_28
Family_35
Family_42
Family_5
Family_14
Family_21
Family_29
Family_36
Family_43
Family_50
Family_15
Family_22
Family_3
Family_37
Family_44
Family_51
Family_16
Family_23
Family_30
Family_38
Family_45
Family_52



Appendix Figure A 4 : Roll across families


Family_1
Family_17
Family_24
Family_31
Family_39
Family_46
Family_10
Family_18
Family_25
Family_32
Family_4
Family_47
Family_11
Family_19
Family_26
Family_33
Family_40
Family_48
Family_12
Family_2
Family_27
Family_34
Family_41
Family_49
Family_13
Family_20
Family_28
Family_35
Family_42
Family_5
Family_14
Family_21
Family_29
Family_36
Family_43
Family_50
Family_15
Family_22
Family_3
Family_37
Family_44
Family_51
Family_16
Family_23
Family_30
Family_38
Family_45
Family_52
Bits

Appendix Figure A 5: Sequence logo of families

Appendix B

Appendix Table B 1 Count of SNVs in overall COSMIC database

		TO			
		A	**C**	**G**	**T**
FROM	**A**	0	94,513	2,006,240	1,121,852
	C	1,965,512	0	990,818	3,788,310
	G	3,788,966	989,193	0	2,017,502
	T	1,124,289	2,007,011	951,101	0

Appendix Table B 2 Counts of SNVs in G4 regions from the COSMIC database

		TO			
		A	**C**	**G**	**T**
FROM	**A**	0	601	7,191	768
	C	2,529	0	2,937	8,094
	G	21,349	4,450	0	10,791
	T	742	1,608	8,861	0

Appendix Table B 3 . Changes in putative G4 from the COSMIC database across both strands before and after mutation. (0: absence of pG4; 1: presence of pG4 in forward strand; -1: presence of pG4 in reverse strand)

G4 with reference allele	G4 with alternate allele	count	%	G4hunter score (reference)	Sd (reference)	G4hunter score (alternate)	Sd (alternate)
-1	-1	13,793	36.77	-1.222	0.387	-1.219	0.4
-1	0	3,581	9.55	-1.026	0.351	-0.901	0.345
-1	1	6	0.02	0.175	0.37	0.279	0.36
0	-1	1,354	3.61	-1.082	0.356	-1.196	0.359
0	1	1,374	3.66	1.085	0.368	1.201	0.371
1	0	3,655	9.74	0.993	0.359	0.871	0.355
1	1	13,753	36.66	1.222	0.376	1.219	0.391

Appendix Table B 4 Count and proportion of variants in experimentally validated G4 regions for different functional regions.

	COSMIC		CLINVAR	
Annotation	**Count**	**Frequency**	**Count**	**Frequency**
CDS	7,569	10.02	2,281	45.38
5' UTR	1,514	2	--	--
3' UTR	3,669	4.86	179	3.56
EXON	13,034	17.26	3,014	59.97
INTRON	26,014	34.44	1,251	24.89
PROMOTER	9,248	12.24	554	11.02
ENHANCER	563	0.75	--	--
CpG ISLAND	5,356	7.09	--	--
GENCODE lncRNA	3,121	4.13	700	13.92
INTERGENIC	5,441	7.2	761	15.14

Appendix Table B 5 Significant GO:BP enrichments for all COSMIC and CLINVAR G4 mutations.

GO ID	GO Description	Universe	COSMIC and CLINVAR	AdjustedP-value
GO:0007399	nervous system development	1648	1087	3.01E-48
GO:0048856	anatomical structure development	4152	2427	6.54E-48
GO:0032502	developmental process	4584	2644	1.02E-46
GO:0048731	system development	2976	1790	1.56E-42
GO:0007275	multicellular organism development	3266	1938	2.46E-41
GO:0009653	anatomical structure morphogenesis	1871	1181	1.16E-38
GO:0048699	generation of neurons	969	670	1.07E-37
GO:0000902	cell morphogenesis	718	520	4.07E-37
GO:0030154	cell differentiation	2860	1706	5.88E-37
GO:0048869	cellular developmental process	2879	1715	1.04E-36
GO:0022008	neurogenesis	1096	736	6.89E-35
GO:0030182	neuron differentiation	923	636	7.71E-35
GO:0048468	cell development	1355	877	4.69E-33
GO:0032501	multicellular organismal process	5319	2948	8.54E-33
GO:0048858	cell projection morphogenesis	459	352	1.36E-32
GO:0048666	neuron development	729	516	2.16E-32
GO:0032989	cellular component morphogenesis	549	407	3.74E-32
GO:0030030	cell projection organization	1164	766	7.19E-32
GO:0120039	plasma membrane bounded cell projection morphogenesis	455	348	9.12E-32
GO:0031175	neuron projection development	656	471	9.40E-32
GO:0120036	plasma membrane bounded cell projection organization	1144	754	1.22E-31
GO:0032990	cell part morphogenesis	469	355	8.10E-31
GO:0048812	neuron projection morphogenesis	441	337	1.53E-30
GO:0000904	cell morphogenesis involved in differentiation	495	367	1.01E-28
GO:0023051	regulation of signaling	2733	1601	2.82E-28
GO:0010646	regulation of cell communication	2727	1596	7.06E-28
GO:0051128	regulation of cellular component organization	1929	1170	3.23E-27
GO:0061564	axon development	323	256	5.08E-27
GO:0030029	actin filament-based process	717	495	1.15E-26
GO:0050793	regulation of developmental process	1810	1101	5.46E-26
GO:0048667	cell morphogenesis involved in neuron differentiation	381	291	6.02E-26
GO:0007409	axonogenesis	298	237	3.00E-25
GO:0023052	signaling	5190	2837	8.54E-25
GO:0007154	cell communication	5221	2844	1.92E-23
GO:0016043	cellular component organization	5370	2912	1.84E-22
GO:0051239	regulation of multicellular organismal process	2117	1248	3.72E-22
GO:0048513	animal organ development	2176	1277	1.06E-21
GO:0030036	actin cytoskeleton organization	637	435	1.52E-21
GO:0035556	intracellular signal transduction	2168	1267	1.65E-20
GO:0007155	cell adhesion	1216	754	1.02E-19
GO:0009966	regulation of signal transduction	2464	1416	2.43E-19
GO:0007010	cytoskeleton organization	1309	799	2.62E-18
GO:0031344	regulation of cell projection organization	467	326	7.18E-18
GO:0050804	modulation of chemical synaptic transmission	252	195	8.78E-18
GO:0120035	regulation of plasma membrane bounded cell projection organization	453	317	1.42E-17
GO:0071840	cellular component organization or biogenesis	5539	2962	1.45E-17
GO:0099177	regulation of trans-synaptic signaling	253	195	1.99E-17
GO:0048522	positive regulation of cellular process	4704	2544	2.87E-17
GO:0099536	synaptic signaling	522	356	5.07E-17
GO:0048523	negative regulation of cellular process	3896	2135	6.91E-17
GO:0034330	cell junction organization	518	353	8.99E-17
GO:0010975	regulation of neuron projection development	288	215	1.91E-16
GO:0022603	regulation of anatomical structure morphogenesis	700	456	2.19E-16
GO:0045595	regulation of cell differentiation	1129	693	2.71E-16
GO:0007417	central nervous system development	584	389	4.80E-16
GO:0050794	regulation of cellular process	9523	4886	6.17E-16
GO:0048518	positive regulation of biological process	5294	2827	8.88E-16
GO:0099537	trans-synaptic signaling	501	340	1.25E-15

GO:0098916	anterograde trans-synaptic signaling	495	336	1.93E-15
GO:0007268	chemical synaptic transmission	495	336	1.93E-15
GO:0007165	signal transduction	4776	2566	2.30E-15
GO:0009987	cellular process	14783	7306	2.42E-15
GO:0048583	regulation of response to stimulus	3327	1835	3.42E-15
GO:0097485	neuron projection guidance	169	137	5.04E-15
GO:0007411	axon guidance	169	137	5.04E-15
GO:0065007	biological regulation	10721	5447	5.19E-15
GO:0048519	negative regulation of biological process	4381	2365	6.60E-15
GO:0065008	regulation of biological quality	2937	1634	6.88E-15
GO:0032879	regulation of localization	1615	947	7.50E-15
GO:0051716	cellular response to stimulus	5982	3159	9.70E-15
GO:0050789	regulation of biological process	10085	5143	1.21E-14
GO:0007267	cell-cell signaling	1269	761	1.63E-14
GO:0051094	positive regulation of developmental process	967	595	6.68E-14
GO:0051179	localization	4343	2335	1.81E-13
GO:0050770	regulation of axonogenesis	100	88	2.65E-13
GO:0051049	regulation of transport	1327	786	3.15E-13
GO:0048870	cell motility	1362	804	4.26E-13
GO:0007167	enzyme-linked receptor protein signaling pathway	795	497	5.75E-13
GO:0009887	animal organ morphogenesis	582	378	7.07E-13
GO:0050808	synapse organization	275	198	3.52E-12
GO:0051960	regulation of nervous system development	257	187	3.93E-12
GO:0016477	cell migration	1210	717	7.09E-12
GO:0009888	tissue development	1239	732	8.83E-12
GO:0003012	muscle system process	305	215	8.94E-12
GO:0055085	transmembrane transport	1060	635	1.60E-11
GO:0051130	positive regulation of cellular component organization	844	518	1.69E-11
GO:0006810	transport	3620	1956	1.92E-11
GO:0065009	regulation of molecular function	2121	1192	2.01E-11
GO:0040011	locomotion	1075	641	4.38E-11
GO:0098660	inorganic ion transmembrane transport	622	394	5.04E-11
GO:0042391	regulation of membrane potential	319	221	7.21E-11
GO:0051234	establishment of localization	3775	2029	7.97E-11
GO:0007015	actin filament organization	395	265	8.57E-11
GO:0045944	positive regulation of transcription by RNA polymerase II	952	573	1.11E-10
GO:0007420	brain development	384	258	1.44E-10
GO:0098655	cation transmembrane transport	656	411	1.57E-10
GO:0023057	negative regulation of signaling	1067	634	1.61E-10
GO:0061061	muscle structure development	407	271	1.72E-10
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	511	330	1.77E-10
GO:0034220	ion transmembrane transport	816	498	2.25E-10
GO:0016310	phosphorylation	1444	833	2.28E-10
GO:1902531	regulation of intracellular signal transduction	1429	825	2.39E-10
GO:0007264	small GTPase mediated signal transduction	389	260	2.86E-10
GO:0010648	negative regulation of cell communication	1060	629	2.89E-10
GO:0050767	regulation of neurogenesis	208	153	3.26E-10
GO:0060322	head development	402	267	4.05E-10
GO:0040012	regulation of locomotion	842	511	4.05E-10
GO:0006936	muscle contraction	260	184	4.79E-10
GO:0022604	regulation of cell morphogenesis	240	172	4.91E-10
GO:0098662	inorganic cation transmembrane transport	572	362	6.98E-10
GO:0097435	supramolecular fiber organization	710	438	7.39E-10
GO:0006811	ion transport	1187	694	9.42E-10
GO:0030001	metal ion transport	678	420	9.61E-10
GO:0048167	regulation of synaptic plasticity	122	98	1.02E-09
GO:2000026	regulation of multicellular organismal development	981	584	1.04E-09
GO:0045597	positive regulation of cell differentiation	624	390	1.13E-09
GO:2000145	regulation of cell motility	818	496	1.14E-09
GO:0030334	regulation of cell migration	767	468	1.35E-09
GO:0048585	negative regulation of response to stimulus	1287	745	2.06E-09
GO:0072359	circulatory system development	729	446	2.71E-09
GO:1903508	positive regulation of nucleic acid-templated transcription	1297	749	3.44E-09
GO:0045893	positive regulation of DNA-templated transcription	1297	749	3.44E-09
GO:0051173	positive regulation of nitrogen compound metabolic process	2565	1404	3.55E-09

GO:0051962	positive regulation of nervous system development	147	113	3.93E-09
GO:0031346	positive regulation of cell projection organization	247	174	3.98E-09
GO:0006812	cation transport	886	530	4.38E-09
GO:0009968	negative regulation of signal transduction	1009	595	6.51E-09
GO:0032970	regulation of actin filament-based process	330	222	6.52E-09
GO:0006468	protein phosphorylation	1268	732	6.97E-09
GO:0071805	potassium ion transmembrane transport	185	136	7.67E-09
GO:1902680	positive regulation of RNA biosynthetic process	1303	750	8.22E-09
GO:0031325	positive regulation of cellular metabolic process	2522	1378	1.21E-08
GO:0044087	regulation of cellular component biogenesis	774	467	1.61E-08
GO:0044057	regulation of system process	392	256	1.94E-08
GO:0060560	developmental growth involved in morphogenesis	135	104	2.36E-08
GO:0048588	developmental cell growth	129	100	3.06E-08
GO:0032535	regulation of cellular component size	256	177	3.24E-08
GO:0060284	regulation of cell development	315	211	4.14E-08
GO:0051056	regulation of small GTPase mediated signal transduction	238	166	4.61E-08
GO:0098609	cell-cell adhesion	743	448	5.00E-08
GO:0031589	cell-substrate adhesion	290	196	6.49E-08
GO:0001667	ameboidal-type cell migration	333	220	1.12E-07
GO:0099587	inorganic ion import across plasma membrane	112	88	1.40E-07
GO:0098659	inorganic cation import across plasma membrane	112	88	1.40E-07
GO:0009893	positive regulation of metabolic process	3177	1700	1.57E-07
GO:0051254	positive regulation of RNA metabolic process	1433	810	1.66E-07
GO:0051240	positive regulation of multicellular organismal process	1137	655	1.76E-07
GO:0050769	positive regulation of neurogenesis	119	92	2.53E-07
GO:0010604	positive regulation of macromolecule metabolic process	2889	1553	2.57E-07
GO:0044093	positive regulation of molecular function	1305	742	2.64E-07
GO:0031324	negative regulation of cellular metabolic process	1858	1028	2.90E-07
GO:0060627	regulation of vesicle-mediated transport	417	266	2.97E-07
GO:0045935	positive regulation of nucleobase-containing compound metabolic process	1620	905	3.26E-07
GO:0006793	phosphorus metabolic process	2354	1281	3.36E-07
GO:0051641	cellular localization	2721	1467	3.45E-07
GO:0035725	sodium ion transmembrane transport	129	98	3.60E-07
GO:0006813	potassium ion transport	200	141	4.40E-07
GO:0033043	regulation of organelle organization	1015	588	5.17E-07
GO:1990138	neuron projection extension	112	87	5.18E-07
GO:0048638	regulation of developmental growth	167	121	5.36E-07
GO:0006796	phosphate-containing compound metabolic process	2336	1270	5.65E-07
GO:0042221	response to chemical	2943	1577	5.88E-07
GO:0001508	action potential	117	90	6.38E-07
GO:0034329	cell junction assembly	341	222	6.89E-07
GO:0023056	positive regulation of signaling	1380	778	7.95E-07
GO:0032956	regulation of actin cytoskeleton organization	296	196	9.18E-07
GO:0030516	regulation of axon extension	61	53	9.20E-07
GO:1905114	cell surface receptor signaling pathway involved in cell-cell signaling	388	248	9.30E-07
GO:0003015	heart process	193	136	9.57E-07
GO:0048646	anatomical structure formation involved in morphogenesis	751	446	1.03E-06
GO:0016049	cell growth	358	231	1.03E-06
GO:0010647	positive regulation of cell communication	1374	774	1.07E-06
GO:0006996	organelle organization	3147	1677	1.11E-06
GO:0061387	regulation of extent of cell growth	67	57	1.13E-06
GO:0051093	negative regulation of developmental process	619	374	1.58E-06
GO:0007507	heart development	338	219	1.67E-06
GO:0048589	developmental growth	273	182	1.68E-06
GO:0003013	circulatory system process	443	277	2.44E-06
GO:0042127	regulation of cell population proliferation	1218	690	2.85E-06
GO:0071495	cellular response to endogenous stimulus	974	562	2.94E-06
GO:0007265	Ras protein signal transduction	278	184	3.20E-06
GO:0060047	heart contraction	187	131	3.55E-06
GO:0009719	response to endogenous stimulus	1076	615	3.74E-06
GO:0060828	regulation of canonical Wnt signaling pathway	211	145	3.96E-06
GO:0051493	regulation of cytoskeleton organization	450	280	4.00E-06
GO:0090066	regulation of anatomical structure size	328	212	4.16E-06
GO:0031327	negative regulation of cellular biosynthetic process	1292	727	4.83E-06

GO:0045892	negative regulation of DNA-templated transcription	1051	601	5.22E-06
GO:0003008	system process	1358	761	5.26E-06
GO:0006814	sodium ion transport	166	118	5.64E-06
GO:0010557	positive regulation of macromolecule biosynthetic process	1492	830	5.80E-06
GO:0009891	positive regulation of biosynthetic process	1593	882	5.89E-06
GO:0086001	cardiac muscle cell action potential	73	60	6.04E-06
GO:0051172	negative regulation of nitrogen compound metabolic process	1927	1053	6.26E-06
GO:1903507	negative regulation of nucleic acid-templated transcription	1056	603	6.56E-06
GO:0009890	negative regulation of biosynthetic process	1315	738	6.62E-06
GO:1902679	negative regulation of RNA biosynthetic process	1057	603	7.97E-06
GO:0098657	import into cell	216	147	8.46E-06
GO:0048675	axon extension	81	65	8.67E-06
GO:0050772	positive regulation of axonogenesis	47	42	9.25E-06
GO:0060537	muscle tissue development	218	148	9.60E-06
GO:0050807	regulation of synapse organization	124	92	9.86E-06
GO:0008283	cell population proliferation	1383	772	9.86E-06
GO:0001558	regulation of cell growth	320	206	1.11E-05
GO:0031328	positive regulation of cellular biosynthetic process	1568	867	1.13E-05
GO:0009790	embryo development	505	308	1.26E-05
GO:0007166	cell surface receptor signaling pathway	2271	1225	1.38E-05
GO:0035637	multicellular organismal signaling	128	94	1.60E-05
GO:0010631	epithelial cell migration	261	172	1.68E-05
GO:0060070	canonical Wnt signaling pathway	254	168	1.68E-05
GO:0010632	regulation of epithelial cell migration	204	139	1.92E-05
GO:0045596	negative regulation of cell differentiation	439	271	2.07E-05
GO:0050896	response to stimulus	7117	3620	2.09E-05
GO:0030111	regulation of Wnt signaling pathway	274	179	2.22E-05
GO:0010558	negative regulation of macromolecule biosynthetic process	1251	701	2.32E-05
GO:0098739	import across plasma membrane	172	120	2.36E-05
GO:0034762	regulation of transmembrane transport	385	241	2.39E-05
GO:0040007	growth	515	312	2.61E-05
GO:0071310	cellular response to organic substance	1734	949	2.81E-05
GO:0050803	regulation of synapse structure or activity	129	94	3.02E-05
GO:0000165	MAPK cascade	599	357	3.11E-05
GO:0014706	striated muscle tissue development	144	103	3.19E-05
GO:0035295	tube development	616	366	3.29E-05
GO:0090132	epithelium migration	263	172	3.75E-05
GO:0050806	positive regulation of synaptic transmission	89	69	3.90E-05
GO:0009892	negative regulation of metabolic process	2418	1295	4.08E-05
GO:0010033	response to organic substance	2106	1137	4.17E-05
GO:0046777	protein autophosphorylation	199	135	4.63E-05
GO:0048738	cardiac muscle tissue development	138	99	4.67E-05
GO:0001505	regulation of neurotransmitter levels	143	102	4.68E-05
GO:0061337	cardiac conduction	91	70	5.39E-05
GO:0016358	dendrite development	155	109	5.39E-05
GO:0090257	regulation of muscle system process	157	110	6.38E-05
GO:0034765	regulation of ion transmembrane transport	325	206	6.51E-05
GO:0051253	negative regulation of RNA metabolic process	1157	649	6.64E-05
GO:0006941	striated muscle contraction	142	101	6.81E-05
GO:0008361	regulation of cell size	119	87	7.72E-05
GO:0030155	regulation of cell adhesion	614	363	7.77E-05
GO:0035239	tube morphogenesis	547	327	7.90E-05
GO:0090130	tissue migration	267	173	8.81E-05
GO:0036211	protein modification process	2994	1581	9.04E-05
GO:0099003	vesicle-mediated transport in synapse	121	88	9.62E-05
GO:0050905	neuromuscular process	73	58	1.07E-04
GO:0030048	actin filament-based movement	113	83	1.10E-04
GO:0042692	muscle cell differentiation	243	159	1.16E-04
GO:0022607	cellular component assembly	2547	1355	1.30E-04
GO:0030900	forebrain development	155	108	1.31E-04
GO:0060078	regulation of postsynaptic membrane potential	61	50	1.38E-04
GO:0007517	muscle organ development	179	122	1.38E-04
GO:0009967	positive regulation of signal transduction	1255	697	1.57E-04
GO:0048639	positive regulation of developmental growth	83	64	1.74E-04
GO:0048729	tissue morphogenesis	348	217	1.83E-04
GO:0070887	cellular response to chemical stimulus	2288	1223	1.83E-04

GO:0043269	regulation of ion transport	458	277	1.97E-04
GO:0035249	synaptic transmission, glutamatergic	66	53	2.05E-04
GO:0099504	synaptic vesicle cycle	114	83	2.07E-04
GO:0040008	regulation of growth	416	254	2.14E-04
GO:0051129	negative regulation of cellular component organization	566	335	2.15E-04
GO:0021537	telencephalon development	101	75	2.19E-04
GO:0098900	regulation of action potential	48	41	2.21E-04
GO:0045934	negative regulation of nucleobase-containing compound metabolic process	1273	705	2.32E-04
GO:0099565	chemical synaptic transmission, postsynaptic	51	43	2.34E-04
GO:0051241	negative regulation of multicellular organismal process	755	435	2.53E-04
GO:0010720	positive regulation of cell development	177	120	2.71E-04
GO:0140352	export from cell	627	367	2.74E-04
GO:1902532	negative regulation of intracellular signal transduction	436	264	3.33E-04
GO:0032409	regulation of transporter activity	244	158	3.37E-04
GO:0010605	negative regulation of macromolecule metabolic process	2244	1198	3.49E-04
GO:0008016	regulation of heart contraction	164	112	3.83E-04
GO:1904062	regulation of cation transmembrane transport	291	184	4.13E-04
GO:0070727	cellular macromolecule localization	1880	1013	4.24E-04
GO:0010594	regulation of endothelial cell migration	154	106	4.36E-04
GO:0043542	endothelial cell migration	194	129	5.00E-04
GO:0048598	embryonic morphogenesis	308	193	5.37E-04
GO:0008104	protein localization	1874	1009	5.38E-04
GO:0044089	positive regulation of cellular component biogenesis	416	252	5.87E-04
GO:0050790	regulation of catalytic activity	1498	817	6.20E-04
GO:0198738	cell-cell signaling by wnt	343	212	6.32E-04
GO:0040017	positive regulation of locomotion	472	282	6.67E-04
GO:0060048	cardiac muscle contraction	111	80	6.71E-04
GO:0008154	actin polymerization or depolymerization	155	106	7.08E-04
GO:1903522	regulation of blood circulation	188	125	7.51E-04
GO:0030335	positive regulation of cell migration	443	266	7.90E-04
GO:0051247	positive regulation of protein metabolic process	1264	696	8.30E-04
GO:0021953	central nervous system neuron differentiation	74	57	8.76E-04
GO:0030100	regulation of endocytosis	159	108	9.23E-04
GO:0016055	Wnt signaling pathway	339	209	9.76E-04
GO:0006836	neurotransmitter transport	137	95	1.04E-03
GO:0016192	vesicle-mediated transport	1326	727	1.05E-03
GO:2000147	positive regulation of cell motility	463	276	1.13E-03
GO:0010243	response to organonitrogen compound	613	356	1.16E-03
GO:0007163	establishment or maintenance of cell polarity	177	118	1.25E-03
GO:0030178	negative regulation of Wnt signaling pathway	146	100	1.30E-03
GO:0099173	postsynapse organization	99	72	1.41E-03
GO:0006937	regulation of muscle contraction	119	84	1.42E-03
GO:0007229	integrin-mediated signaling pathway	104	75	1.44E-03
GO:0070252	actin-mediated cell contraction	86	64	1.50E-03
GO:0008360	regulation of cell shape	116	82	1.72E-03
GO:0008015	blood circulation	363	221	1.74E-03
GO:0060429	epithelium development	689	395	1.79E-03
GO:0043087	regulation of GTPase activity	301	187	1.79E-03
GO:0032412	regulation of ion transmembrane transporter activity	222	143	1.84E-03
GO:0018193	peptidyl-amino acid modification	1035	575	1.97E-03
GO:0055001	muscle cell development	118	83	2.04E-03
GO:0070848	response to growth factor	516	303	2.08E-03
GO:0043085	positive regulation of catalytic activity	960	536	2.08E-03
GO:0031623	receptor internalization	98	71	2.09E-03
GO:0051050	positive regulation of transport	709	405	2.15E-03
GO:0010977	negative regulation of neuron projection development	85	63	2.27E-03
GO:0031345	negative regulation of cell projection organization	125	87	2.31E-03
GO:1901699	cellular response to nitrogen compound	468	277	2.40E-03
GO:0086065	cell communication involved in cardiac conduction	58	46	2.49E-03
GO:0031532	actin cytoskeleton reorganization	105	75	2.56E-03
GO:0051146	striated muscle cell differentiation	170	113	2.56E-03
GO:0051649	establishment of localization in cell	1698	913	2.68E-03
GO:0046903	secretion	638	367	2.72E-03
GO:0046578	regulation of Ras protein signal transduction	139	95	2.74E-03

GO:0031399	regulation of protein modification process	1259	689	2.83E-03
GO:0071417	cellular response to organonitrogen compound	413	247	2.92E-03
GO:0048813	dendrite morphogenesis	97	70	3.06E-03
GO:0060291	long-term synaptic potentiation	49	40	3.09E-03
GO:0071526	semaphorin-plexin signaling pathway	40	34	3.13E-03
GO:0050771	negative regulation of axonogenesis	43	36	3.22E-03
GO:0022898	regulation of transmembrane transporter activity	229	146	3.28E-03
GO:0032940	secretion by cell	571	331	3.39E-03
GO:0051174	regulation of phosphorus metabolic process	1126	620	3.43E-03
GO:0042325	regulation of phosphorylation	1004	557	3.54E-03
GO:0098901	regulation of cardiac muscle cell action potential	30	27	3.62E-03
GO:0040013	negative regulation of locomotion	278	173	3.69E-03
GO:0007416	synapse assembly	126	87	3.77E-03
GO:0019220	regulation of phosphate metabolic process	1125	619	3.93E-03
GO:0043408	regulation of MAPK cascade	534	311	3.94E-03
GO:0045229	external encapsulating structure organization	237	150	4.35E-03
GO:0051966	regulation of synaptic transmission, glutamatergic	51	41	4.76E-03
GO:0006898	receptor-mediated endocytosis	214	137	4.77E-03
GO:0086003	cardiac muscle cell contraction	62	48	5.00E-03
GO:0098703	calcium ion import across plasma membrane	36	31	5.04E-03
GO:0086002	cardiac muscle cell action potential involved in contraction	48	39	5.09E-03
GO:0033036	macromolecule localization	2245	1186	5.15E-03
GO:0001944	vasculature development	492	288	5.17E-03
GO:0050890	cognition	156	104	5.19E-03
GO:0030198	extracellular matrix organization	234	148	5.31E-03
GO:0060079	excitatory postsynaptic potential	45	37	5.32E-03
GO:1901888	regulation of cell junction assembly	144	97	5.54E-03
GO:0050678	regulation of epithelial cell proliferation	236	149	5.67E-03
GO:1902903	regulation of supramolecular fiber organization	311	190	6.64E-03
GO:0043062	extracellular structure organization	235	148	7.38E-03
GO:0002009	morphogenesis of an epithelium	275	170	7.74E-03
GO:0042592	homeostatic process	1249	680	7.85E-03
GO:0043412	macromolecule modification	3198	1659	8.65E-03
GO:0086091	regulation of heart rate by cardiac conduction	41	34	9.11E-03
GO:0110053	regulation of actin filament organization	216	137	9.48E-03
GO:0031400	negative regulation of protein modification process	409	242	1.05E-02
GO:0001568	blood vessel development	469	274	1.06E-02
GO:0090090	negative regulation of canonical Wnt signaling pathway	118	81	1.13E-02
GO:0048592	eye morphogenesis	81	59	1.14E-02
GO:0051246	regulation of protein metabolic process	2072	1095	1.17E-02
GO:0007423	sensory organ development	271	167	1.19E-02
GO:2001257	regulation of cation channel activity	153	101	1.31E-02
GO:0070588	calcium ion transmembrane transport	235	147	1.31E-02
GO:0007611	learning or memory	115	79	1.37E-02
GO:0048863	stem cell differentiation	155	102	1.44E-02
GO:0031401	positive regulation of protein modification process	814	454	1.44E-02
GO:0060341	regulation of cellular localization	766	429	1.48E-02
GO:0021954	central nervous system neuron development	40	33	1.51E-02
GO:0071363	cellular response to growth factor stimulus	499	289	1.51E-02
GO:1901379	regulation of potassium ion transmembrane transport	70	52	1.53E-02
GO:0030041	actin filament polymerization	129	87	1.53E-02
GO:0048514	blood vessel morphogenesis	439	257	1.60E-02
GO:0050773	regulation of dendrite development	62	47	1.62E-02
GO:0009792	embryo development ending in birth or egg hatching	198	126	1.68E-02
GO:0051961	negative regulation of nervous system development	85	61	1.69E-02
GO:0090596	sensory organ morphogenesis	119	81	1.79E-02
GO:0001501	skeletal system development	289	176	1.82E-02
GO:0120031	plasma membrane bounded cell projection assembly	483	280	1.85E-02
GO:0006816	calcium ion transport	313	189	1.85E-02
GO:0010634	positive regulation of epithelial cell migration	133	89	1.92E-02
GO:0031098	stress-activated protein kinase signaling cascade	202	128	1.93E-02
GO:0009628	response to abiotic stimulus	712	400	1.97E-02
GO:1990573	potassium ion import across plasma membrane	48	38	1.99E-02
GO:0150104	transport across blood-brain barrier	87	62	2.03E-02
GO:0010232	vascular transport	87	62	2.03E-02
GO:0007160	cell-matrix adhesion	195	124	2.04E-02
GO:0032880	regulation of protein localization	672	379	2.06E-02
GO:0043254	regulation of protein-containing complex assembly	332	199	2.11E-02

GO:0051258	protein polymerization	233	145	2.14E-02
GO:0000122	negative regulation of transcription by RNA polymerase II	751	420	2.14E-02
GO:0007612	learning	53	41	2.48E-02
GO:0045664	regulation of neuron differentiation	96	67	2.77E-02
GO:0010959	regulation of metal ion transport	298	180	2.80E-02
GO:0030010	establishment of cell polarity	108	74	2.86E-02
GO:0007269	neurotransmitter secretion	91	64	2.87E-02
GO:0099643	signal release from synapse	91	64	2.87E-02
GO:0051963	regulation of synapse assembly	58	44	2.89E-02
GO:0016079	synaptic vesicle exocytosis	58	44	2.89E-02
GO:0150063	visual system development	207	130	3.03E-02
GO:0051403	stress-activated MAPK cascade	198	125	3.03E-02
GO:0061572	actin filament bundle organization	136	90	3.22E-02
GO:0048762	mesenchymal cell differentiation	191	121	3.23E-02
GO:0007215	glutamate receptor signaling pathway	44	35	3.51E-02
GO:0086009	membrane repolarization	44	35	3.51E-02
GO:0048880	sensory system development	213	133	3.64E-02
GO:0032878	regulation of establishment or maintenance of cell polarity	26	23	3.65E-02
GO:0001654	eye development	204	128	3.68E-02
GO:0048640	negative regulation of developmental growth	60	45	3.78E-02
GO:0030031	cell projection assembly	498	286	3.80E-02
GO:0043549	regulation of kinase activity	599	339	3.81E-02
GO:0050881	musculoskeletal movement	41	33	3.82E-02
GO:0050879	multicellular organismal movement	41	33	3.82E-02
GO:1990778	protein localization to cell periphery	279	169	3.89E-02
GO:0051017	actin filament bundle assembly	133	88	3.94E-02
GO:0086005	ventricular cardiac muscle cell action potential	29	25	4.10E-02
GO:0050919	negative chemotaxis	35	29	4.26E-02
GO:0030517	negative regulation of axon extension	32	27	4.29E-02
GO:0048878	chemical homeostasis	769	427	4.30E-02
GO:0042327	positive regulation of phosphorylation	663	372	4.31E-02
GO:0050768	negative regulation of neurogenesis	80	57	4.32E-02
GO:0031323	regulation of cellular metabolic process	4966	2520	4.38E-02
GO:0043009	chordate embryonic development	183	116	4.45E-02
GO:0010721	negative regulation of cell development	109	74	4.53E-02
GO:0045785	positive regulation of cell adhesion	369	217	4.75E-02
GO:0006887	exocytosis	267	162	4.91E-02
GO:0045216	cell-cell junction organization	169	108	4.96E-02
GO:0001764	neuron migration	87	61	4.98E-02

Appendix Table B 6 Significant GO:BP enrichments for all CLINVAR G4 mutations.

GO ID	GO Description	Universe	CLINVAR	Adjusted P-value
GO:0006941	striated muscle contraction	142	32	1.1E-11
GO:0060048	cardiac muscle contraction	111	27	2.1E-10
GO:0050905	neuromuscular process	73	22	4.5E-10
GO:0035637	multicellular organismal signaling	128	28	1.1E-09
GO:0001508	action potential	117	26	5.2E-09
GO:0030048	actin filament-based movement	113	25	1.5E-08
GO:0070252	actin-mediated cell contraction	86	20	7.0E-07
GO:0061337	cardiac conduction	91	20	1.9E-06
GO:0086001	cardiac muscle cell action potential	73	18	2.0E-06
GO:0006937	regulation of muscle contraction	119	22	7.3E-06
GO:0086003	cardiac muscle cell contraction	62	15	5.8E-05
GO:0051899	membrane depolarization	64	15	8.9E-05
GO:1903115	regulation of actin filament-based movement	32	11	9.0E-05
GO:0060415	muscle tissue morphogenesis	47	13	9.1E-05
GO:0048644	muscle organ morphogenesis	47	13	9.1E-05
GO:0002027	regulation of heart rate	84	17	1.1E-04
GO:0098900	regulation of action potential	48	13	1.2E-04
GO:0086002	cardiac muscle cell action potential involved in contraction	48	13	1.2E-04
GO:0050954	sensory perception of mechanical stimulus	106	18	5.7E-04
GO:0086065	cell communication involved in cardiac conduction	58	13	1.0E-03
GO:0006942	regulation of striated muscle contraction	78	15	1.1E-03
GO:0007605	sensory perception of sound	100	17	1.1E-03
GO:0086091	regulation of heart rate by cardiac conduction	41	11	1.2E-03
GO:0050881	musculoskeletal movement	41	11	1.2E-03
GO:0050879	multicellular organismal movement	41	11	1.2E-03
GO:0019226	transmission of nerve impulse	42	11	1.5E-03
GO:0055008	cardiac muscle tissue morphogenesis	42	11	1.5E-03
GO:0048738	cardiac muscle tissue development	138	20	1.5E-03
GO:0055001	muscle cell development	118	18	2.5E-03
GO:0003009	skeletal muscle contraction	36	10	2.5E-03
GO:0014706	striated muscle tissue development	144	20	2.8E-03
GO:0086005	ventricular cardiac muscle cell action potential	29	9	3.1E-03
GO:0050885	neuromuscular process controlling balance	16	7	3.4E-03
GO:0098901	regulation of cardiac muscle cell action potential	30	9	4.2E-03
GO:0003229	ventricular cardiac muscle tissue development	39	10	5.3E-03
GO:0048592	eye morphogenesis	81	14	8.1E-03
GO:0055010	ventricular cardiac muscle tissue morphogenesis	33	9	9.3E-03
GO:0055117	regulation of cardiac muscle contraction	61	12	9.8E-03
GO:0086019	cell-cell signaling involved in cardiac conduction	34	9	1.2E-02
GO:0008306	associative learning	19	7	1.2E-02
GO:0086010	membrane depolarization during action potential	35	9	1.5E-02
GO:0050953	sensory perception of light stimulus	138	18	1.9E-02
GO:0086004	regulation of cardiac muscle cell contraction	28	8	2.0E-02
GO:0043589	skin morphogenesis	5	4	2.6E-02
GO:0002028	regulation of sodium ion transport	68	12	2.7E-02
GO:0048661	positive regulation of smooth muscle cell proliferation	49	10	3.6E-02
GO:0048483	autonomic nervous system development	31	8	4.1E-02
GO:0060348	bone development	108	15	4.5E-02

Appendix Table B 7 Significant GO:BP enrichments for COSMIC and CLINVAR G4 mutations leading to the loss of a G4.

GO ID	GO Description	Universe	COSMIC and CLINVAR	Adjusted P-value
GO:0032502	developmental process	4584	1138	3.23E-24
GO:0048856	anatomical structure development	4152	1043	2.82E-23
GO:0007399	nervous system development	1648	488	1.15E-22
GO:0009653	anatomical structure morphogenesis	1871	539	2.18E-22
GO:0048731	system development	2976	784	9.30E-22
GO:0007275	multicellular organism development	3266	842	1.00E-20
GO:0032501	multicellular organismal process	5319	1270	1.40E-20
GO:0030154	cell differentiation	2860	742	3.10E-18
GO:0048699	generation of neurons	969	309	3.31E-18
GO:0048869	cellular developmental process	2879	745	5.24E-18
GO:0016043	cellular component organization	5370	1259	6.78E-17
GO:0022008	neurogenesis	1096	335	1.06E-16
GO:0030182	neuron differentiation	923	292	1.76E-16
GO:0071840	cellular component organization or biogenesis	5539	1277	2.38E-14
GO:0023051	regulation of signaling	2733	694	3.06E-14
GO:0010646	regulation of cell communication	2727	692	4.16E-14
GO:0048666	neuron development	729	236	4.22E-14
GO:0000904	cell morphogenesis involved in differentiation	495	175	7.97E-14
GO:0048468	cell development	1355	385	9.16E-14
GO:0048513	animal organ development	2176	567	3.84E-13
GO:0023052	signaling	5190	1196	9.50E-13
GO:0032989	cellular component morphogenesis	549	186	1.02E-12
GO:0007154	cell communication	5221	1200	2.02E-12
GO:0031175	neuron projection development	656	212	2.53E-12
GO:0048667	cell morphogenesis involved in neuron differentiation	381	140	3.85E-12
GO:0034330	cell junction organization	518	176	4.81E-12
GO:0000902	cell morphogenesis	718	226	6.91E-12
GO:0048812	neuron projection morphogenesis	441	155	9.70E-12
GO:0007010	cytoskeleton organization	1309	365	1.42E-11
GO:0009966	regulation of signal transduction	2464	621	1.59E-11
GO:0009887	animal organ morphogenesis	582	190	2.87E-11
GO:0007155	cell adhesion	1216	342	3.26E-11
GO:0120036	plasma membrane bounded cell projection organization	1144	325	3.84E-11
GO:0120039	plasma membrane bounded cell projection morphogenesis	455	157	3.85E-11
GO:0048858	cell projection morphogenesis	459	157	8.90E-11
GO:0030030	cell projection organization	1164	328	9.33E-11
GO:0032990	cell part morphogenesis	469	159	1.42E-10
GO:0030029	actin filament-based process	717	219	5.92E-10
GO:0061564	axon development	323	118	1.06E-09
GO:0007409	axonogenesis	298	111	1.25E-09
GO:0050808	synapse organization	275	104	2.31E-09
GO:0050793	regulation of developmental process	1810	465	4.44E-09
GO:0051128	regulation of cellular component organization	1929	490	6.77E-09
GO:0007165	signal transduction	4776	1078	4.27E-08
GO:0016477	cell migration	1210	325	7.79E-08
GO:0048870	cell motility	1362	359	8.30E-08
GO:0098609	cell-cell adhesion	743	217	8.39E-08
GO:0051716	cellular response to stimulus	5982	1316	9.92E-08
GO:0035556	intracellular signal transduction	2168	534	1.26E-07
GO:0048583	regulation of response to stimulus	3327	777	1.74E-07
GO:0051239	regulation of multicellular organismal process	2117	522	1.84E-07
GO:0099537	trans-synaptic signaling	501	157	2.02E-07
GO:0099536	synaptic signaling	522	162	2.28E-07
GO:0009888	tissue development	1239	329	2.43E-07
GO:0098916	anterograde trans-synaptic signaling	495	155	2.81E-07
GO:0007268	chemical synaptic transmission	495	155	2.81E-07
GO:0065007	biological regulation	10721	2221	3.01E-07
GO:0072359	circulatory system development	729	211	3.69E-07
GO:0065008	regulation of biological quality	2937	692	5.65E-07
GO:0050794	regulation of cellular process	9523	1993	7.40E-07
GO:0007267	cell-cell signaling	1269	333	7.84E-07
GO:0003008	system process	1358	351	1.64E-06
GO:0007417	central nervous system development	584	173	2.68E-06
GO:0009987	cellular process	14783	2936	4.03E-06

GO:0040011	locomotion	1075	285	5.65E-06
GO:0050789	regulation of biological process	10085	2089	8.77E-06
GO:1905114	cell surface receptor signaling pathway involved in cell-cell signaling	388	123	1.03E-05
GO:0030036	actin cytoskeleton organization	637	183	1.15E-05
GO:0050804	modulation of chemical synaptic transmission	252	88	1.18E-05
GO:0030048	actin filament-based movement	113	49	1.44E-05
GO:0099177	regulation of trans-synaptic signaling	253	88	1.46E-05
GO:0006812	cation transport	886	240	1.55E-05
GO:0030001	metal ion transport	678	192	1.58E-05
GO:0010975	regulation of neuron projection development	288	97	1.61E-05
GO:0060047	heart contraction	187	70	1.80E-05
GO:0006811	ion transport	1187	307	1.99E-05
GO:0048518	positive regulation of biological process	5294	1158	2.05E-05
GO:0034329	cell junction assembly	341	110	2.35E-05
GO:0032879	regulation of localization	1615	400	2.54E-05
GO:0044057	regulation of system process	392	122	3.78E-05
GO:0007411	axon guidance	169	64	4.59E-05
GO:0097485	neuron projection guidance	169	64	4.59E-05
GO:0051179	localization	4343	964	4.59E-05
GO:0048522	positive regulation of cellular process	4704	1036	5.30E-05
GO:0007507	heart development	338	108	5.46E-05
GO:0006936	muscle contraction	260	88	6.00E-05
GO:0055085	transmembrane transport	1060	276	6.09E-05
GO:0034220	ion transmembrane transport	816	221	6.22E-05
GO:0042391	regulation of membrane potential	319	103	6.34E-05
GO:0048523	negative regulation of cellular process	3896	872	6.65E-05
GO:0003015	heart process	193	70	7.53E-05
GO:0007166	cell surface receptor signaling pathway	2271	536	8.29E-05
GO:0003012	muscle system process	305	99	8.73E-05
GO:1902531	regulation of intracellular signal transduction	1429	356	9.28E-05
GO:0061061	muscle structure development	407	124	1.07E-04
GO:0040012	regulation of locomotion	842	225	1.46E-04
GO:0098655	cation transmembrane transport	656	182	1.81E-04
GO:0048646	anatomical structure formation involved in morphogenesis	751	203	2.65E-04
GO:0048519	negative regulation of biological process	4381	964	2.73E-04
GO:0048729	tissue morphogenesis	348	108	2.78E-04
GO:0007416	synapse assembly	126	50	2.85E-04
GO:0022603	regulation of anatomical structure morphogenesis	700	191	3.10E-04
GO:0051960	regulation of nervous system development	257	85	3.10E-04
GO:0030334	regulation of cell migration	767	206	3.51E-04
GO:0050905	neuromuscular process	73	34	3.51E-04
GO:0031589	cell-substrate adhesion	290	93	3.99E-04
GO:0050896	response to stimulus	7117	1503	4.02E-04
GO:0031344	regulation of cell projection organization	467	136	4.37E-04
GO:0003013	circulatory system process	443	130	5.12E-04
GO:0065009	regulation of molecular function	2121	498	5.42E-04
GO:0006810	transport	3620	807	5.65E-04
GO:0120035	regulation of plasma membrane bounded cell projection organization	453	132	6.34E-04
GO:0008015	blood circulation	363	110	7.79E-04
GO:0030111	regulation of Wnt signaling pathway	274	88	7.79E-04
GO:0007420	brain development	384	115	8.24E-04
GO:0051130	positive regulation of cellular component organization	844	221	9.67E-04
GO:2000145	regulation of cell motility	818	215	1.04E-03
GO:0006996	organelle organization	3147	708	1.04E-03
GO:0070252	actin-mediated cell contraction	86	37	1.15E-03
GO:0051056	regulation of small GTPase mediated signal transduction	238	78	1.44E-03
GO:0050877	nervous system process	755	200	1.46E-03
GO:0051234	establishment of localization	3775	834	1.55E-03
GO:0060048	cardiac muscle contraction	111	44	1.57E-03
GO:0051094	positive regulation of developmental process	967	247	1.64E-03
GO:0098662	inorganic cation transmembrane transport	572	158	1.65E-03
GO:0051963	regulation of synapse assembly	58	28	1.69E-03
GO:0060322	head development	402	118	1.74E-03
GO:0001667	ameboidal-type cell migration	333	101	2.14E-03
GO:0048638	regulation of developmental growth	167	59	2.18E-03
GO:0007167	enzyme-linked receptor protein signaling pathway	795	208	2.19E-03
GO:0008016	regulation of heart contraction	164	58	2.57E-03
GO:0051049	regulation of transport	1327	324	2.69E-03

GO:0006941	striated muscle contraction	142	52	2.76E-03
GO:0045595	regulation of cell differentiation	1129	281	2.82E-03
GO:0097435	supramolecular fiber organization	710	188	3.17E-03
GO:0035637	multicellular organismal signaling	128	48	3.20E-03
GO:0098660	inorganic ion transmembrane transport	622	168	3.22E-03
GO:0001505	regulation of neurotransmitter levels	143	52	3.47E-03
GO:0048598	embryonic morphogenesis	308	94	3.66E-03
GO:0010647	positive regulation of cell communication	1374	333	3.80E-03
GO:0006816	calcium ion transport	313	95	4.16E-03
GO:0009967	positive regulation of signal transduction	1255	307	4.32E-03
GO:0099565	chemical synaptic transmission, postsynaptic	51	25	4.41E-03
GO:0009719	response to endogenous stimulus	1076	268	4.62E-03
GO:0099587	inorganic ion import across plasma membrane	112	43	5.31E-03
GO:0098659	inorganic cation import across plasma membrane	112	43	5.31E-03
GO:0009790	embryo development	505	140	5.37E-03
GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	195	65	5.93E-03
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	511	141	6.44E-03
GO:0023057	negative regulation of signaling	1067	265	6.57E-03
GO:1903522	regulation of blood circulation	188	63	6.84E-03
GO:0050807	regulation of synapse organization	124	46	7.05E-03
GO:0001508	action potential	117	44	7.54E-03
GO:0023056	positive regulation of signaling	1380	332	7.60E-03
GO:0051962	positive regulation of nervous system development	147	52	8.38E-03
GO:0198738	cell-cell signaling by wnt	343	101	8.47E-03
GO:0010631	epithelial cell migration	261	81	8.86E-03
GO:0043269	regulation of ion transport	458	128	9.01E-03
GO:0016358	dendrite development	155	54	9.18E-03
GO:0050803	regulation of synapse structure or activity	129	47	9.61E-03
GO:1903115	regulation of actin filament-based movement	32	18	9.75E-03
GO:1901888	regulation of cell junction assembly	144	51	9.80E-03
GO:0040008	regulation of growth	416	118	9.86E-03
GO:0007612	learning	53	25	1.00E-02
GO:0022607	cellular component assembly	2547	575	1.03E-02
GO:0010648	negative regulation of cell communication	1060	262	1.06E-02
GO:0042221	response to chemical	2943	656	1.07E-02
GO:0086001	cardiac muscle cell action potential	73	31	1.16E-02
GO:0071495	cellular response to endogenous stimulus	974	243	1.17E-02
GO:0090130	tissue migration	267	82	1.17E-02
GO:0090132	epithelium migration	263	81	1.19E-02
GO:0090596	sensory organ morphogenesis	119	44	1.21E-02
GO:0048738	cardiac muscle tissue development	138	49	1.34E-02
GO:0007269	neurotransmitter secretion	91	36	1.45E-02
GO:0099643	signal release from synapse	91	36	1.45E-02
GO:0016055	Wnt signaling pathway	339	99	1.47E-02
GO:0040007	growth	515	140	1.52E-02
GO:0048589	developmental growth	273	83	1.55E-02
GO:0098703	calcium ion import across plasma membrane	36	19	1.76E-02
GO:0099003	vesicle-mediated transport in synapse	121	44	1.91E-02
GO:0014706	striated muscle tissue development	144	50	2.13E-02
GO:0016310	phosphorylation	1444	342	2.13E-02
GO:0007517	muscle organ development	179	59	2.24E-02
GO:0034762	regulation of transmembrane transport	385	109	2.30E-02
GO:0006836	neurotransmitter transport	137	48	2.39E-02
GO:0044087	regulation of cellular component biogenesis	774	197	2.41E-02
GO:0086003	cardiac muscle cell contraction	62	27	2.50E-02
GO:0007611	learning or memory	115	42	2.61E-02
GO:0009968	negative regulation of signal transduction	1009	248	2.74E-02
GO:0090257	regulation of muscle system process	157	53	2.86E-02
GO:0035249	synaptic transmission, glutamatergic	66	28	3.09E-02
GO:0060560	developmental growth involved in morphogenesis	135	47	3.44E-02
GO:0002009	morphogenesis of an epithelium	275	82	3.61E-02
GO:0007264	small GTPase mediated signal transduction	389	109	3.61E-02
GO:0002027	regulation of heart rate	84	33	3.80E-02
GO:0043542	endothelial cell migration	194	62	3.82E-02
GO:0051668	localization within membrane	570	150	4.30E-02
GO:0086091	regulation of heart rate by cardiac conduction	41	20	4.31E-02
GO:0060828	regulation of canonical Wnt signaling pathway	211	66	4.44E-02
GO:0051965	positive regulation of synapse assembly	35	18	4.46E-02

GO:0048588	developmental cell growth	129	45	4.73E-02
GO:0099504	synaptic vesicle cycle	114	41	4.75E-02
GO:0042127	regulation of cell population proliferation	1218	291	4.82E-02
GO:0050890	cognition	156	52	4.83E-02
GO:0007158	neuron cell-cell adhesion	16	11	4.94E-02

Appendix Table B 8 Significant GO:BP enrichments for COSMIC G4 mutations leading to the loss of a G4.

GO ID	GO Description	Universe	COSMIC	Adjusted P-value
GO:0009653	anatomical structure morphogenesis	1871	381	1.50E-10
GO:0016043	cellular component organization	5370	931	4.07E-10
GO:0032502	developmental process	4584	804	5.37E-09
GO:0071840	cellular component organization or biogenesis	5539	945	1.77E-08
GO:0048856	anatomical structure development	4152	734	1.85E-08
GO:0048699	generation of neurons	969	216	1.86E-08
GO:0007399	nervous system development	1648	333	2.19E-08
GO:0030182	neuron differentiation	923	204	1.65E-07
GO:0032501	multicellular organismal process	5319	904	2.10E-07
GO:0048731	system development	2976	543	2.40E-07
GO:0022008	Neurogenesis	1096	233	3.56E-07
GO:0007275	multicellular organism development	3266	587	4.31E-07
GO:0048666	neuron development	729	167	5.80E-07
GO:0030154	cell differentiation	2860	521	8.72E-07
GO:0048869	cellular developmental process	2879	523	1.24E-06
GO:0048468	cell development	1355	273	2.92E-06
GO:0010646	regulation of cell communication	2727	493	9.55E-06
GO:0023051	regulation of signaling	2733	492	1.83E-05
GO:0034330	cell junction organization	518	122	2.98E-05
GO:0031175	neuron projection development	656	147	3.28E-05
GO:0009966	regulation of signal transduction	2464	447	4.09E-05
GO:0048513	animal organ development	2176	401	4.17E-05
GO:0051128	regulation of cellular component organization	1929	360	6.48E-05
GO:0000904	cell morphogenesis involved in differentiation	495	116	9.06E-05
GO:0007010	cytoskeleton organization	1309	256	0.000164105
GO:0023052	Signaling	5190	860	0.000200273
GO:0007154	cell communication	5221	863	0.000302145
GO:0007155	cell adhesion	1216	239	0.000305302
GO:0050808	synapse organization	275	72	0.000421197
GO:0009887	animal organ morphogenesis	582	129	0.000483951
GO:0032989	cellular component morphogenesis	549	123	0.000504882
GO:0048667	cell morphogenesis involved in neuron differentiation	381	91	0.001066523
GO:0061564	axon development	323	80	0.001103984
GO:0065007	biological regulation	10721	1650	0.001653583
GO:0048812	neuron projection morphogenesis	441	101	0.002100476
GO:0120036	plasma membrane bounded cell projection organization	1144	222	0.002358738
GO:0048583	regulation of response to stimulus	3327	568	0.002363352
GO:0007409	Axonogenesis	298	74	0.002622931
GO:0003008	system process	1358	257	0.002627434
GO:0030030	cell projection organization	1164	225	0.002676372
GO:0035556	intracellular signal transduction	2168	387	0.002733973
GO:0032879	regulation of localization	1615	298	0.003442374
GO:0051239	regulation of multicellular organismal process	2117	378	0.003636668
GO:0000902	cell morphogenesis	718	149	0.003637049
GO:0007165	signal transduction	4776	786	0.003677398
GO:0098609	cell-cell adhesion	743	153	0.004206573
GO:0120039	plasma membrane bounded cell projection morphogenesis	455	102	0.00520267
GO:0006811	ion transport	1187	227	0.005379723
GO:0051179	Localization	4343	718	0.007233245
GO:0044057	regulation of system process	392	90	0.007238929
GO:0051716	cellular response to stimulus	5982	962	0.007709961
GO:0048858	cell projection morphogenesis	459	102	0.007839606
GO:0065008	regulation of biological quality	2937	503	0.00859953
GO:0097485	neuron projection guidance	169	47	0.009226901
GO:0007411	axon guidance	169	47	0.009226901
GO:0009888	tissue development	1239	234	0.009565902
GO:0050794	regulation of cellular process	9523	1473	0.010601198
GO:0006812	cation transport	886	175	0.011607738
GO:0032990	cell part morphogenesis	469	103	0.012164938
GO:0055085	transmembrane transport	1060	203	0.017138186
GO:0048598	embryonic morphogenesis	308	73	0.018555392
GO:0007417	central nervous system development	584	122	0.022813937
GO:0009790	embryo development	505	108	0.02527435
GO:0031589	cell-substrate adhesion	290	69	0.028200129

GO ID	GO Description	Universe Count	COSMIC Count	Adjusted P-value
GO:0040011	Locomotion	1075	204	0.029753799
GO:0050793	regulation of developmental process	1810	322	0.033576014
GO:0050905	neuromuscular process	73	25	0.036424405
GO:0048523	negative regulation of cellular process	3896	643	0.039903379
GO:0010975	regulation of neuron projection development	288	68	0.042350124
GO:0052697	xenobiotic glucuronidation	11	8	0.042820644
GO:0016477	cell migration	1210	225	0.043542025
GO:0003013	circulatory system process	443	96	0.043990329
GO:0048646	anatomical structure formation involved in morphogenesis	751	149	0.045881349
GO:0003012	muscle system process	305	71	0.046497064
GO:0048519	negative regulation of biological process	4381	715	0.049876587

Appendix Table B 9 Significant GO:BP enrichments for CLINVAR G4 mutations leading to the loss of a G4.

GO ID	GO Description	Universe	CLINVAR	Adjusted P-value
GO:0006936	muscle contraction	260	21	1.07E-06
GO:0003012	muscle system process	305	22	1.86E-06
GO:0060048	cardiac muscle contraction	111	14	5.31E-06
GO:0006941	striated muscle contraction	142	15	9.47E-06
GO:0060047	heart contraction	187	16	3.82E-05
GO:0003015	heart process	193	16	4.82E-05
GO:0086003	cardiac muscle cell contraction	62	10	6.95E-05
GO:0086002	cardiac muscle cell action potential involved in contraction	48	9	9.09E-05
GO:0070252	actin-mediated cell contraction	86	11	1.12E-04
GO:0030048	actin filament-based movement	113	12	1.84E-04
GO:0086001	cardiac muscle cell action potential	73	10	2.19E-04
GO:0008016	regulation of heart contraction	164	13	8.37E-04
GO:0001508	action potential	117	11	1.31E-03
GO:0035637	multicellular organismal signaling	128	11	2.48E-03
GO:0008015	blood circulation	363	18	2.56E-03
GO:0003013	circulatory system process	443	20	2.56E-03
GO:1903522	regulation of blood circulation	188	13	2.64E-03
GO:0060348	bone development	108	10	3.32E-03
GO:0086005	ventricular cardiac muscle cell action potential	29	6	4.45E-03
GO:0061337	cardiac conduction	91	9	5.54E-03
GO:0044057	regulation of system process	392	17	1.65E-02
GO:0002027	regulation of heart rate	84	8	2.02E-02
GO:0001501	skeletal system development	289	14	2.76E-02
GO:0030279	negative regulation of ossification	26	5	2.93E-02
GO:0060537	muscle tissue development	218	12	3.16E-02

Appendix Table B 10 Significant GO:BP enrichments for COSMIC and CLINVAR G4 mutations leading to the gain of a G4.

GO ID	GO Description	Universe	COSMIC And CLINVAR	Adjusted P-value
GO:0048856	anatomical structure development	4152	588	5.78E-18
GO:0032502	developmental process	4584	634	2.01E-17
GO:0048731	system development	2976	448	5.64E-17
GO:0007275	multicellular organism development	3266	480	1.76E-16
GO:0007399	nervous system development	1648	279	1.92E-15
GO:0009653	anatomical structure morphogenesis	1871	303	2.95E-14
GO:0048869	cellular developmental process	2879	419	7.27E-13
GO:0030154	cell differentiation	2860	416	1.07E-12
GO:0007155	cell adhesion	1216	209	2.10E-11
GO:0032501	multicellular organismal process	5319	684	2.28E-11
GO:0030029	actin filament-based process	717	138	2.88E-10
GO:1903508	positive regulation of nucleic acid-templated transcription	1297	215	3.71E-10
GO:0045893	positive regulation of DNA-templated transcription	1297	215	3.71E-10
GO:1902680	positive regulation of RNA biosynthetic process	1303	215	5.95E-10
GO:0045944	positive regulation of transcription by RNA polymerase II	952	167	2.58E-09
GO:0010646	regulation of cell communication	2727	384	3.03E-09
GO:0000902	cell morphogenesis	718	135	3.17E-09
GO:0023051	regulation of signaling	2733	384	4.14E-09
GO:0050793	regulation of developmental process	1810	274	5.89E-09
GO:0032989	cellular component morphogenesis	549	110	7.78E-09
GO:0051254	positive regulation of RNA metabolic process	1433	227	7.96E-09
GO:0030036	actin cytoskeleton organization	637	121	2.50E-08
GO:0048858	cell projection morphogenesis	459	95	3.88E-08
GO:0032990	cell part morphogenesis	469	96	5.73E-08
GO:0045935	positive regulation of nucleobase-containing compound metabolic process	1620	245	1.21E-07
GO:0010557	positive regulation of macromolecule biosynthetic process	1492	229	1.43E-07
GO:0048522	positive regulation of cellular process	4704	594	2.02E-07
GO:0031328	positive regulation of cellular biosynthetic process	1568	237	2.77E-07
GO:0048812	neuron projection morphogenesis	441	90	2.81E-07
GO:0007267	cell-cell signaling	1269	200	2.89E-07
GO:0120039	plasma membrane bounded cell projection morphogenesis	455	92	2.89E-07
GO:0048513	animal organ development	2176	309	3.69E-07
GO:0009891	positive regulation of biosynthetic process	1593	239	4.60E-07
GO:0009887	animal organ morphogenesis	582	109	6.23E-07
GO:0023052	signaling	5190	642	8.45E-07
GO:0048699	generation of neurons	969	160	8.71E-07
GO:0000904	cell morphogenesis involved in differentiation	495	96	1.14E-06
GO:0048468	cell development	1355	208	1.18E-06
GO:0007154	cell communication	5221	644	1.30E-06
GO:0030182	neuron differentiation	923	153	1.64E-06
GO:0022008	neurogenesis	1096	175	1.72E-06
GO:0030030	cell projection organization	1164	183	2.34E-06
GO:0120036	plasma membrane bounded cell projection organization	1144	180	3.05E-06
GO:0007010	cytoskeleton organization	1309	199	6.53E-06
GO:0099536	synaptic signaling	522	96	1.86E-05
GO:0048667	cell morphogenesis involved in neuron differentiation	381	76	2.12E-05
GO:0048518	positive regulation of biological process	5294	642	2.88E-05
GO:0031175	neuron projection development	656	113	3.77E-05
GO:0050804	modulation of chemical synaptic transmission	252	56	4.32E-05
GO:0099177	regulation of trans-synaptic signaling	253	56	4.97E-05
GO:0009966	regulation of signal transduction	2464	330	5.42E-05
GO:0022603	regulation of anatomical structure morphogenesis	700	118	6.09E-05
GO:0051173	positive regulation of nitrogen compound metabolic process	2565	341	6.44E-05
GO:0051239	regulation of multicellular organismal process	2117	289	8.50E-05

GO:0016043	cellular component organization	5370	646	9.41E-05
GO:0048666	neuron development	729	121	9.62E-05
GO:0031325	positive regulation of cellular metabolic process	2522	335	9.64E-05
GO:0099537	trans-synaptic signaling	501	90	1.58E-04
GO:0048589	developmental growth	273	57	2.93E-04
GO:0048523	negative regulation of cellular process	3896	484	2.94E-04
GO:0048870	cell motility	1362	197	3.32E-04
GO:0007268	chemical synaptic transmission	495	88	3.49E-04
GO:0098916	anterograde trans-synaptic signaling	495	88	3.49E-04
GO:0009987	cellular process	14783	1556	6.39E-04
GO:0034330	cell junction organization	518	90	7.04E-04
GO:0009888	tissue development	1239	180	8.89E-04
GO:0035556	intracellular signal transduction	2168	288	1.05E-03
GO:0007417	central nervous system development	584	98	1.09E-03
GO:0010604	positive regulation of macromolecule metabolic process	2889	369	1.09E-03
GO:0045595	regulation of cell differentiation	1129	166	1.22E-03
GO:0051128	regulation of cellular component organization	1929	260	1.31E-03
GO:0099587	inorganic ion import across plasma membrane	112	30	1.40E-03
GO:0098659	inorganic cation import across plasma membrane	112	30	1.40E-03
GO:0060322	head development	402	73	1.69E-03
GO:0071840	cellular component organization or biogenesis	5539	653	1.79E-03
GO:0048519	negative regulation of biological process	4381	529	2.36E-03
GO:0007409	axonogenesis	298	58	2.40E-03
GO:0009893	positive regulation of metabolic process	3177	398	2.59E-03
GO:0016477	cell migration	1210	174	2.61E-03
GO:0051716	cellular response to stimulus	5982	697	3.27E-03
GO:0060560	developmental growth involved in morphogenesis	135	33	3.51E-03
GO:0007420	brain development	384	69	4.70E-03
GO:0048729	tissue morphogenesis	348	64	4.93E-03
GO:0007165	signal transduction	4776	568	5.32E-03
GO:0061564	axon development	323	60	7.13E-03
GO:0014074	response to purine-containing compound	75	22	7.97E-03
GO:0048670	regulation of collateral sprouting	9	7	8.28E-03
GO:0048588	developmental cell growth	129	31	9.81E-03
GO:1990138	neuron projection extension	112	28	1.23E-02
GO:0050794	regulation of cellular process	9523	1052	1.23E-02
GO:0007167	enzyme-linked receptor protein signaling pathway	795	120	1.26E-02
GO:0098657	import into cell	216	44	1.26E-02
GO:0009719	response to endogenous stimulus	1076	154	1.31E-02
GO:0032970	regulation of actin filament-based process	330	60	1.36E-02
GO:0031344	regulation of cell projection organization	467	78	1.63E-02
GO:0071495	cellular response to endogenous stimulus	974	141	1.80E-02
GO:0120035	regulation of plasma membrane bounded cell projection organization	453	76	1.80E-02
GO:0098739	import across plasma membrane	172	37	1.85E-02
GO:0034329	cell junction assembly	341	61	1.88E-02
GO:0051094	positive regulation of developmental process	967	140	1.91E-02
GO:0048583	regulation of response to stimulus	3327	407	1.94E-02
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	63	19	2.15E-02
GO:0007015	actin filament organization	395	68	2.20E-02
GO:0032535	regulation of cellular component size	256	49	2.21E-02
GO:0048638	regulation of developmental growth	167	36	2.28E-02
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	511	83	2.39E-02
GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	195	40	2.62E-02
GO:0051093	negative regulation of developmental process	619	96	3.43E-02
GO:0050789	regulation of biological process	10085	1103	3.55E-02
GO:0010243	response to organonitrogen compound	613	95	3.86E-02
GO:0098609	cell-cell adhesion	743	111	4.07E-02
GO:0048639	positive regulation of developmental growth	83	22	4.10E-02
GO:1905114	cell surface receptor signaling pathway involved in cell-cell signaling	388	66	4.13E-02
GO:0051172	negative regulation of nitrogen compound metabolic process	1927	249	4.44E-02
GO:0072359	circulatory system development	729	109	4.60E-02

Appendix Table B 11 Significant GO:BP enrichments for CLINVAR G4 mutations leading to the gain of a G4.

GO ID	GO Description	Universe	CLINVAR	Adjusted P-value
GO:0001508	action potential	117	6	4.30E-02
GO:0086001	cardiac muscle cell action potential	73	5	4.96E-02

Appendix Table B 12 . Significant GO:CC enrichments for COSMIC and CLINVAR G4 mutations.

GO ID	GO Description	Universe	COSMIC and CLINVAR	Adjusted P-value
GO:0030054	cell junction	1385	893	2.16E-34
GO:0071944	cell periphery	5199	2884	2.79E-34
GO:0042995	cell projection	1567	979	3.89E-30
GO:0005886	plasma membrane	4762	2634	7.87E-29
GO:0120025	plasma membrane bounded cell projection	1540	959	1.11E-28
GO:0098590	plasma membrane region	794	529	4.09E-24
GO:0043005	neuron projection	839	553	1.42E-23
GO:0045202	synapse	652	445	2.85E-23
GO:0016020	membrane	7611	3986	9.36E-21
GO:0098797	plasma membrane protein complex	540	364	2.72E-17
GO:0030424	axon	339	243	4.48E-16
GO:0031226	intrinsic component of plasma membrane	1632	951	2.93E-15
GO:0005887	integral component of plasma membrane	1560	913	2.97E-15
GO:0015629	actin cytoskeleton	441	300	7.75E-15
GO:0036477	somatodendritic compartment	444	300	3.31E-14
GO:0005856	cytoskeleton	1812	1033	5.15E-13
GO:0031252	cell leading edge	289	206	5.23E-13
GO:0097060	synaptic membrane	160	126	9.52E-13
GO:0070161	anchoring junction	707	442	2.07E-12
GO:0043235	receptor complex	395	266	3.53E-12
GO:0098794	postsynapse	293	204	3.32E-11
GO:0005911	cell-cell junction	369	247	8.94E-11
GO:0097447	dendritic tree	341	230	1.79E-10
GO:0030425	dendrite	338	228	2.23E-10
GO:0005737	cytoplasm	10633	5335	1.52E-09
GO:0005912	adherens junction	144	110	1.96E-09
GO:0043228	non-membrane-bounded organelle	4394	2308	4.06E-09
GO:0043232	intracellular non-membrane-bounded organelle	4393	2307	4.57E-09
GO:0031224	intrinsic component of membrane	2600	1406	1.40E-08
GO:0005938	cell cortex	209	147	2.90E-08
GO:0099572	postsynaptic specialization	163	119	4.37E-08
GO:0005829	cytosol	5137	2662	6.95E-08
GO:0099081	supramolecular polymer	659	396	1.20E-07
GO:0098793	presynapse	241	164	1.37E-07
GO:0016021	integral component of membrane	2485	1339	1.60E-07
GO:0014069	postsynaptic density	151	110	2.64E-07
GO:0045211	postsynaptic membrane	106	82	3.26E-07
GO:0098984	neuron to neuron synapse	163	117	3.38E-07
GO:0099512	supramolecular fiber	650	389	3.45E-07
GO:0031982	vesicle	3535	1861	4.51E-07
GO:0042383	sarcolemma	77	63	5.11E-07
GO:0098978	glutamatergic synapse	82	66	7.67E-07
GO:0012505	endomembrane system	3988	2083	8.67E-07
GO:0032279	asymmetric synapse	158	113	9.15E-07
GO:0043025	neuronal cell body	209	143	1.00E-06
GO:0098588	bounding membrane of organelle	1644	905	1.02E-06
GO:0031253	cell projection membrane	215	146	1.50E-06
GO:0097708	intracellular vesicle	1967	1068	1.67E-06
GO:0031410	cytoplasmic vesicle	1966	1067	1.92E-06
GO:0043226	organelle	12740	6294	1.95E-06
GO:0019898	extrinsic component of membrane	282	184	1.95E-06
GO:0034702	ion channel complex	247	164	2.08E-06
GO:0044297	cell body	237	158	2.57E-06
GO:0034703	cation channel complex	218	147	2.71E-06
GO:0150034	distal axon	144	103	4.39E-06
GO:0005884	actin filament	105	79	5.83E-06
GO:0012506	vesicle membrane	881	504	8.55E-06
GO:0042734	presynaptic membrane	53	45	1.25E-05
GO:0030027	lamellipodium	135	96	2.09E-05
GO:0045177	apical part of cell	267	172	2.18E-05
GO:0005622	intracellular anatomical structure	13666	6714	2.42E-05
GO:0030659	cytoplasmic vesicle membrane	868	494	2.79E-05
GO:0019897	extrinsic component of plasma membrane	156	108	3.06E-05

GO:0099080	supramolecular complex	978	549	6.16E-05

GO ID	GO Description	Universe	COSMIC and CLINVAR	Adjusted P-value
GO:0009925	basal plasma membrane	177	119	8.22E-05
GO:0045178	basal part of cell	184	123	8.39E-05
GO:0042641	actomyosin	64	51	8.50E-05
GO:0031256	leading edge membrane	111	80	1.00E-04
GO:0098802	plasma membrane signaling receptor complex	176	118	1.14E-04
GO:0030863	cortical cytoskeleton	82	62	1.36E-04
GO:0044304	main axon	39	34	1.57E-04
GO:0048786	presynaptic active zone	39	34	1.57E-04
GO:0099634	postsynaptic specialization membrane	45	38	1.83E-04
GO:0016323	basolateral plasma membrane	158	107	1.85E-04
GO:0008328	ionotropic glutamate receptor complex	32	29	1.85E-04
GO:0098839	postsynaptic density membrane	35	31	2.38E-04
GO:0030055	cell-substrate junction	395	238	2.60E-04
GO:0030426	growth cone	88	65	2.96E-04
GO:0030427	site of polarized growth	93	68	3.11E-04
GO:0099513	polymeric cytoskeletal fiber	482	284	3.57E-04
GO:0005901	caveola	61	48	3.61E-04
GO:0005925	focal adhesion	387	233	3.62E-04
GO:0098796	membrane protein complex	1115	614	3.62E-04
GO:0016324	apical plasma membrane	234	149	4.01E-04
GO:0097517	contractile actin filament bundle	58	46	4.01E-04
GO:0001725	stress fiber	58	46	4.01E-04
GO:0000785	chromatin	1214	663	5.89E-04
GO:0098858	actin-based cell projection	133	91	6.17E-04
GO:0005604	basement membrane	65	50	7.08E-04
GO:0009898	cytoplasmic side of plasma membrane	160	106	9.53E-04
GO:0032432	actin filament bundle	64	49	1.11E-03
GO:0043292	contractile fiber	173	113	1.26E-03
GO:0005794	Golgi apparatus	1365	737	1.26E-03
GO:0045121	membrane raft	209	133	1.45E-03
GO:0005768	endosome	810	452	1.57E-03
GO:0030016	myofibril	165	108	1.74E-03
GO:0098862	cluster of actin-based cell projections	96	68	1.83E-03
GO:0098857	membrane microdomain	210	133	2.05E-03
GO:0044309	neuron spine	90	64	2.72E-03
GO:0044853	plasma membrane raft	87	62	3.30E-03
GO:0070382	exocytic vesicle	151	99	3.62E-03
GO:0043197	dendritic spine	89	63	3.94E-03
GO:0030017	sarcomere	146	96	4.06E-03
GO:0043229	intracellular organelle	11898	5852	4.14E-03
GO:0034705	potassium channel complex	91	64	4.68E-03
GO:0044291	cell-cell contact zone	50	39	4.80E-03
GO:0048471	perinuclear region of cytoplasm	440	255	5.80E-03
GO:0032589	neuron projection membrane	38	31	7.16E-03
GO:0031674	I band	97	67	7.44E-03
GO:0016342	catenin complex	32	27	7.57E-03
GO:0005769	early endosome	295	177	7.71E-03
GO:0030864	cortical actin cytoskeleton	62	46	8.23E-03
GO:0001726	ruffle	118	79	8.51E-03
GO:1902495	transmembrane transporter complex	331	196	8.99E-03
GO:0098878	neurotransmitter receptor complex	37	30	1.18E-02
GO:0030018	Z disc	86	60	1.19E-02
GO:0032420	stereocilium	31	26	1.29E-02
GO:0008021	synaptic vesicle	128	84	1.42E-02
GO:0031234	extrinsic component of cytoplasmic side of plasma membrane	95	65	1.48E-02
GO:0015630	microtubule cytoskeleton	1159	622	1.58E-02
GO:0030315	T-tubule	36	29	1.92E-02
GO:0016010	dystrophin-associated glycoprotein complex	17	16	2.25E-02
GO:0032281	AMPA glutamate receptor complex	20	18	3.07E-02
GO:0031941	filamentous actin	29	24	3.62E-02
GO:0005815	microtubule organizing center	724	397	3.66E-02
GO:0043194	axon initial segment	12	12	4.29E-02
GO:0008076	voltage-gated potassium channel complex	82	56	4.75E-02
GO:0034399	nuclear periphery	96	64	5.00E-02

GO ID	GO Description	Universe	COSMIC	Adjusted P-value
GO:0071944	cell periphery	5199	2843	5.77E-34
GO:0030054	cell junction	1385	879	3.22E-33
GO:0042995	cell projection	1567	967	4.83E-30
GO:0005886	plasma membrane	4762	2598	7.93E-29
GO:0120025	plasma membrane bounded cell projection	1540	947	1.55E-28
GO:0045202	synapse	652	440	3.61E-23
GO:0098590	plasma membrane region	794	520	5.03E-23
GO:0043005	neuron projection	839	545	5.77E-23
GO:0016020	membrane	7611	3925	2.66E-20
GO:0030424	axon	339	242	1.07E-16
GO:0098797	plasma membrane protein complex	540	358	1.46E-16
GO:0005887	integral component of plasma membrane	1560	900	5.34E-15
GO:0015629	actin cytoskeleton	441	297	6.5E-15
GO:0031226	intrinsic component of plasma membrane	1632	936	9.92E-15
GO:0036477	somatodendritic compartment	444	296	6.2E-14
GO:0097060	synaptic membrane	160	126	2.17E-13
GO:0031252	cell leading edge	289	204	4.97E-13
GO:0005856	cytoskeleton	1812	1018	8.96E-13
GO:0070161	anchoring junction	707	436	3.3E-12
GO:0098794	postsynapse	293	202	3.02E-11
GO:0043235	receptor complex	395	259	8.91E-11
GO:0097447	dendritic tree	341	227	2.74E-10
GO:0030425	dendrite	338	225	3.47E-10
GO:0005911	cell-cell junction	369	242	5.32E-10
GO:0005912	adherens junction	144	110	5.79E-10
GO:0043228	non-membrane-bounded organelle	4394	2279	1.51E-09
GO:0043232	intracellular non-membrane-bounded organelle	4393	2278	1.7E-09
GO:0005737	cytoplasm	10633	5250	4.16E-09
GO:0099572	postsynaptic specialization	163	119	1.28E-08
GO:0005938	cell cortex	209	146	1.83E-08
GO:0031224	intrinsic component of membrane	2600	1385	2.07E-08
GO:0014069	postsynaptic density	151	110	8.66E-08
GO:0016021	integral component of membrane	2485	1322	1.01E-07
GO:0045211	postsynaptic membrane	106	82	1.31E-07
GO:0005829	cytosol	5137	2619	1.67E-07
GO:0098793	presynapse	241	162	1.7E-07
GO:0098984	neuron to neuron synapse	163	116	2.92E-07
GO:0099081	supramolecular polymer	659	389	3.35E-07
GO:0098978	glutamatergic synapse	82	66	3.5E-07
GO:0019898	extrinsic component of membrane	282	184	4.45E-07
GO:0034702	ion channel complex	247	164	5.27E-07
GO:0034703	cation channel complex	218	147	7.57E-07
GO:0032279	asymmetric synapse	158	112	8.12E-07
GO:0099512	supramolecular fiber	650	382	9.7E-07
GO:0042383	sarcolemma	77	62	1.08E-06
GO:0043025	neuronal cell body	209	141	1.5E-06
GO:0043226	organelle	12740	6199	1.51E-06
GO:0012505	endomembrane system	3988	2050	1.58E-06
GO:0031253	cell projection membrane	215	144	2.15E-06
GO:0005884	actin filament	105	79	2.5E-06
GO:0098588	bounding membrane of organelle	1644	889	3E-06
GO:0044297	cell body	237	156	3.16E-06
GO:0031982	vesicle	3535	1825	3.48E-06
GO:0097708	intracellular vesicle	1967	1050	3.97E-06
GO:0150034	distal axon	144	102	4.31E-06
GO:0031410	cytoplasmic vesicle	1966	1049	4.54E-06
GO:0042734	presynaptic membrane	53	45	7.12E-06
GO:0012506	vesicle membrane	881	497	9.91E-06
GO:0019897	extrinsic component of plasma membrane	156	108	1.15E-05
GO:0030027	lamellipodium	135	95	2.2E-05
GO:0005622	intracellular anatomical structure	13666	6611	2.35E-05
GO:0030659	cytoplasmic vesicle membrane	868	487	3.36E-05
GO:0042641	actomyosin	64	51	4.76E-05
GO:0099080	supramolecular complex	978	542	5.51E-05
GO:0030863	cortical cytoskeleton	82	62	7.1E-05

GO ID	GO Description	Universe	COSMIC	Adjusted P-value
GO:0048786	presynaptic active zone	39	34	0.000103
GO:0044304	main axon	39	34	0.000103
GO:0099634	postsynaptic specialization membrane	45	38	0.000116
GO:0031256	leading edge membrane	111	79	0.000126
GO:0008328	ionotropic glutamate receptor complex	32	29	0.000128
GO:0045177	apical part of cell	267	167	0.000152
GO:0030426	growth cone	88	65	0.000155
GO:0030427	site of polarized growth	93	68	0.000159
GO:0098839	postsynaptic density membrane	35	31	0.000161
GO:0001725	stress fiber	58	46	0.000242
GO:0097517	contractile actin filament bundle	58	46	0.000242
GO:0030055	cell-substrate junction	395	235	0.000271
GO:0045178	basal part of cell	184	120	0.000291
GO:0009925	basal plasma membrane	177	116	0.000304
GO:0098796	membrane protein complex	1115	606	0.000328
GO:0005925	focal adhesion	387	230	0.000389
GO:0009898	cytoplasmic side of plasma membrane	160	106	0.000405
GO:0005604	basement membrane	65	50	0.000416
GO:0098802	plasma membrane signaling receptor complex	176	115	0.000416
GO:0098858	actin-based cell projection	133	90	0.000646
GO:0099513	polymeric cytoskeletal fiber	482	279	0.000661
GO:0032432	actin filament bundle	64	49	0.000663
GO:0000785	chromatin	1214	653	0.000752
GO:0016323	basolateral plasma membrane	158	104	0.000776
GO:0005901	caveola	61	47	0.000777
GO:0016324	apical plasma membrane	234	146	0.000888
GO:0043292	contractile fiber	173	112	0.001057
GO:0005768	endosome	810	447	0.001123
GO:0030016	myofibril	165	107	0.001521
GO:0070382	exocytic vesicle	151	99	0.001679
GO:0098862	cluster of actin-based cell projections	96	67	0.002472
GO:0034705	potassium channel complex	91	64	0.002626
GO:0043229	intracellular organelle	11898	5765	0.002825
GO:1902495	transmembrane transporter complex	331	196	0.002936
GO:0045121	membrane raft	209	130	0.003563
GO:0030017	sarcomere	146	95	0.003962
GO:0005769	early endosome	295	176	0.004405
GO:0001726	ruffle	118	79	0.004484
GO:0044853	plasma membrane raft	87	61	0.004808
GO:0098857	membrane microdomain	210	130	0.00492
GO:0032589	neuron projection membrane	38	31	0.005071
GO:0030864	cortical actin cytoskeleton	62	46	0.005237
GO:0016342	catenin complex	32	27	0.005544
GO:0005794	Golgi apparatus	1365	720	0.006418
GO:0008021	synaptic vesicle	128	84	0.007447
GO:0098878	neurotransmitter receptor complex	37	30	0.008507
GO:0031234	extrinsic component of cytoplasmic side of plasma membrane	95	65	0.008596
GO:0044309	neuron spine	90	62	0.009397
GO:0031674	I band	97	66	0.009766
GO:0048471	perinuclear region of cytoplasm	440	250	0.011477
GO:0043197	dendritic spine	89	61	0.013481
GO:0030315	T-tubule	36	29	0.014156
GO:0030018	Z disc	86	59	0.016795
GO:0032281	AMPA glutamate receptor complex	20	18	0.02496
GO:0015630	microtubule cytoskeleton	1159	611	0.027121
GO:1990351	transporter complex	356	204	0.02757
GO:0031941	filamentous actin	29	24	0.027799
GO:0008076	voltage-gated potassium channel complex	82	56	0.029877
GO:0044291	cell-cell contact zone	50	37	0.033809
GO:0043194	axon initial segment	12	12	0.036977
GO:0016363	nuclear matrix	79	54	0.037414
GO:0099240	intrinsic component of synaptic membrane	34	27	0.037971
GO:0032420	stereocilium	31	25	0.042583
GO:0005815	microtubule organizing center	724	391	0.043147

Appendix Table B 13 Significant GO:CC enrichments for CLINVAR G4 mutations

GO ID	GO Description	Universe	CLINVAR	Adjusted P-value
GO:0030017	sarcomere	146	26	9.1E-08
GO:0031674	I band	97	16	5.6E-04
GO:0042383	sarcolemma	77	14	7.7E-04
GO:0036379	myofilament	25	8	1.6E-03
GO:0030315	T-tubule	36	9	3.3E-03
GO:0005865	striated muscle thin filament	21	7	4.4E-03
GO:0014704	intercalated disc	32	8	9.8E-03
GO:0030018	Z disc	86	13	1.1E-02
GO:0043202	lysosomal lumen	87	13	1.2E-02
GO:0033268	node of Ranvier	12	5	2.1E-02
GO:0043194	axon initial segment	12	5	2.1E-02
GO:1990584	cardiac Troponin complex	3	3	2.2E-02
GO:0044291	cell-cell contact zone	50	9	4.0E-02
GO:0005861	troponin complex	8	4	4.8E-02

Appendix Table B 14 Significant KEGG enrichments for COSMIC and CLINVAR G4 mutations

KEGG ID	**KEGG Description**	**Universe**	**COSMIC and CLINVAR**	**Adjusted P-value**
KEGG:04360	Axon guidance	181	137	9.52E-13
KEGG:04921	Oxytocin signaling pathway	154	120	1.19E-12
KEGG:05412	Arrhythmogenic right ventricular cardiomyopathy	77	68	1.05E-11
KEGG:04010	MAPK signaling pathway	294	201	2.43E-11
KEGG:04724	Glutamatergic synapse	114	92	4.31E-11
KEGG:04015	Rap1 signaling pathway	210	151	5.83E-11
KEGG:05200	Pathways in cancer	529	330	1.01E-10
KEGG:04510	Focal adhesion	200	143	5.00E-10
KEGG:04929	GnRH secretion	64	57	5.44E-10
KEGG:04728	Dopaminergic synapse	132	100	4.36E-09
KEGG:04261	Adrenergic signaling in cardiomyocytes	150	111	4.46E-09
KEGG:04072	Phospholipase D signaling pathway	147	109	5.47E-09
KEGG:04014	Ras signaling pathway	234	159	1.91E-08
KEGG:01522	Endocrine resistance	95	75	4.97E-08
KEGG:04725	Cholinergic synapse	113	86	7.34E-08
KEGG:04020	Calcium signaling pathway	239	160	9.32E-08
KEGG:04912	GnRH signaling pathway	93	73	1.37E-07
KEGG:05414	Dilated cardiomyopathy	95	74	1.96E-07
KEGG:04713	Circadian entrainment	97	75	2.75E-07
KEGG:04720	Long-term potentiation	67	55	9.00E-07
KEGG:05215	Prostate cancer	97	74	9.96E-07
KEGG:04934	Cushing syndrome	153	107	1.96E-06
KEGG:04750	Inflammatory mediator regulation of TRP channels	98	74	2.12E-06
KEGG:05410	Hypertrophic cardiomyopathy	90	69	2.26E-06
KEGG:04810	Regulation of actin cytoskeleton	216	143	2.26E-06
KEGG:04012	ErbB signaling pathway	84	65	3.13E-06
KEGG:04925	Aldosterone synthesis and secretion	98	73	6.95E-06
KEGG:04151	PI3K-Akt signaling pathway	353	216	1.11E-05
KEGG:04722	Neurotrophin signaling pathway	119	85	1.35E-05
KEGG:04911	Insulin secretion	86	65	1.49E-05
KEGG:04022	cGMP-PKG signaling pathway	166	112	1.84E-05
KEGG:05213	Endometrial cancer	58	47	2.40E-05
KEGG:04370	VEGF signaling pathway	59	47	6.10E-05
KEGG:04730	Long-term depression	59	47	6.10E-05
KEGG:04660	T cell receptor signaling pathway	103	74	6.17E-05
KEGG:05214	Glioma	75	57	6.46E-05
KEGG:04927	Cortisol synthesis and secretion	64	50	7.72E-05
KEGG:04070	Phosphatidylinositol signaling system	97	70	9.54E-05
KEGG:04919	Thyroid hormone signaling pathway	121	84	1.09E-04
KEGG:04662	B cell receptor signaling pathway	79	59	1.11E-04
KEGG:01521	EGFR tyrosine kinase inhibitor resistance	79	59	1.11E-04
KEGG:05165	Human papillomavirus infection	331	200	1.17E-04
KEGG:04928	Parathyroid hormone synthesis, secretion and action	106	75	1.33E-04
KEGG:05166	Human T-cell leukemia virus 1 infection	219	139	1.37E-04
KEGG:04330	Notch signaling pathway	59	46	2.38E-04
KEGG:04380	Osteoclast differentiation	125	85	3.47E-04
KEGG:05224	Breast cancer	147	98	2.42E-04
KEGG:05031	Amphetamine addiction	69	52	3.10E-04
KEGG:04727	GABAergic synapse	89	64	3.38E-04
KEGG:04380	Osteoclast differentiation	125	85	3.47E-04
KEGG:04390	Hippo signaling pathway	157	103	3.97E-04
KEGG:05225	Hepatocellular carcinoma	166	108	4.11E-04
KEGG:05205	Proteoglycans in cancer	205	129	6.14E-04
KEGG:05222	Small cell lung cancer	92	65	7.28E-04
KEGG:04935	Growth hormone synthesis, secretion and action	120	81	8.95E-04
KEGG:04658	Th1 and Th2 cell differentiation	89	63	9.14E-04
KEGG:04666	Fc gamma R-mediated phagocytosis	96	67	1.01E-03
KEGG:04611	Platelet activation	124	83	1.13E-03
KEGG:04971	Gastric acid secretion	76	55	1.22E-03
KEGG:05220	Chronic myeloid leukemia	76	55	1.22E-03
KEGG:05210	Colorectal cancer	86	59	7.15E-03
KEGG:04540	Gap junction	88	60	8.16E-03

KEGG:05135	Yersinia infection	136	87	8.37E-03
KEGG:04520	Adherens junction	71	50	9.16E-03
KEGG:04310	Wnt signaling pathway	170	108	1.92E-03
KEGG:04926	Relaxin signaling pathway	129	85	2.15E-03
KEGG:04371	Apelin signaling pathway	138	90	2.24E-03
KEGG:04512	ECM-receptor interaction	88	62	1.36E-03
KEGG:05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	89	62	2.34E-03
KEGG:05231	Choline metabolism in cancer	98	67	2.83E-03
KEGG:04910	Insulin signaling pathway	137	89	3.06E-03
KEGG:04922	Glucagon signaling pathway	107	72	3.22E-03
KEGG:04144	Endocytosis	251	151	3.24E-03
KEGG:05226	Gastric cancer	148	95	3.33E-03
KEGG:04659	Th17 cell differentiation	105	70	6.43E-03
KEGG:05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	89	62	2.34E-03
KEGG:04930	Type II diabetes mellitus	46	35	9.21E-03
KEGG:05032	Morphine addiction	90	63	1.60E-03
KEGG:00562	Inositol phosphate metabolism	73	51	1.07E-02
KEGG:04931	Insulin resistance	108	71	1.10E-02
KEGG:04916	Melanogenesis	101	67	1.14E-02
KEGG:04024	cAMP signaling pathway	220	132	1.16E-02
KEGG:04917	Prolactin signaling pathway	70	49	1.35E-02
KEGG:05230	Central carbon metabolism in cancer	70	49	1.35E-02
KEGG:04960	Aldosterone-regulated sodium reabsorption	37	29	1.51E-02
KEGG:05223	Non-small cell lung cancer	72	50	1.56E-02
KEGG:05218	Melanoma	72	50	1.56E-02
KEGG:05163	Human cytomegalovirus infection	223	133	1.57E-02
KEGG:04664	Fc epsilon RI signaling pathway	67	47	1.68E-02
KEGG:04625	C-type lectin receptor signaling pathway	104	68	1.90E-02
KEGG:04721	Synaptic vesicle cycle	78	53	2.32E-02
KEGG:04071	Sphingolipid signaling pathway	119	76	2.41E-02
KEGG:04152	AMPK signaling pathway	121	77	2.59E-02
KEGG:04726	Serotonergic synapse	112	72	2.62E-02
KEGG:05219	Bladder cancer	41	31	2.74E-02
KEGG:04270	Vascular smooth muscle contraction	134	84	2.83E-02
KEGG:04933	AGE-RAGE signaling pathway in diabetic complications	100	65	3.26E-02

Appendix Table B 15 Significant KEGG enrichments for COSMIC G4 mutations.

GO ID	GO Description	Universe	COSMIC	Adjusted P-value
KEGG:04611	Platelet activation	124	81	0.002339
KEGG:05226	Gastric cancer	148	94	0.002757
KEGG:04144	Endocytosis	251	149	0.002928
KEGG:05166	Human T-cell leukemia virus 1 infection	219	132	0.003096
KEGG:04926	Relaxin signaling pathway	129	83	0.00414
KEGG:04024	cAMP signaling pathway	220	132	0.004174
KEGG:04935	Growth hormone synthesis, secretion and action	120	78	0.004212
KEGG:04520	Adherens junction	71	50	0.005182
KEGG:04930	Type II diabetes mellitus	46	35	0.005861
KEGG:00562	Inositol phosphate metabolism	73	51	0.006052
KEGG:04666	Fc gamma R-mediated phagocytosis	96	64	0.006737
KEGG:04658	Th1 and Th2 cell differentiation	89	60	0.007009
KEGG:05231	Choline metabolism in cancer	98	65	0.007447
KEGG:04960	Aldosterone-regulated sodium reabsorption	37	29	0.010324
KEGG:04910	Insulin signaling pathway	137	86	0.010713
KEGG:04916	Melanogenesis	101	66	0.012641
KEGG:04721	Synaptic vesicle cycle	78	53	0.013364
KEGG:05135	Yersinia infection	136	85	0.014406
KEGG:05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	89	59	0.015883
KEGG:04922	Glucagon signaling pathway	107	69	0.015977
KEGG:05230	Central carbon metabolism in cancer	70	48	0.019532
KEGG:05210	Colorectal cancer	86	57	0.020522
KEGG:05218	Melanoma	72	49	0.022161
KEGG:04931	Insulin resistance	108	69	0.023655
KEGG:04152	AMPK signaling pathway	121	76	0.025197
KEGG:04270	Vascular smooth muscle contraction	134	83	0.025742
KEGG:05163	Human cytomegalovirus infection	223	130	0.026284
KEGG:05220	Chronic myeloid leukemia	76	51	0.027645
KEGG:04659	Th17 cell differentiation	105	67	0.030128
KEGG:04625	C-type lectin receptor signaling pathway	104	66	0.041098
KEGG:05030	Cocaine addiction	49	35	0.043077
KEGG:04071	Sphingolipid signaling pathway	119	74	0.045066
KEGG:04917	Prolactin signaling pathway	70	47	0.04558
KEGG:04540	Gap junction	88	57	0.047383

Appendix Table B 16 Significant KEGG enrichments for CLINVAR G4 mutations

KEGG ID	KEGG Description	Universe	CLINVAR	Adjusted P-value
KEGG:05414	Dilated cardiomyopathy	95	21	6.7E-07
KEGG:05410	Hypertrophic cardiomyopathy	90	20	1.4E-06
KEGG:05412	Arrhythmogenic right ventricular cardiomyopathy	77	18	3.4E-06
KEGG:04261	Adrenergic signaling in cardiomyocytes	150	23	1.0E-04
KEGG:04512	ECM-receptor interaction	88	16	5.1E-04
KEGG:04260	Cardiac muscle contraction	87	15	1.8E-03
KEGG:05221	Acute myeloid leukemia	67	12	7.9E-03
KEGG:05230	Central carbon metabolism in cancer	70	12	1.2E-02
KEGG:04919	Thyroid hormone signaling pathway	121	16	2.0E-02
KEGG:05220	Chronic myeloid leukemia	76	12	2.4E-02
KEGG:04912	GnRH signaling pathway	93	13	4.1E-02

Appendix Table B 17 Significant KEGG enrichments for COSMIC and CLINVAR G4 mutations leading to a G4 loss.

KEGG ID	KEGG Description	Universe	COSMIC and CLINVAR	Adjusted P-value
KEGG:05412	Arrhythmogenic right ventricular cardiomyopathy	77	44	4.37E-11
KEGG:05414	Dilated cardiomyopathy	95	50	7.37E-11
KEGG:05410	Hypertrophic cardiomyopathy	90	48	1.15E-10
KEGG:04261	Adrenergic signaling in cardiomyocytes	150	61	1.83E-07
KEGG:04921	Oxytocin signaling pathway	154	60	1.62E-06
KEGG:04010	MAPK signaling pathway	294	95	7.23E-06
KEGG:04015	Rap1 signaling pathway	210	73	1.11E-05
KEGG:04510	Focal adhesion	200	70	1.52E-05
KEGG:04020	Calcium signaling pathway	239	80	1.63E-05
KEGG:04725	Cholinergic synapse	113	44	1.41E-04
KEGG:04151	PI3K-Akt signaling pathway	353	104	2.35E-04
KEGG:04514	Cell adhesion molecules	153	53	6.96E-04
KEGG:05200	Pathways in cancer	529	142	9.19E-04
KEGG:04022	cGMP-PKG signaling pathway	166	56	9.35E-04
KEGG:04024	cAMP signaling pathway	220	69	1.41E-03
KEGG:04512	ECM-receptor interaction	88	34	2.64E-03
KEGG:04810	Regulation of actin cytoskeleton	216	67	2.74E-03
KEGG:04925	Aldosterone synthesis and secretion	98	36	5.22E-03
KEGG:04360	Axon guidance	181	57	6.73E-03
KEGG:04929	GnRH secretion	64	26	8.78E-03
KEGG:04713	Circadian entrainment	97	35	9.76E-03
KEGG:04911	Insulin secretion	86	32	1.00E-02
KEGG:04662	B cell receptor signaling pathway	79	30	1.07E-02
KEGG:04934	Cushing syndrome	153	49	1.34E-02
KEGG:04724	Glutamatergic synapse	114	39	1.42E-02
KEGG:05165	Human papillomavirus infection	331	91	1.55E-02
KEGG:04014	Ras signaling pathway	234	68	1.95E-02
KEGG:04728	Dopaminergic synapse	132	43	2.21E-02
KEGG:05032	Morphine addiction	90	32	2.49E-02
KEGG:04390	Hippo signaling pathway	157	49	2.56E-02
KEGG:05224	Breast cancer	147	46	3.61E-02
KEGG:04730	Long-term depression	59	23	4.14E-02
KEGG:04727	GABAergic synapse	89	31	4.49E-02

Appendix Table B 18 Significant KEGG enrichments for COSMIC G4 mutations leading to a G4 loss.

KEGG ID	KEGG Description	Universe	COSMIC	Adjusted P-value
KEGG:05410	Hypertrophic cardiomyopathy	90	37	9.29E-08
KEGG:05414	Dilated cardiomyopathy	95	38	1.37E-07
KEGG:05412	Arrhythmogenic right ventricular cardiomyopathy	77	32	9.49E-07
KEGG:04261	Adrenergic signaling in cardiomyocytes	150	43	5.39E-04
KEGG:04020	Calcium signaling pathway	239	59	1.65E-03
KEGG:04921	Oxytocin signaling pathway	154	42	2.51E-03
KEGG:04360	Axon guidance	181	45	1.33E-02
KEGG:04022	cGMP-PKG signaling pathway	166	42	1.47E-02
KEGG:04015	Rap1 signaling pathway	210	50	1.76E-02
KEGG:04510	Focal adhesion	200	48	1.92E-02
KEGG:04024	cAMP signaling pathway	220	51	2.94E-02
KEGG:04151	PI3K-Akt signaling pathway	353	74	3.78E-02

.

Appendix Table B 19 Significant KEGG enrichments for CLINVAR G4 mutations leading to a G4 loss.

KEGG ID	KEGG Description	Universe	CLINVAR	Adjusted P-value
KEGG:05410	Hypertrophic cardiomyopathy	90	10	5.17E-04
KEGG:05414	Dilated cardiomyopathy	95	10	7.44E-04
KEGG:05412	Arrhythmogenic right ventricular cardiomyopathy	77	9	9.61E-04
KEGG:04261	Adrenergic signaling in cardiomyocytes	150	9	4.67E-02
KEGG:05221	Acute myeloid leukemia	67	6	5.00E-02

Appendix Table B 20 Significant GO:CC enrichments for COSMIC and CLINVAR G4 mutations leading to a G4 gain.

KEGG ID	KEGG Description	Universe	COSMIC and CLINVAR	Adjusted P-value
KEGG:04929	GnRH secretion	64	22	1.99E-05
KEGG:05200	Pathways in cancer	529	89	5.62E-05
KEGG:04724	Glutamatergic synapse	114	30	8.27E-05
KEGG:04725	Cholinergic synapse	113	28	6.58E-04
KEGG:04919	Thyroid hormone signaling pathway	121	29	8.96E-04
KEGG:04015	Rap1 signaling pathway	210	42	1.20E-03
KEGG:04261	Adrenergic signaling in cardiomyocytes	150	33	1.45E-03
KEGG:05202	Transcriptional misregulation in cancer	192	39	1.70E-03
KEGG:04713	Circadian entrainment	97	24	3.03E-03
KEGG:04930	Type II diabetes mellitus	46	15	3.49E-03
KEGG:05210	Colorectal cancer	86	22	3.76E-03
KEGG:04010	MAPK signaling pathway	294	52	3.76E-03
KEGG:04928	Parathyroid hormone synthesis, secretion and action	106	25	4.76E-03
KEGG:04072	Phospholipase D signaling pathway	147	31	5.89E-03
KEGG:04730	Long-term depression	59	17	5.93E-03
KEGG:04921	Oxytocin signaling pathway	154	32	6.02E-03
KEGG:05218	Melanoma	72	19	7.97E-03
KEGG:04728	Dopaminergic synapse	132	28	1.16E-02
KEGG:04934	Cushing syndrome	153	31	1.22E-02
KEGG:04512	ECM-receptor interaction	88	21	1.52E-02
KEGG:05213	Endometrial cancer	58	16	1.62E-02
KEGG:04961	Endocrine and other factor-regulated calcium reabsorption	53	15	1.89E-02
KEGG:04510	Focal adhesion	200	37	1.91E-02
KEGG:04974	Protein digestion and absorption	103	23	2.09E-02
KEGG:04810	Regulation of actin cytoskeleton	216	39	2.15E-02
KEGG:05030	Cocaine addiction	49	14	2.72E-02
KEGG:04720	Long-term potentiation	67	17	2.91E-02
KEGG:04911	Insulin secretion	86	20	3.02E-02
KEGG:04151	PI3K-Akt signaling pathway	353	56	3.34E-02
KEGG:05165	Human papillomavirus infection	331	53	3.80E-02
KEGG:04540	Gap junction	88	20	4.06E-02

Appendix Table B 21 Significant KEGG enrichments for COSMIC G4 mutations leading to a G4 gain.

KEGG ID	KEGG Description	Universe	COSMIC	Adjusted P-value
KEGG:05218	Melanoma	72	13	1.45E-02
KEGG:04072	Phospholipase D signaling pathway	147	20	1.59E-02
KEGG:05030	Cocaine addiction	49	10	2.90E-02

Appendix Table B 22 Significant INTERPRO enrichments for COSMIC and CLINVAR G4 mutations

INTERPRO ID	COSMIC and CLINVAR	UNIVERSE	FDR
IPR011993:Pleckstrin homology-like domain	150	446	1.41E-10
IPR001849:Pleckstrin homology domain	95	277	1.4E-06
IPR011009:Protein kinase-like domain	158	547	4.88E-06
IPR000719:Protein kinase, catalytic domain	146	502	9.19E-06
IPR013098:Immunoglobulin I-set	52	140	0.000295
IPR008271:Serine/threonine-protein kinase, active site	96	316	0.000295
IPR017441:Protein kinase, ATP binding site	113	390	0.000331
IPR017970:Homeobox, conserved site	65	193	0.000418
IPR001781:Zinc finger, LIM-type	32	75	0.001657
IPR001452:Src homology-3 domain	72	230	0.001677
IPR002219:Protein kinase C-like, phorbol ester/diacylglycerol binding	29	67	0.003026
IPR003598:Immunoglobulin subtype 2	76	254	0.004198
IPR013164:Cadherin, N-terminal	28	65	0.004198
IPR008936:Rho GTPase activation protein	35	95	0.013732
IPR015425:Actin-binding FH2	11	15	0.013732
IPR000008:C2 calcium-dependent membrane targeting	48	148	0.018196
IPR020479:Homeodomain, metazoa	34	93	0.018196
IPR013088:Zinc finger, NHR/GATA-type	24	57	0.021614
IPR001025:Bromo adjacent homology (BAH) domain	9	11	0.025203
IPR000536:Nuclear hormone receptor, ligand-binding, core	21	48	0.031914
IPR009057:Homeodomain-like	95	360	0.04254
IPR001478:PDZ domain	50	163	0.04254
IPR001628:Zinc finger, nuclear hormone receptor-type	20	46	0.045727

Appendix Table B 23 Significant INTERPRO enrichments for COSMIC G4 mutations.

INTERPRO ID	COSMIC	Universe	FDR
IPR011993:Pleckstrin homology-like domain	147	446	1.30E-10
IPR001849:Pleckstrin homology domain	93	277	1.46E-06
IPR011009:Protein kinase-like domain	153	547	1.13E-05
IPR000719:Protein kinase, catalytic domain	141	502	2.55E-05
IPR013098:Immunoglobulin I-set	50	140	7.98E-04
IPR017441:Protein kinase, ATP binding site	109	390	7.98E-04
IPR017970:Homeobox, conserved site	63	193	7.98E-04
IPR008271:Serine/threonine-protein kinase, active site	91	316	1.39E-03
IPR002219:Protein kinase C-like, phorbol ester/diacylglycerol binding	29	67	1.98E-03
IPR001781:Zinc finger, LIM-type	31	75	2.42E-03
IPR013164:Cadherin, N-terminal	28	65	2.72E-03
IPR001452:Src homology-3 domain	69	230	3.73E-03
IPR008936:Rho GTPase activation protein	35	95	7.92E-03
IPR015425:Actin-binding FH2	11	15	1.09E-02
IPR020479:Homeodomain, metazoa	34	93	1.09E-02
IPR003598:Immunoglobulin subtype 2	72	254	1.39E-02
IPR000008:C2 calcium-dependent membrane targeting	47	148	1.68E-02
IPR001025:Bromo adjacent homology (BAH) domain	9	11	2.07E-02
IPR001478:PDZ domain	50	163	2.28E-02
IPR013088:Zinc finger, NHR/GATA-type	23	57	3.61E-02
IPR009057:Homeodomain-like	93	360	3.71E-02
IPR015919:Cadherin-like	39	121	4.08E-02
IPR002126:Cadherin	38	118	4.83E-02

Appendix Table B 24 Significant INTERPRO enrichments for CLINVAR G4 mutations.

INTERPRO ID	CLINVAR	UNIVERSE	FDR
IPR000595:Cyclic nucleotide-binding domain	7	36	0.009358
IPR018490:Cyclic nucleotide-binding-like	7	39	0.009358

Appendix Table B 25 :Top 50 significant transcription factor enrichments for COSMIC and CLINVAR G4.

Transcription Factor ID	Transcription Factor Description	Universe	COSMIC and CLINVAR	FDR
TF:M09636_1	Factor: MAZ; motif: GGGMGGGGSSGGGGGGGGGGG; match class: 1	14379	7641	6.97E-262
TF:M09973_1	Factor: CPBP; motif: GNNRGGGHGGGGNNGGGRN; match class: 1	6788	4243	1.01E-254
TF:M09826_1	Factor: BTEB3; motif: CCNNSCCNSCCCCKCCCCC; match class: 1	7694	4675	1.41E-249
TF:M07289_1	Factor: GKLF; motif: NNNRGGNGNGGSN; match class: 1	10800	6111	2.57E-248
TF:M07039_1	Factor: ETF; motif: CCCCGCCCCYN; match class: 1	13890	7403	5.24E-241
TF:M09973	Factor: CPBP; motif: GNNRGGGHGGGGNNGGGRN	11087	6214	2.70E-238
TF:M09984	Factor: MAZ; motif: GGGGGAGGGGGNGRGRRRGNRG	9762	5614	4.46E-236
TF:M12351_1	Factor: TIEG1; motif: NCCCNSNCCCCGCCCCC; match class: 1	8412	4966	9.71E-228
TF:M09723	Factor: BTEB1; motif: GGGGGCGGGGCNGSGGGNGS	10228	5801	4.64E-226
TF:M09826	Factor: BTEB3; motif: CCNNSCCNSCCCCKCCCCC	11731	6462	1.50E-224
TF:M09984_1	Factor: MAZ; motif: GGGGGAGGGGGNGRGRRRGNRG; match class: 1	5696	3615	3.02E-220
TF:M10026	Factor: PATZ; motif: GGGGNGGGGGMKGGRRNGGNRN	8607	5037	2.42E-219
TF:M07040_1	Factor: GKLF; motif: NNRRGRRNGNSNNN; match class: 1	8337	4909	3.67E-219
TF:M00986_1	Factor: Churchill; motif: CGGGNN; match class: 1	10609	5947	4.14E-216
TF:M09723_1	Factor: BTEB1; motif: GGGGGCGGGGCNGSGGGNGS; match class: 1	6131	3819	2.47E-213
TF:M12160_1	Factor: KLF15; motif: RCCMCRCCCMCN; match class: 1	8212	4823	1.49E-208
TF:M10432_1	Factor: MAZ; motif: GGGMGGGGS; match class: 1	4484	2948	2.74E-203
TF:M12351	Factor: TIEG1; motif: NCCCNSNCCCCGCCCCC	12580	6766	2.59E-200
TF:M10432	Factor: MAZ; motif: GGGMGGGGS	9496	5399	2.34E-199
TF:M00933	Factor: Sp1; motif: CCCCGCCCCN	9913	5589	4.40E-199
TF:M10529	Factor: Sp1; motif: RGGGMGGRGSNGGGG	7039	4230	1.45E-197
TF:M04953	Factor: Sp1; motif: GGNDGGRGGCGGGG	8852	5093	1.56E-196
TF:M02089_1	Factor: E2F-3; motif: GGCGGGN; match class: 1	9606	5440	8.00E-196
TF:M10112	Factor: Miz-1; motif: NNRGGWGGGGGAGGGGMRR	8878	5103	9.30E-196
TF:M12160	Factor: KLF15; motif: RCCMCRCCCMCN	12959	6914	1.26E-195
TF:M09636	Factor: MAZ; motif: GGGMGGGGSSGGGGGGGGGGG	16533	8315	1.88E-194
TF:M10026_1	Factor: PATZ; motif: GGGGNGGGGGMKGGRRNGGNRN; match class: 1	5053	3219	9.11E-192
TF:M01104_1	Factor: MOVO-B; motif: GNGGGGG; match class: 1	5798	3599	1.13E-191
TF:M00932_1	Factor: Sp1; motif: NNGGGGCGGGGNN; match class: 1	6212	3805	5.40E-191
TF:M07395_1	Factor: Sp1; motif: NGGGGCGGGGN; match class: 1	6529	3956	1.81E-188
TF:M00931	Factor: Sp1; motif: GGGGCGGGGC	10524	5832	8.08E-187
TF:M00933_1	Factor: Sp1; motif: CCCCGCCCCN; match class: 1	5316	3340	7.30E-186
TF:M09834	Factor: ZNF148; motif: NNNNNNCCNNCCCCTCCCCCACCCN	7099	4227	3.40E-185
TF:M00932	Factor: Sp1; motif: NNGGGGCGGGGNN	10669	5892	5.01E-185
TF:M01303	Factor: SP1; motif: GGGGYGGGGNS	8089	4697	5.64E-183
TF:M03876_1	Factor: Kaiso; motif: GCMGGGRGCRGS; match class: 1	9311	5267	3.63E-182
TF:M07436	Factor: WT1; motif: NNGGGNGGGSGN	6637	3990	4.64E-181
TF:M07226	Factor: SP1; motif: NCCCCKCCCCC	8460	4865	2.87E-180

TF:M07397	Factor: ZBP89; motif: CCCCKCCCCCNN	7289	4306	2.95E-180
TF:M07289	Factor: GKLF; motif: NNNRGGNGNGGSN	14918	7675	3.00E-180
TF:M00196	Factor: Sp1; motif: NGGGGGCGGGGYN	10479	5792	1.13E-179
TF:M10071	Factor: Sp1; motif: NGGGGGCGGGGCCNGGGGGGGG	8705	4978	1.41E-179
TF:M00931_1	Factor: Sp1; motif: GGGGCGGGGC; match class: 1	6075	3707	2.15E-179
TF:M11529_1	Factor: E2F-2; motif: GCGCGCGCNCS; match class: 1	14789	7621	5.14E-179
TF:M07395	Factor: Sp1; motif: NGGGGCGGGGN	10901	5975	1.19E-177
TF:M00196_1	Factor: Sp1; motif: NGGGGGCGGGGYN; match class: 1	6084	3703	4.14E-176
TF:M09970	Factor: KLF3; motif: NNNNNNGGGCGGGGCNNGN	7907	4589	3.12E-175
TF:M07039	Factor: ETF; motif: CCCCGCCCCYN	16656	8320	3.08E-174
TF:M01104	Factor: MOVO-B; motif: GNGGGGG	10486	5777	3.21E-173
TF:M12703_1	Factor: ZNF383; motif: SSNGGGMGGNGSNGGS; match class: 1	4453	2863	3.94E-173

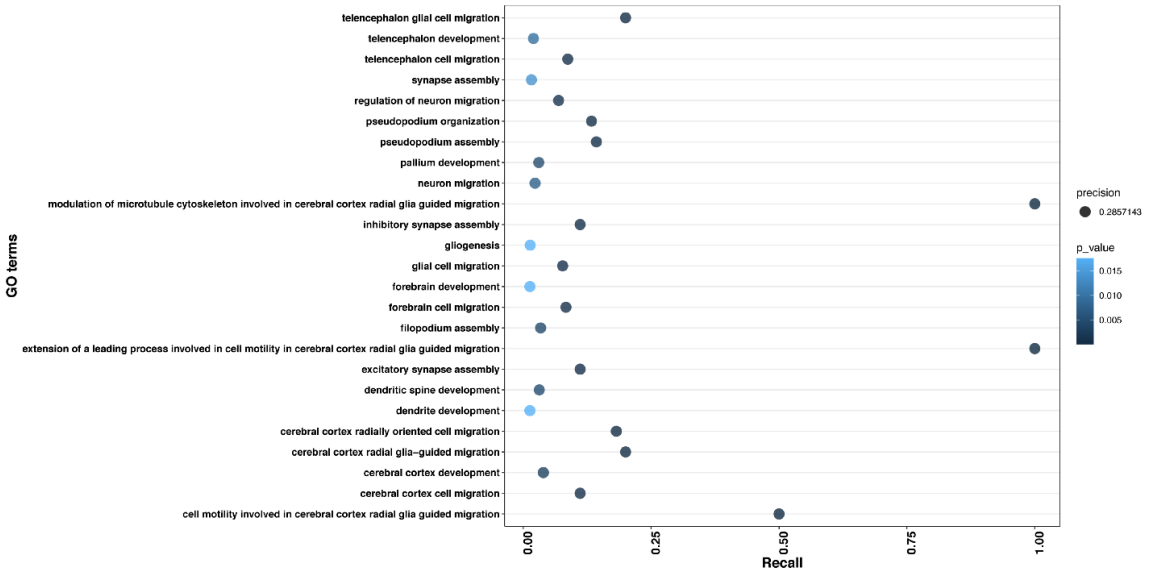
Appendix Table B 26 . Count and percentage of effect of SNV calculated by thermodynamic MFE and ED changes in the G quadruplex sequence

Change in Stability by SNV	Change in multiconfirm by SNV	Frequency	Percentage
Further stabilized	less diversity	6,417	17.105
no change	less diversity	3,835	10.222
Destabilized	less diversity	5,383	14.349
Further stabilized	no change	34	0.091
no change	no change	5,378	14.335
Destabilized	no change	39	0.104
Further stabilized	more diversity	3,984	10.619
no change	more diversity	2,849	7.594
Destabilized	more diversity	9,597	25.581

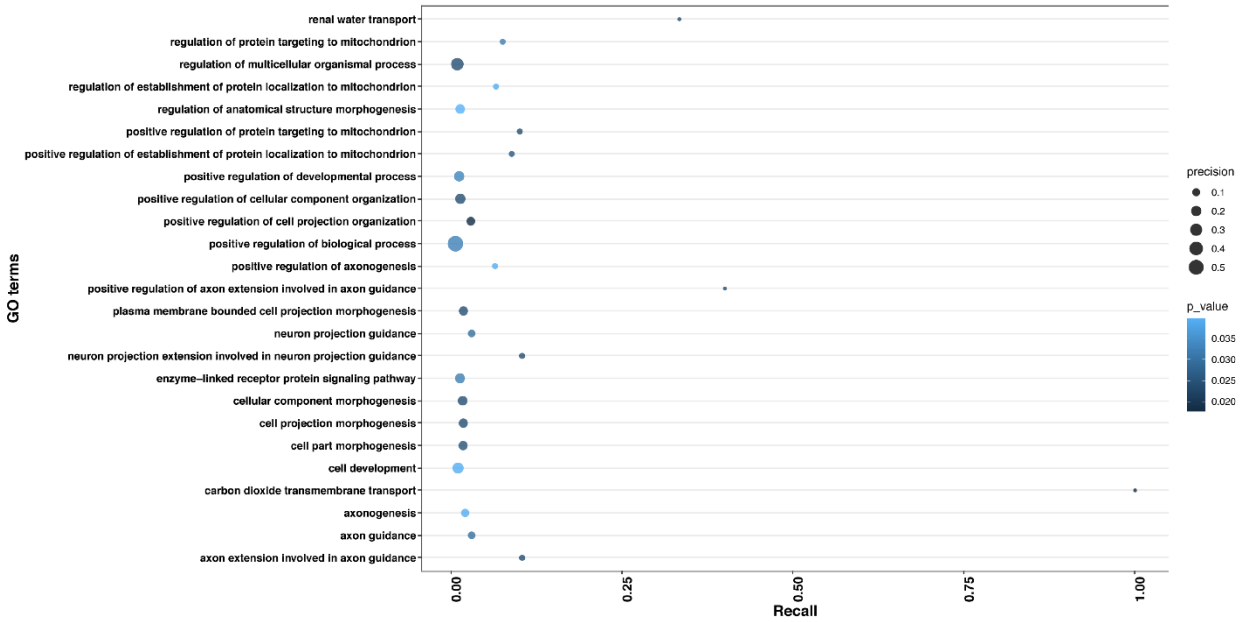
.

Appendix Table B 27 Effect of transition mutation G→A in chr10:122,143,482 on potential binding for multiple transcription factors. All effects are strong

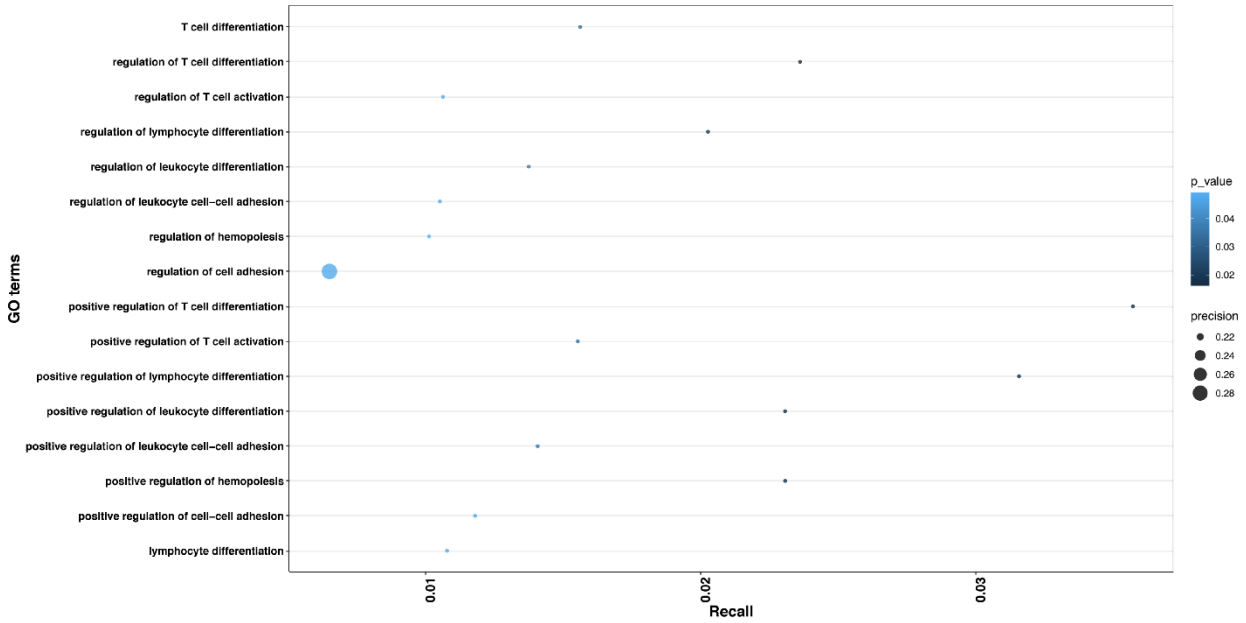
Motif Position	Gene Symbol	Transcription factor binding match	Reference P-value	Alternate P-value	Allele Difference	Allele Effect Size
-3|6	NHLH1	tgtgtgggcAggtgggttg	0.0018	2.86E-05	2.1672	0.1850
-8|2	FOXO3	atgtgtgggcAggtgggttgg	0.0033	0.0001	2.3166	0.1431
-3|6	TAL1	tgtgtgggcAggtgggttg	0.0045	0.0001	2.3166	0.1814
-12|7	TP53	gtccattccatgtgtgggcAggtgggttgggtggggtga	0.0031	0.0001	2.3166	0.1149
-4|7	HES5	catgtgtgggcAggtgggttggg	0.0040	0.0001	2.2697	0.1547
-4|7	HES7	catgtgtgggcAggtgggttggg	0.0041	0.0002	2.2817	0.14880
1|8	USF2	gtgtgggcAggtgggtt	0.0026	0.0002	1.8992	0.1282
-11|3	EGR3	ttccatgtgtgggcGggtgggttgggtgg	3.47E-06	0.0002	-1.8709	-0.1170
-12|2	EGR3	ttccatgtgtgggcGggtgggttgggtgg	5.21E-06	0.0002	-1.8638	-0.1082
-11|2	EGR1	tccatgtgtgggcGggtgggttgggtg	6.63E-06	0.0002	-1.8447	-0.1058
-9|1	EGR2	atgtgtgggcGggtgggttgg	5.96E-06	0.0002	-1.4785	-0.1112
-11|2	EGR1	tccatgtgtgggcGggtgggttgggtg	4.59E-06	0.0003	-1.6940	-0.1215
-11|3	EGR2	ttccatgtgtgggcGggtgggttgggtgg	5.46E-06	0.0003	-1.7277	-0.1245
-12|3	EGR1	attccatgtgtgggcGggtgggttgggtggg	1.35E-05	0.0004	-1.9116	-0.1130
-8|1	EGR1	tgtgtgggcGggtgggttg	1.62E-05	0.0006	-1.9401	-0.1347
-6|3	ZNF740	tgtgtgggcGggtgggttg	4.86E-05	0.0011	-1.6642	-0.1248
-6|3	SP1	tgtgtgggcGggtgggttg	0.0001	0.0017	-1.0496	-0.1072
-6|4	KLF16	atgtgtgggcGggtgggttgg	0.0001	0.0021894	-1.5426264	-0.12738754
-7|9	SP4	cattccatgtgtgggcGggtgggttgggtgggg	0.0002	0.0023592	-1.76102043	-0.10166014
-6|4	SP1	atgtgtgggcGggtgggttgg	0.0002	0.00291348	-1.15354231	-0.11339262
-3|6	SP1	tgtgtgggcGggtgggttg	0.0002	0.00323868	-1.21127823	-0.11785306
-6|3	ZNF740	tgtgtgggcGggtgggttg	0.0002	0.00376701	-1.71699463	-0.13807642
-9|2	ZBTB7A	catgtgtgggcGggtgggttggg	0.0004	0.0038684	-1.3008656	-0.11661897
-11|5	SP4	cattccatgtgtgggcGggtgggttgggtgggg	0.0002	0.0047031	-1.86702905	-0.14590419
-6|4	SP3	atgtgtgggcGggtgggttgg	0.0003	0.00499582	-1.29081923	-0.13319603



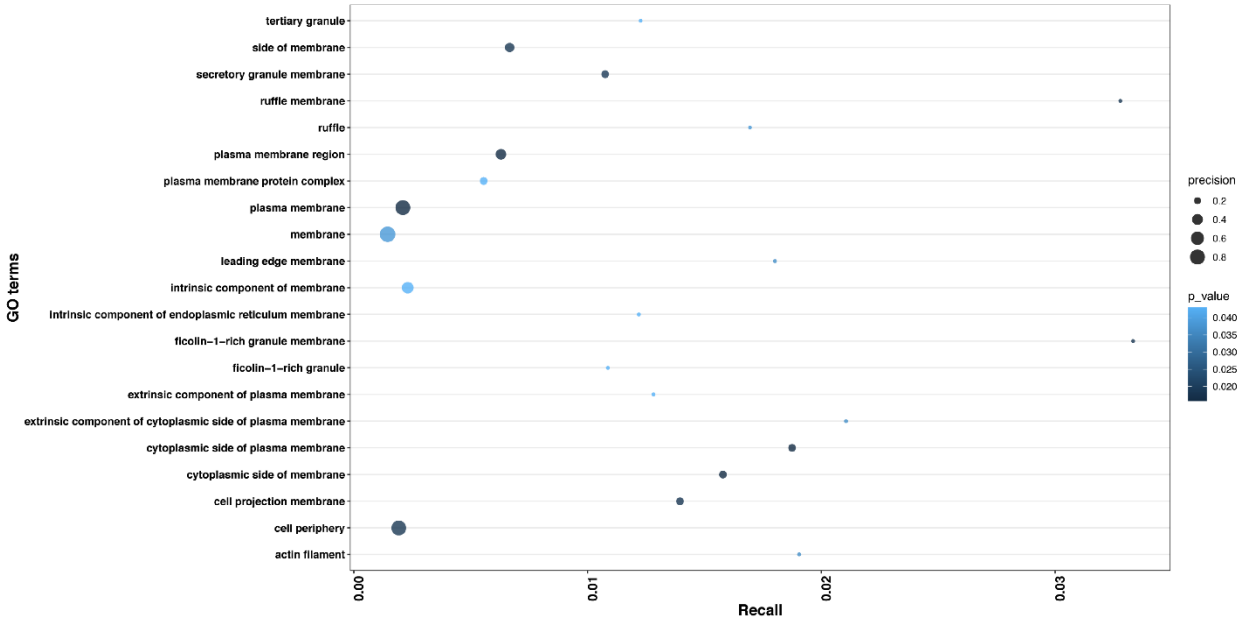
Appendix Figure A 6 Top 25 GO: BP enrichments for Family 4



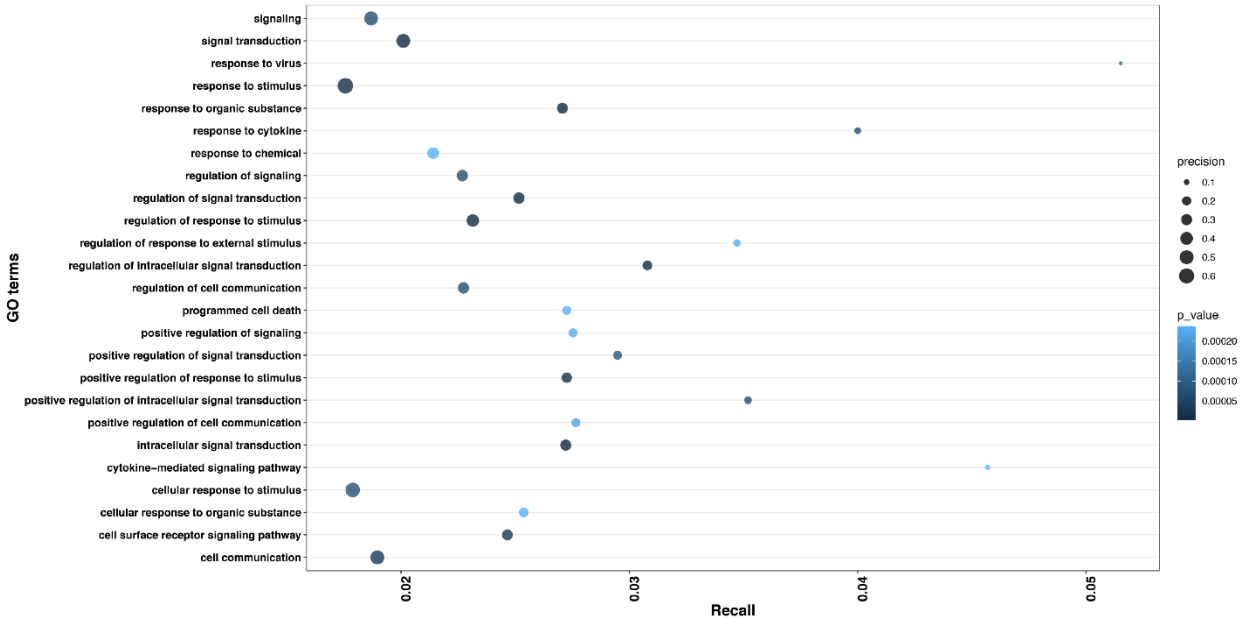
Appendix Figure A 7 Top 25 GO: BP enrichments for Family 32



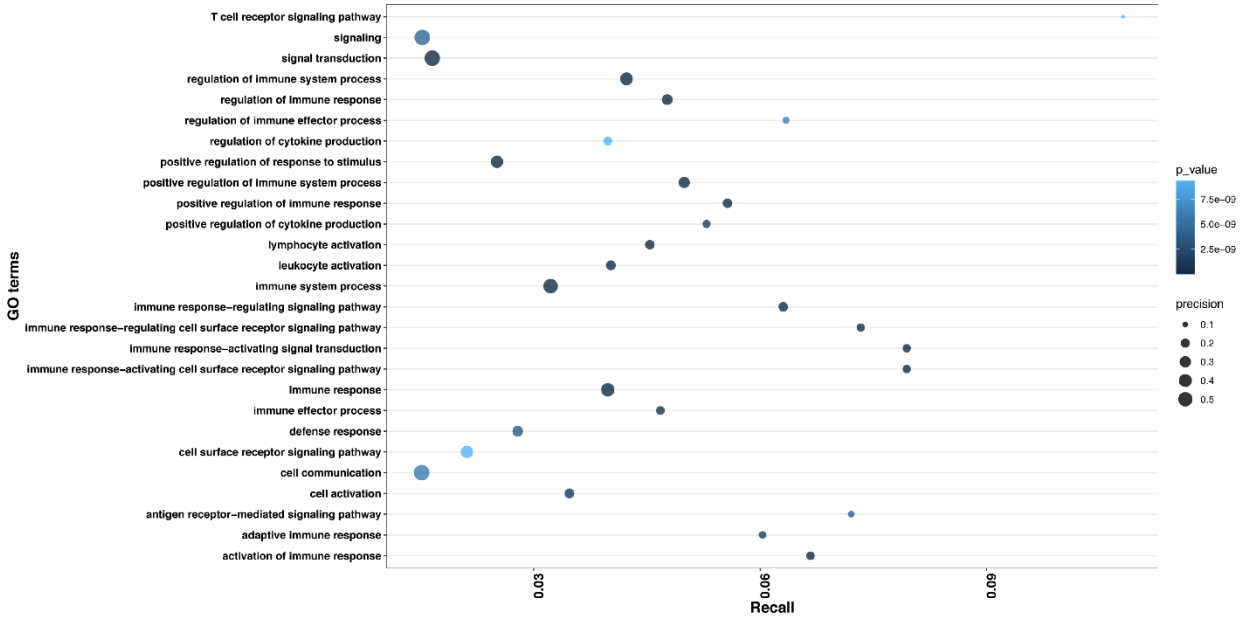
Appendix Figure A 8 Top 25 GO: BP enrichments for Family 75



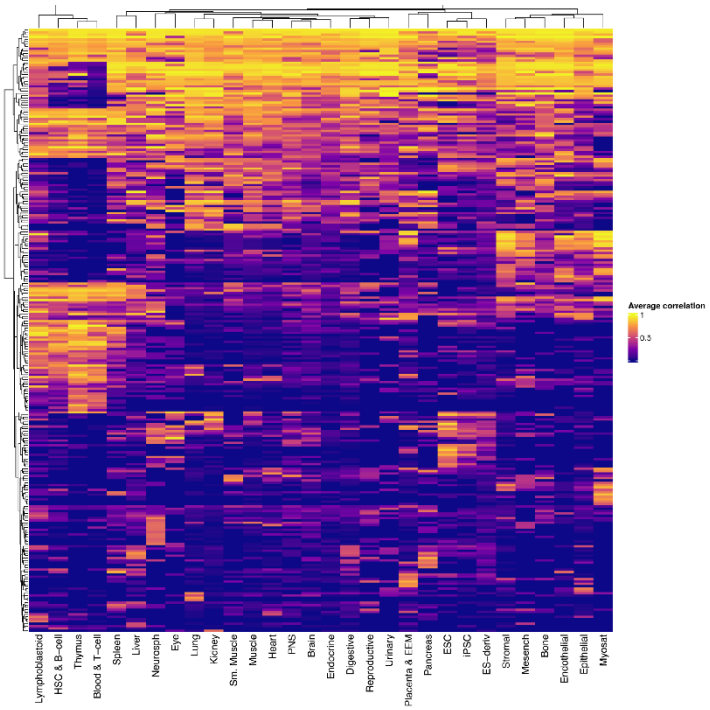
Appendix Figure A 9 Top 25 GO: BP enrichments for Family 80



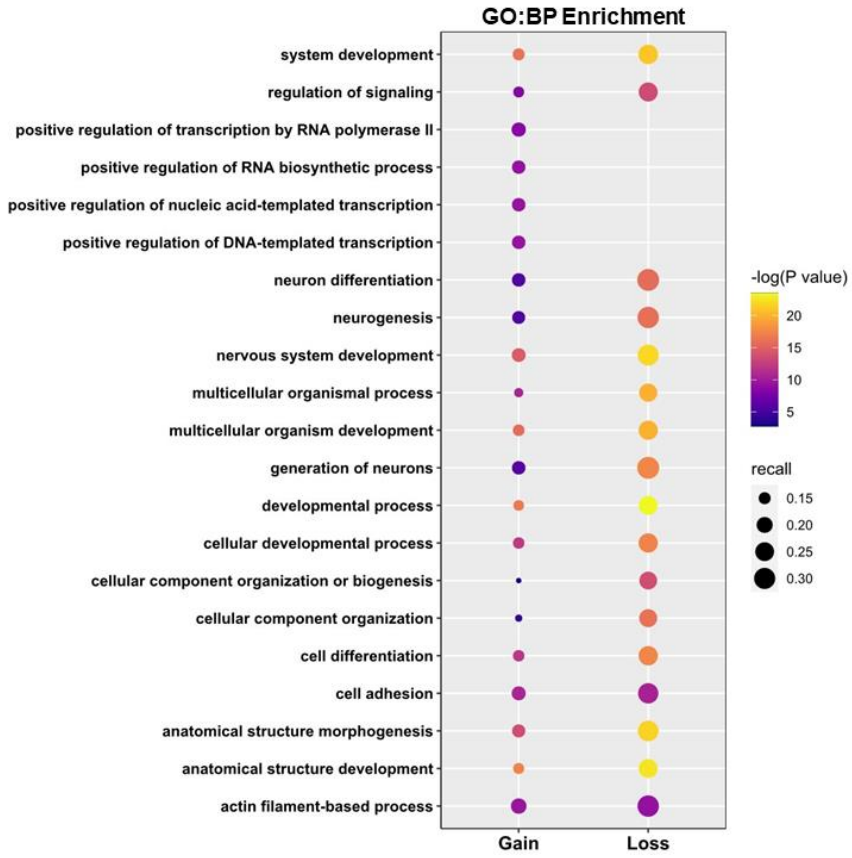
Appendix Figure A 10 Top 25 GO: BP enrichments for experimentally validated G4s overlapping enhancers, group 1.



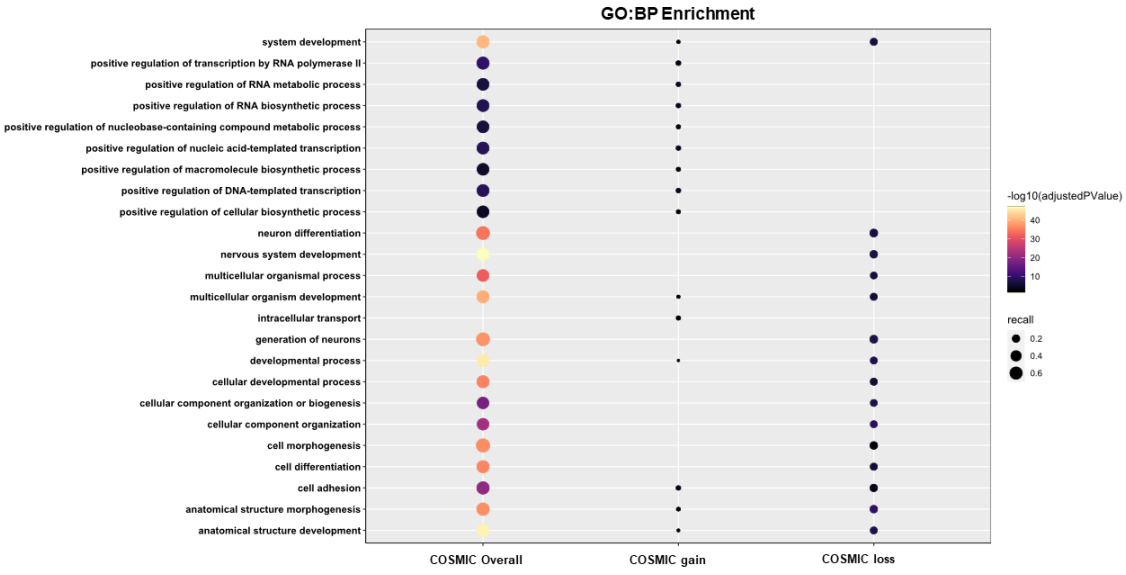
Appendix Figure A 11 Supplemental Figure 6. Top 25 GO: BP enrichments for experimentally validated G4s overlapping enhancers, group 2.



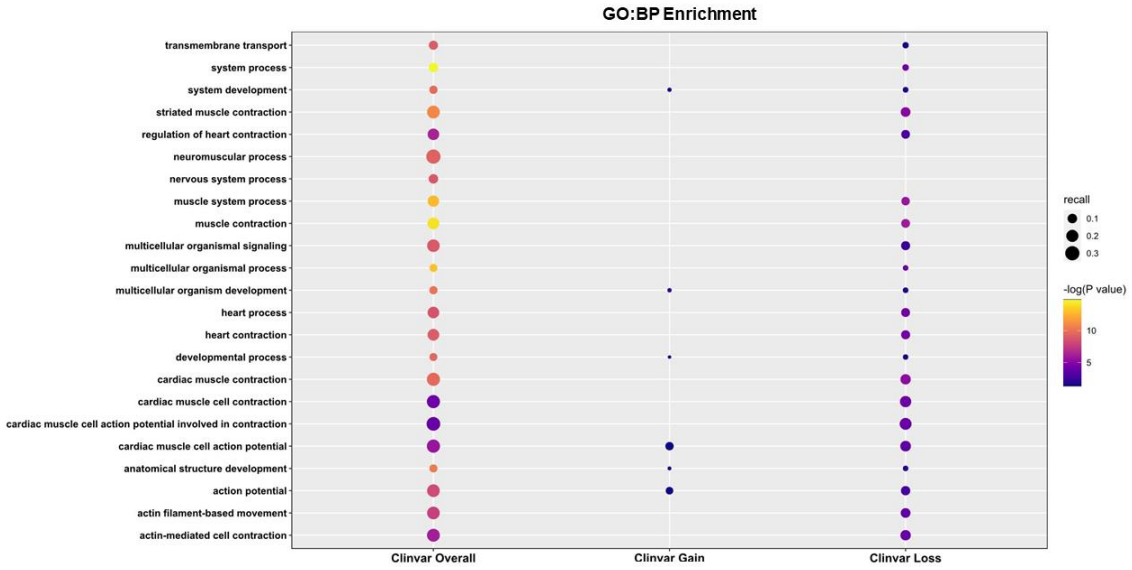
Appendix Figure A 12 Correlation of selected enhancers consisting of pG4 with gene expression in multiple cell types utilizing the epimap correlation group-link data.



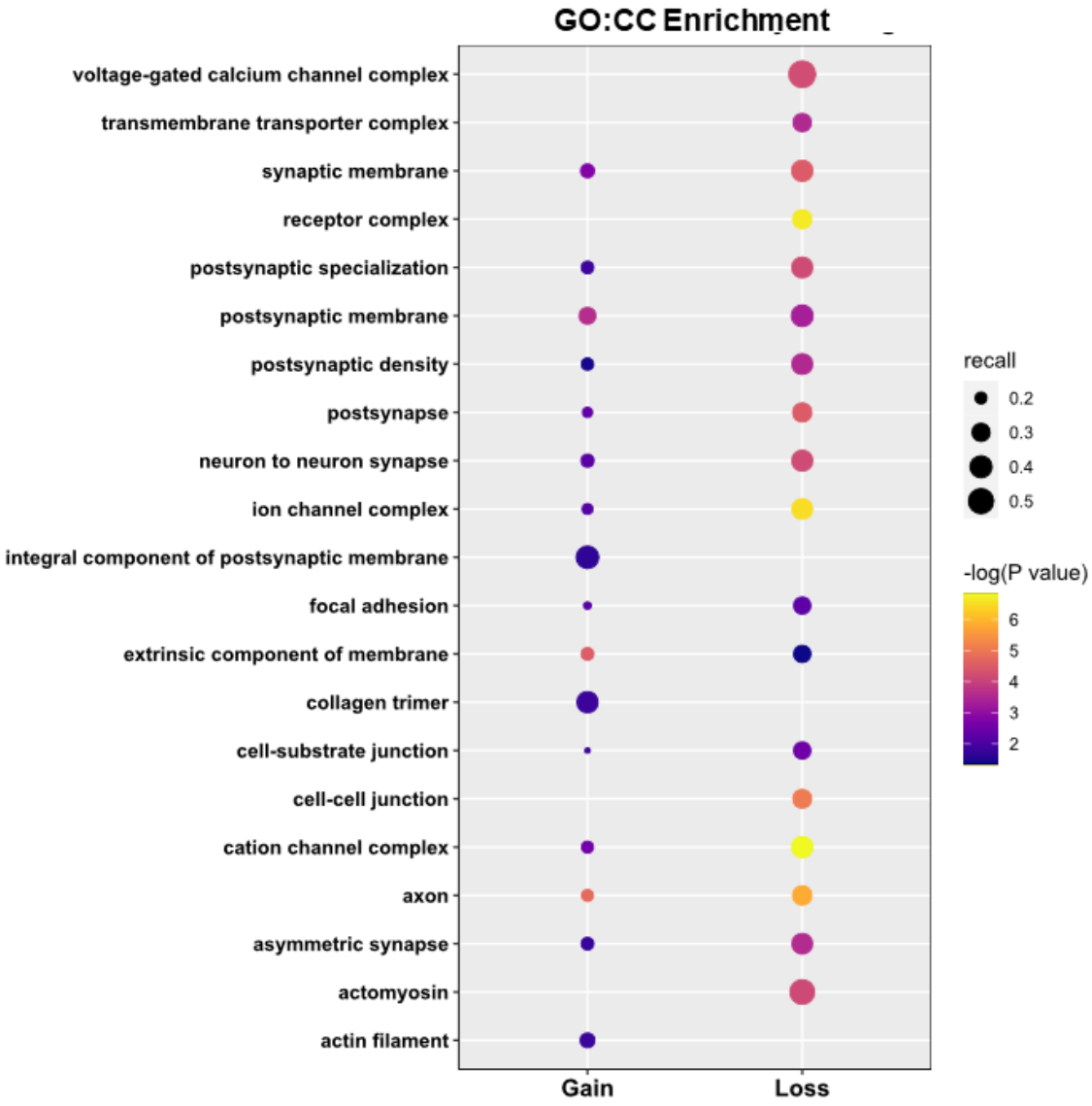
Appendix Figure B 1 Top 25 enriched GO: BP terms for COSMIC and CLINVAR G4 mutations.



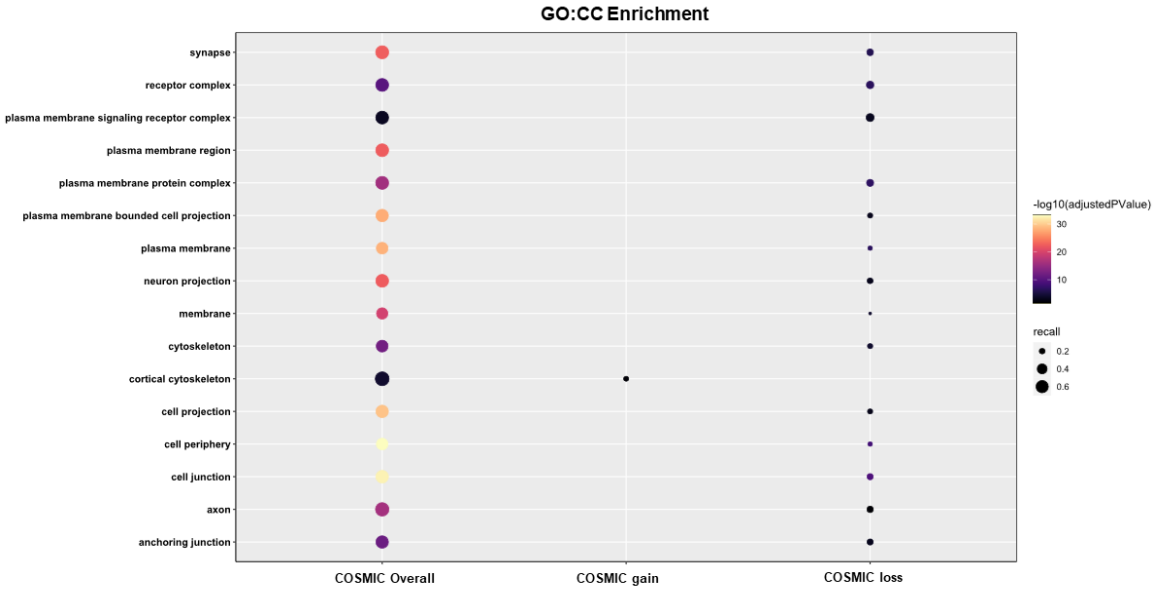
Appendix Figure B 2: Enriched GO:BP terms for G4 mutations.`



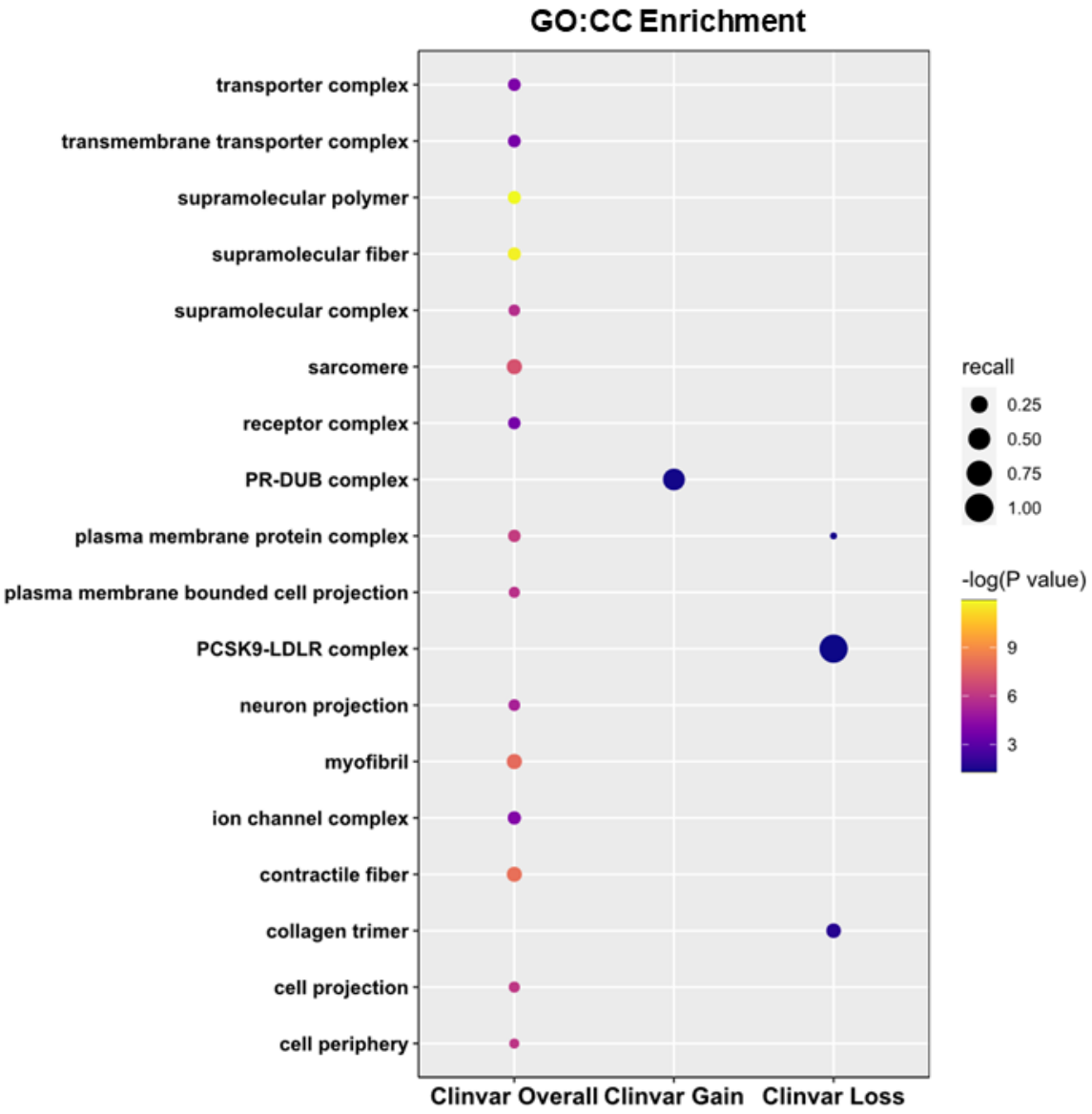
Appendix Figure B 3: Top 25 enriched GO:BP terms for CLINVAR G4 mutations.



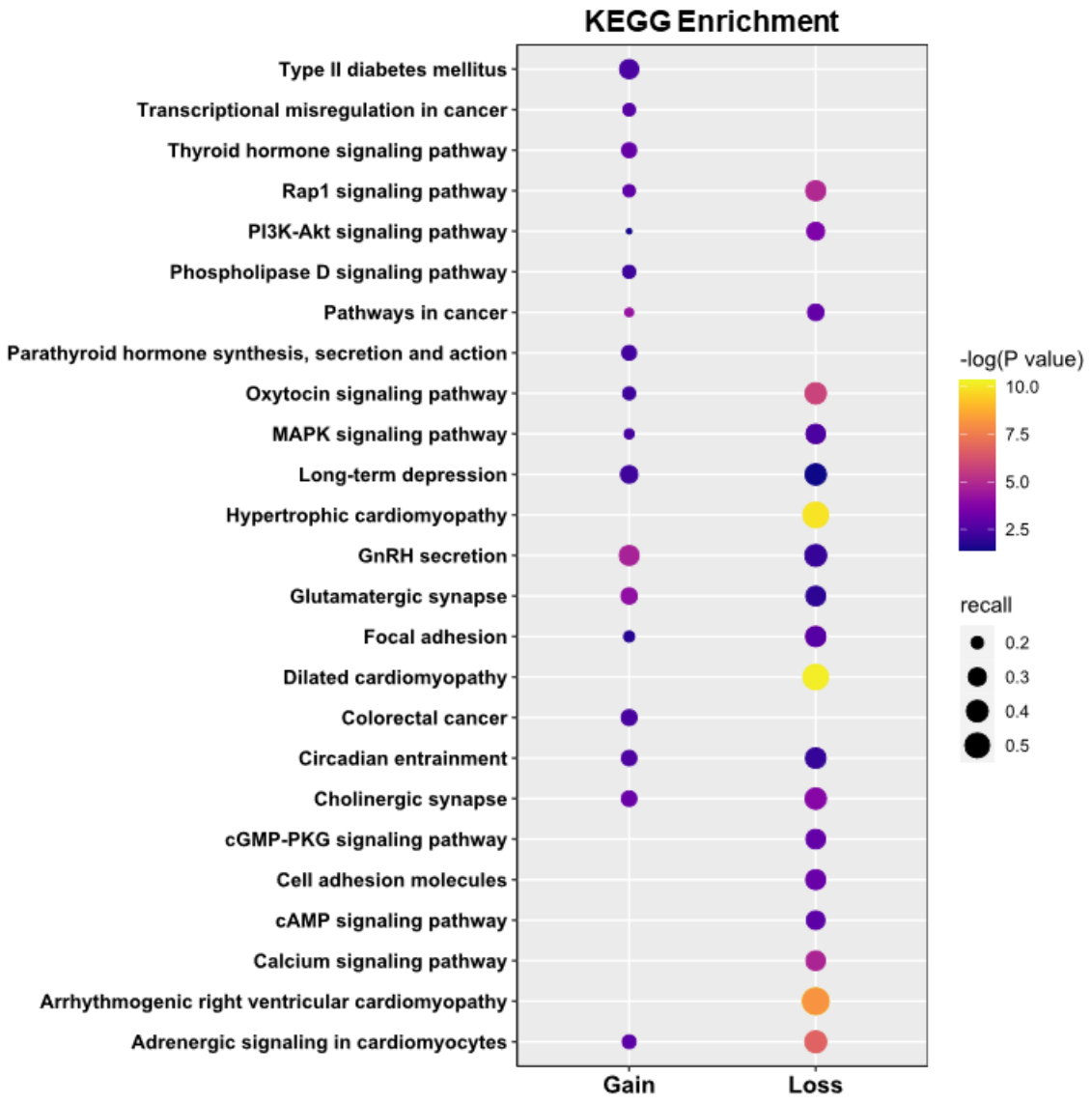
Appendix Figure B 4: Top25 enriched GO:CC terms for COSMIC and CLINVAR G4 mutations.



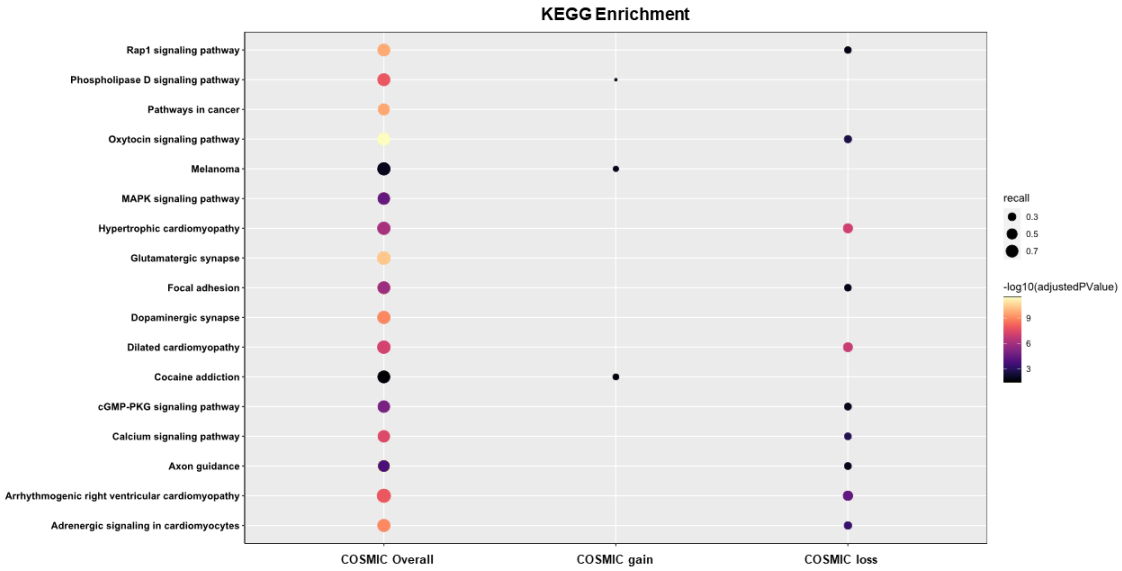
Appendix Figure B 5:Top 25 enriched GO:CC terms for COSMIC G4 mutations.



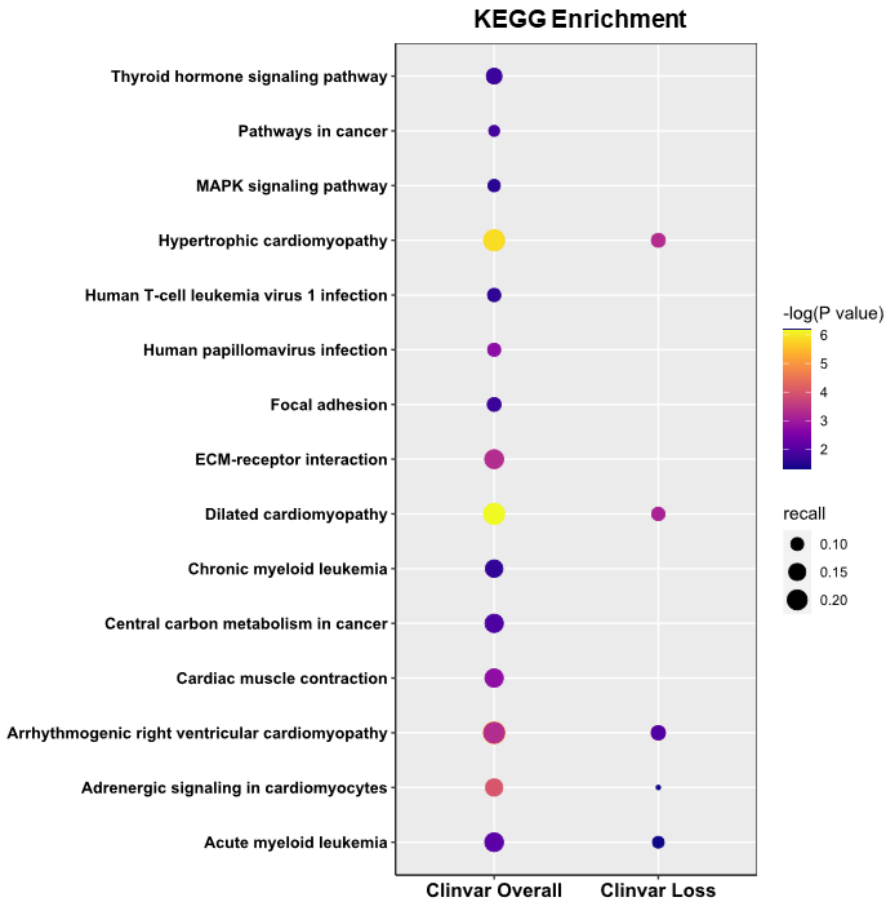
Appendix Figure B 6: Top 25 enriched GO:CC terms for CLINVAR G4 mutations.



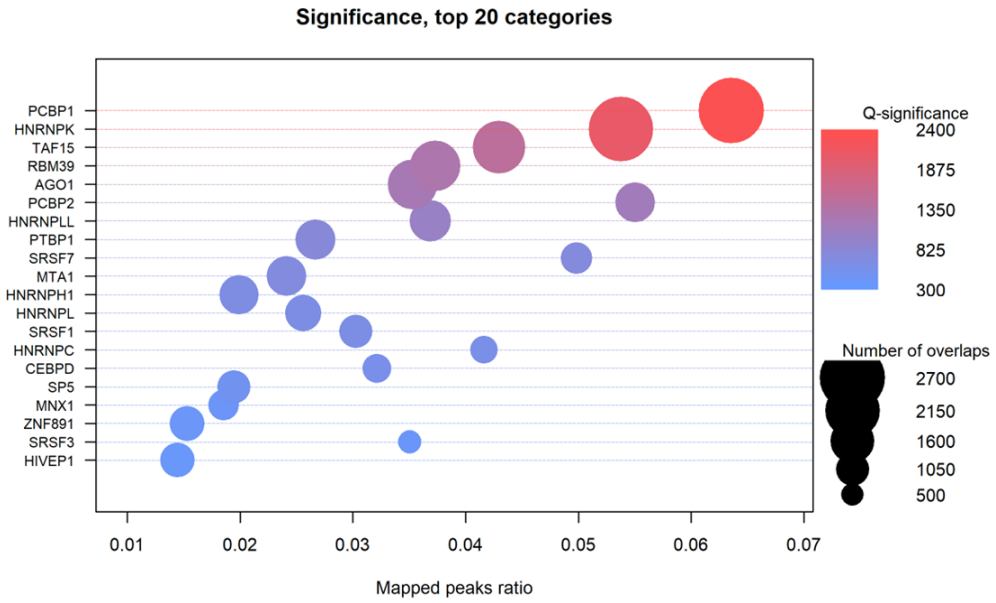
Appendix Figure B 7: Top 25 enriched KEGG terms for COSMIC and CLINVAR Gain and loss muttions



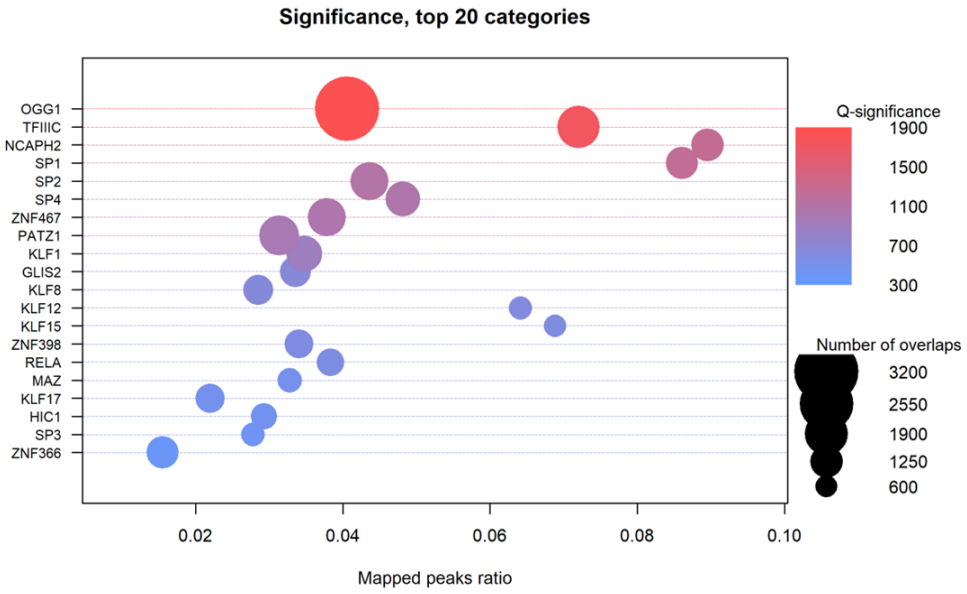
Appendix Figure B 8: Top 25 enriched KEGG terms for COSMIC G4 mutations.



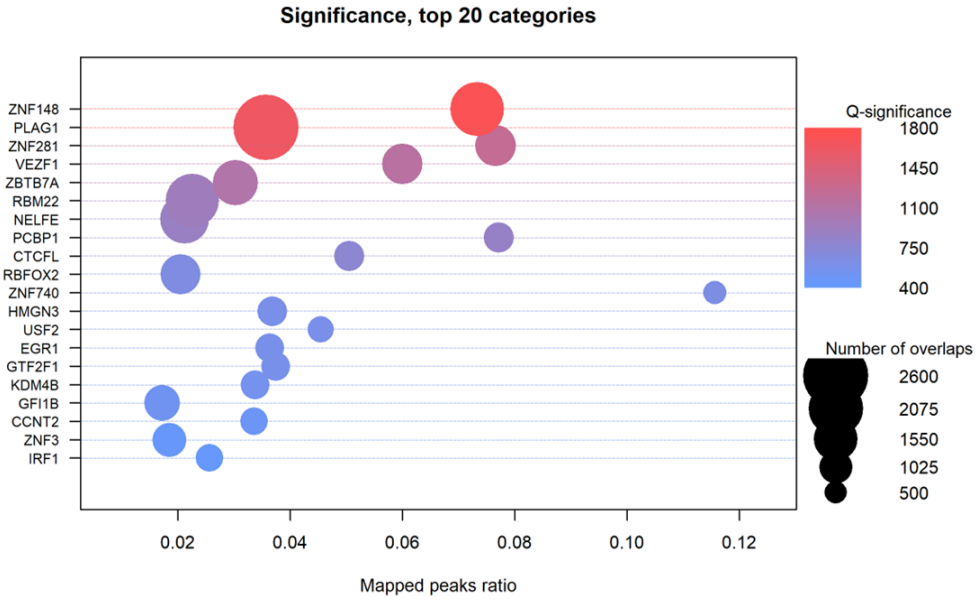
Appendix Figure B 9: Top 25 enriched KEGG terms for CLINVAR G4 mutations.



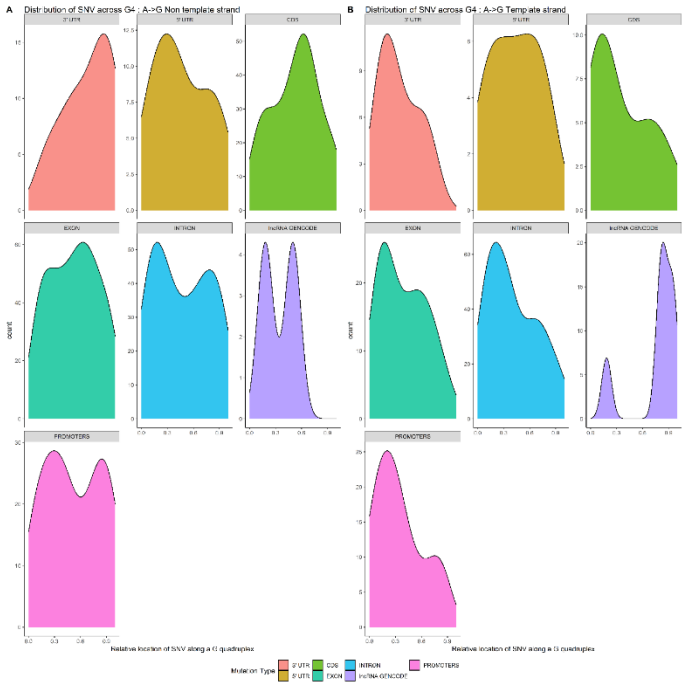
Appendix Figure B 10: Top 20 enriched transcription factors with overlapping ChIP-seq peaks for COSMIC and CLINVAR G4 SNVs in the HEK293 cell line.



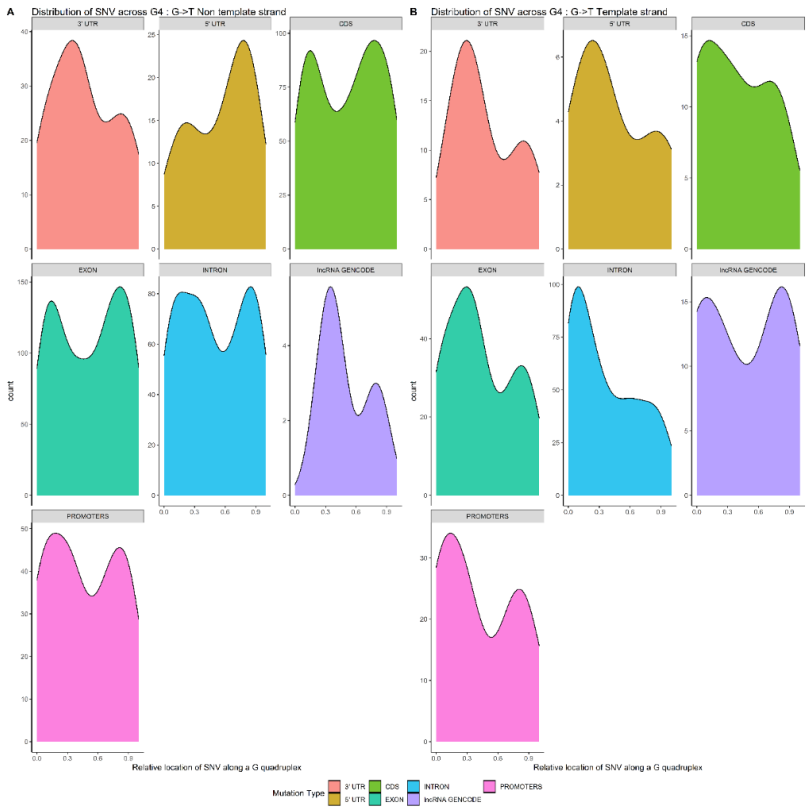
Appendix Figure B 11: Top 20 enriched transcription factors with overlapping ChIP-seq peaks for COSMIC and CLINVAR G4 SNVs in the K562 cell line.



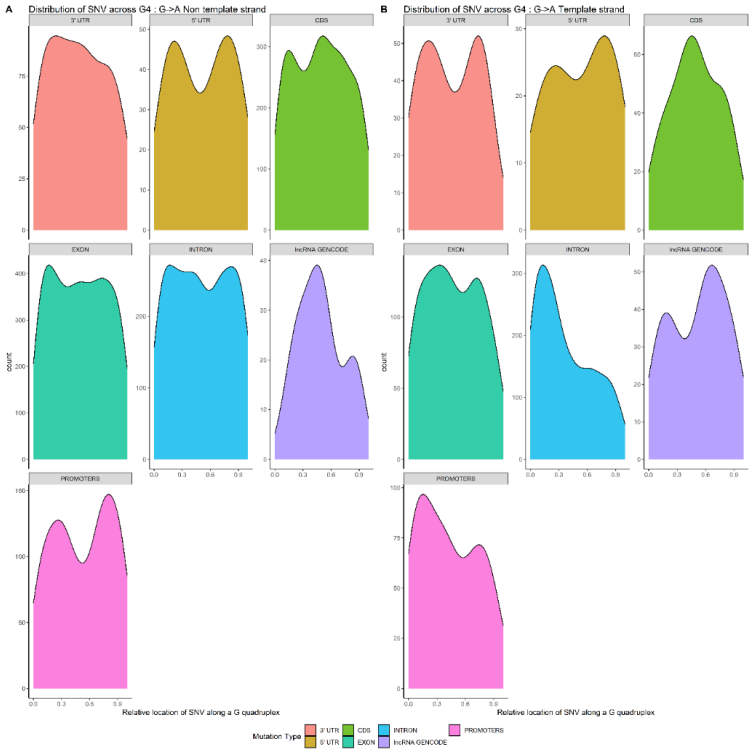
Appendix Figure B 12: Top 20 enriched transcription factors with overlapping ChIP-seq peaks for COSMIC and CLINVAR G4 SNVs in the Hep-G2 cell line.



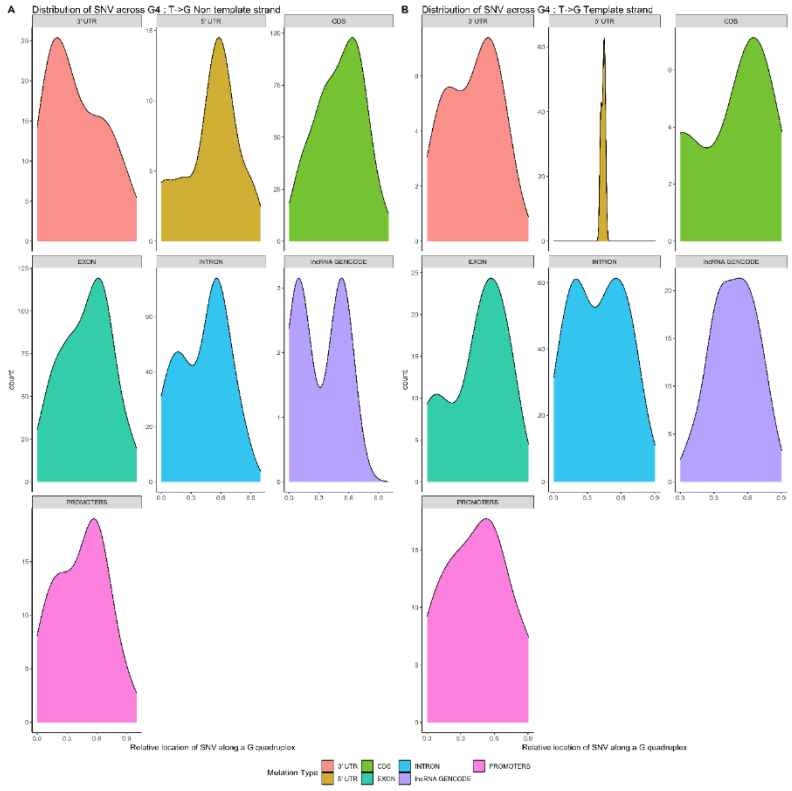
Appendix Figure B 13: Distribution of A→G SNVs across the G4 region for different features on (A) the non-template and (B) template strand.



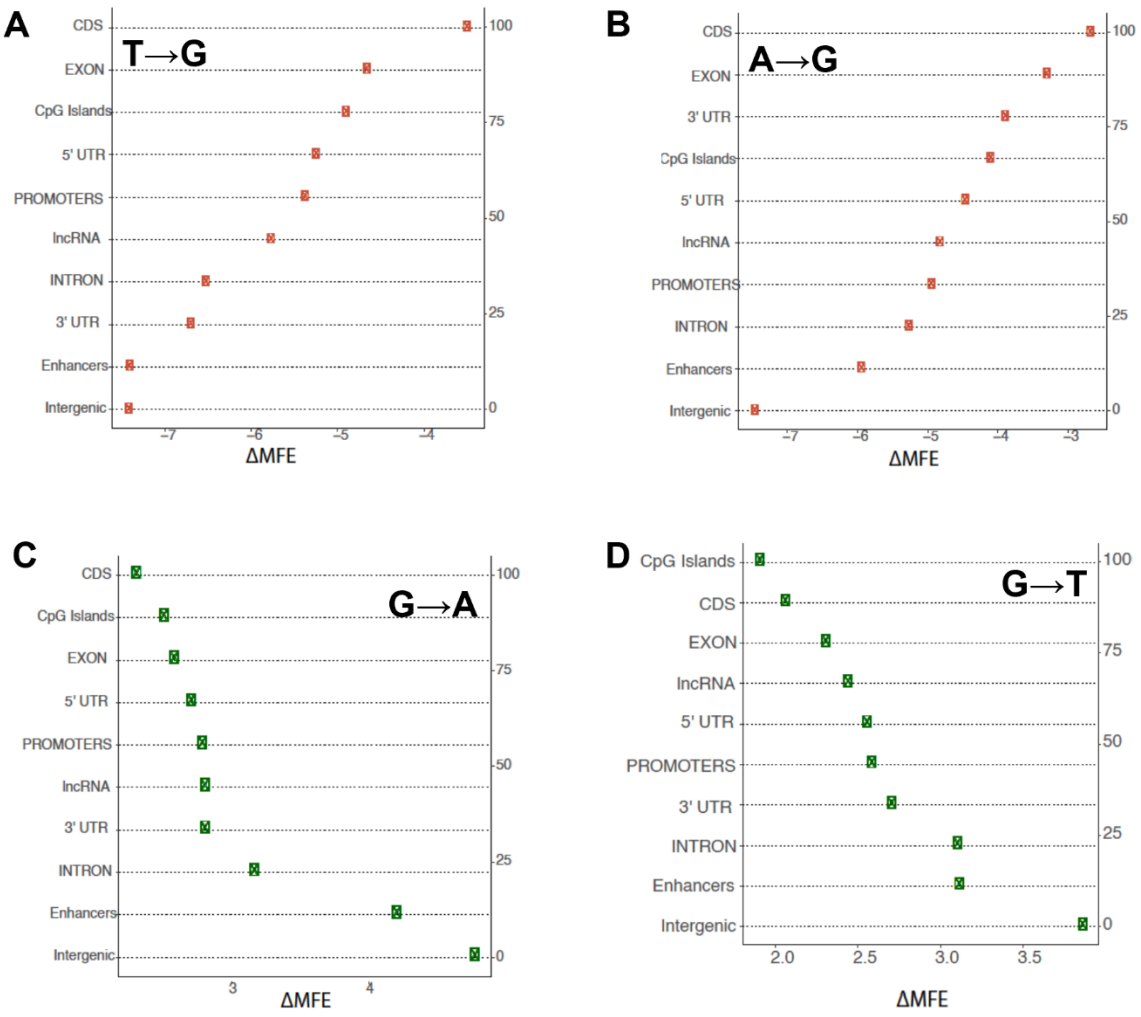
Appendix Figure B 14: Distribution of G→T SNVs across the G4 region for different features on (A) the non-template and (B) template strand.



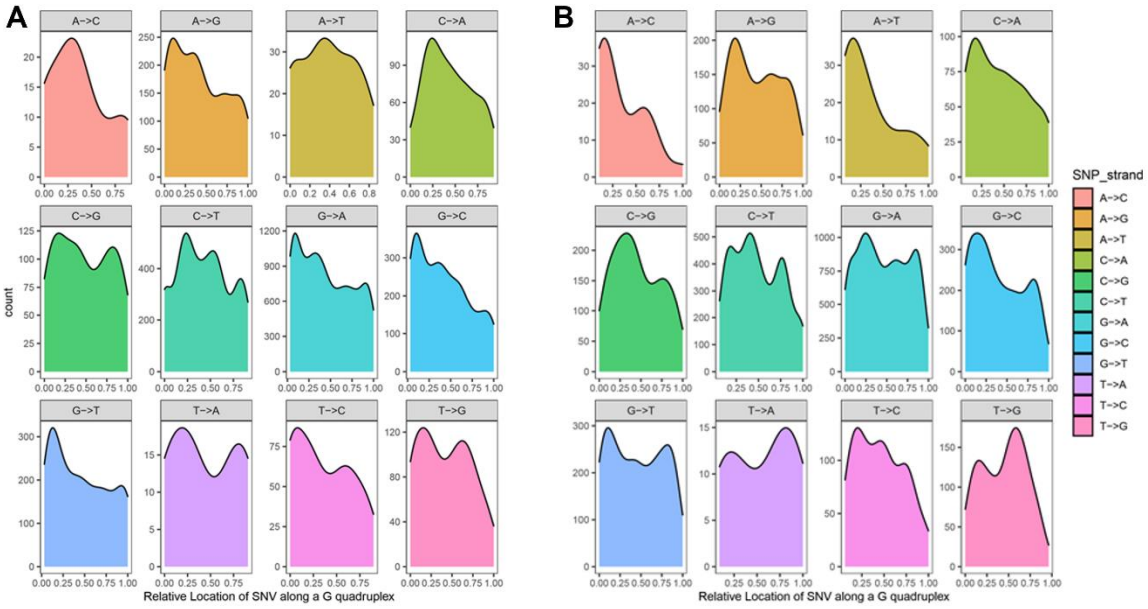
Appendix Figure B 15: Distribution of G→A SNVs across the G4 region for different features on (A) the non-template and (B) template strand.



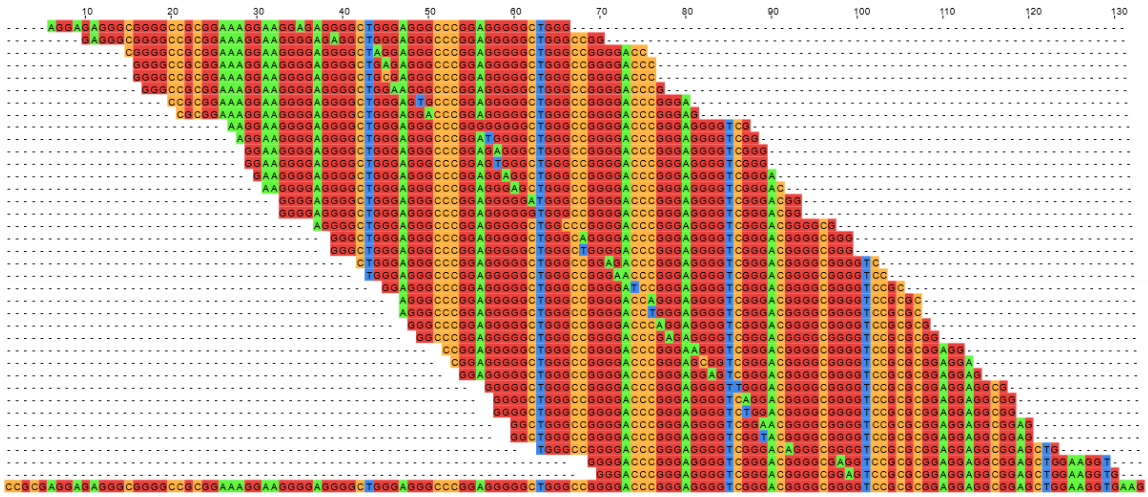
Appendix Figure B 16: Distribution of T→G SNVs across the G4 region for different features on (A) the non-template and (B) template strand.



Appendix Figure B 17: Effect of each SNV on Δ MFE of G4 on different annotations with percentile of the counts shown in the secondary y axis. Shown is (A) T→G SNVs; (B) A→G SNVs; (C) G→A SNVs; and (D) G→T SNVs.



Appendix Figure B 18: Distribution of SNVs across G-Quadruplex regions for the (A) forward and (B) reverse strands for SNVs detected in the CLINVAR database.



Appendix Figure B 19: G4 sequence along with variants along a TERT promoter

CURRICULUM VITAE

NAME: Vitae Aryan Neupane

ADDRESS: Bioinformatics Lab, Duthie Center For Engineering,

222 Eastern Pkwy, Louisville, KY 40208

DOB: Kathmandu, Nepal- July 7,1995

EDUCATION & TRAINING:

Bachelor of Science,
Biotechnology,Kathmandu University, Nepal
Aug 2012 - Nov 2016

AWARDS:

Best Poster Award in Bioinformatics and Data Science,
2019 Southeast Regional IDeA Conference,
November 6-8,2019

PUBLICATIONS:

Structural and Functional Classification of G-Quadruplex Families within the Human Genome, Mdpi Genes 2023, 14(3), 645;
https://doi.org/10.3390/genes14030645

Aryan Neupane, Julia H. Chariker, Eric C. Rouchka, Analysis of nucleotide variations in human g-quadruplex forming regions associated with disease states [Submitted, Feb 2, 2023]
https://doi.org/10.1101/2023.01.30.526341

Optical properties of natural dyes: prospect of application in dye sensitized solar cells (DSSCs) and organic light emitting diodes (OLEDs):
https://doi.org/10.26656/fr.2017.2(5).096