Investigating the Frequency of DNA Words near Loci Associated with Human Complex Traits

by

KARTIKAY CHADHA

A thesis submitted in conformity with the requirements for the degree of Master of Science
Institute of Medical Science
University of Toronto

© Copyright by Kartikay Chadha 2018
Investigating the Frequency of DNA Words near Loci Associated with Human Complex Traits

Kartikay Chadha
Master of Science
Institute of Medical Science
University of Toronto
2018

Abstract

Genome-wide association studies and expression quantitative trait loci (eQTL) studies have identified thousands of variants associated with complex diseases and gene expression levels. The frequency of DNA words associated with these variants has not been extensively evaluated. These words may help understand the biological role of trait-associated variants and also enable their identification in future studies.

An exact word-counting method was developed to investigate the hypothesis that short DNA words have different frequencies near single nucleotide polymorphisms (SNPs) associated with (1) Alzheimer’s disease and (2) thyroid eQTLs, compared to the rest of the genome.

No significant DNA words were found near AD associated SNPs. Some words enriched in GC content have significantly higher frequency around thyroid’s eQTLs compared to controls. These DNA words were no longer significant when the controls were matched for nucleotide frequency, but this is likely due to over-matching.
Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisor Dr. Jo Knight and co-supervisor Dr. Andrew Paterson for their continuous support, patience, motivation, enthusiasm, and immense knowledge, throughout the research during my masters program at the University of Toronto, Canada. Their guidance helped me throughout the research and writing of this thesis. I could not have imagined having better supervisors and mentors for my graduate studies. Besides my supervisor, I would like to thank Dr. Michael Wilson and Dr. Mario Masellis for their insightful feedback and guidance as members of my Program Advisory Committee at the Institute of Medical Science (IMS).

My sincere thanks also goes to Dr. Shelley Bull and Dr. Lei Sun without whom this could not have been possible. I also thank Dr. Michael Hoffman, Dr. Boris Steipe at the University of Toronto, and Dr. James Engert at McGill University for their insightful feedback on my research work, in capacity of my examiners. I thank my fellow lab mates at the Centre for Addiction and Mental Health: Dr. Sarah Gagliano, Ms. Sejal Patel, Mr. Eric Huang, Ms. Victoria Marshe, Ms. Venuja Sri, Ms. Amanda Lisoway and Mr. Sheraz Cheema for the stimulating discussions and for all the fun we have had in the last few years. I thank my on-site manager Ms. Mehrab Ali for her constant support during my masters program. Also I thank my friends in Toronto (Mr. Vincent Lai, Mr. Devansh Dabral, Ms. Devina Aggarwal and Ms. Mandi Ta), United Kingdom (Mr. Patrick Pearson, Ms. Holly Philip and Ms. Nicole Jenson), and India (Mr. Nikhil Mehra, Mr. Abhijeet Singh Dogra and Ms. Mayuri Jain). In particular, I am grateful to Dr. Martha Ladly, Dr. Rao Sethumadhavan and Dr. Febin Prabhu Dass J for enlightening me the first glance of research. I would also like to special thank the director of education at IMS, Dr. Howard Mount, and graduate coordinator Ms. Michelle Rosen for their encouragement and support. Last but not least, I would like to thank my mom (Ms. Punam Chadha), my dad (Mr. Rajeev K. Chadha), my aunt (Ms. Uma Chadha) and my sisters (Ms. Kriti Kakkar and Ms. Shruti Kapoor) for their unbelievable and constant support throughout the course of my studies.
# Table of Contents

Acknowledgments .................................................................................................................. iii

Table of Contents .................................................................................................................. iv

List of Tables ......................................................................................................................... ix

List of Figures ......................................................................................................................... x

List of Abbreviations ........................................................................................................... xii

Chapter 1 Literature Review .................................................................................................. 1

1.1. Genes and gene expression ......................................................................................... 2

1.2. Variations in the genome ............................................................................................ 2

  1.2.1. Type of genetic variations ....................................................................................... 3

  1.2.2. Correlation between SNPs ...................................................................................... 4

1.3. Association between variants and complex traits or diseases .................................. 5

  1.3.1. GWAS study design ................................................................................................. 5

  1.3.2. Outcome and impact of GWAS ............................................................................... 6

1.4. Alzheimer’s disease (AD) ............................................................................................ 6

  1.4.1. Clinical and pathological diagnosis of AD ................................................................. 7

  1.4.2. Stages of AD ........................................................................................................... 7

  1.4.3. Demographics and genetics of AD .......................................................................... 8

    1.4.3.1. Causative genes for AD ...................................................................................... 8

    1.4.3.2. Apolipoprotein E (APOE): The strongest genetic risk factor associated with late
            onset of AD ................................................................................................................. 9

  1.4.4. GWAS for AD ......................................................................................................... 10

1.5. Expression Quantitative Trait Loci (eQTLs) ............................................................... 13

  1.5.1. Identification of eQTLs .......................................................................................... 14

  1.5.2. GWAS and eQTLs ................................................................................................. 14

1.6. Exploring GWAS hits and eQTLs ............................................................................... 15

  1.6.1. Fine Mapping ........................................................................................................ 16

  1.6.2. Prioritizing variations based on functional annotations ....................................... 16

1.7. Motif discovery: Potential method to analyze trait associated variants .................... 17

  1.7.1. Motif analysis tools ............................................................................................... 18
1.7.1.1. Discriminative Regular Expression Motif Elicitation (DREME) .................... 19
1.7.1.2. Position weight matrix ................................................................................. 19
1.7.1.3. Regulatory Sequence Analysis Tool (RSAT) ............................................ 20

1.8. Summary and rationale ......................................................................................... 21

Chapter 2 Thesis Aims and Hypotheses ..................................................................... 22
2.1. Aims and Hypotheses ........................................................................................... 23
2.2. Structure of the Thesis .......................................................................................... 25

Chapter 3 DNA words near Alzheimer’s disease (AD) loci identified by GWAS ............ 26
3.1. Introduction: The largest meta-GWAS for AD .................................................... 27
3.2. Methods: DNA word discovery ............................................................................ 30
   3.2.1. Data selection ................................................................................................ 30
      3.2.1.1. Variants of interest (VOI) .................................................................... 30
      3.2.1.2. Control variants .................................................................................... 31
      3.2.1.3. SNPs in high Linkage disequilibrium (LD) ........................................ 31
      3.2.1.4. Validating matching parameters .......................................................... 32
      3.2.1.5. Exploring nucleotide substitutions ....................................................... 33
      3.2.1.6. Summary of data selection .................................................................. 33
      3.2.1.7. DNA sequences near VOI and control variants .................................. 33
      3.2.2. SNPs with overlapping DNA sequences .................................................. 33
         3.2.2.1. Algorithm for selecting SNPs with non-overlapping DNA ............. 34
         3.2.2.2. Implementation of the algorithm to select SNPs with non-overlapping DNA ... 35
      3.2.3. Calculating the frequency of DNA words ............................................... 36
      3.2.4. Statistical analysis ....................................................................................... 37
         3.2.4.1. Fisher’s exact test ............................................................................. 37
         3.2.4.2. Trend test ......................................................................................... 37
      3.2.5. Multiple test correction ............................................................................ 38
      3.2.6. Deviation from the null hypothesis ............................................................ 39
         3.2.6.1. More than expected significant results .............................................. 40
   3.3. Results: DNA words frequency near AD loci ................................................... 41
      3.3.1. SNPs selected for analysis ........................................................................ 41
         3.3.1.1. Extracting LD proxies ...................................................................... 41
3.3.1.2. Validating matching parameters .......................................................... 41
3.3.1.3. Exploring nucleotide substitutions ....................................................... 44
3.3.2. SNPs with no overlapping DNA ............................................................... 44
3.3.3. Statistical findings ................................................................................. 47
  3.3.3.1. Fisher’s exact test ............................................................................. 47
  3.3.3.2. Trend Test ....................................................................................... 47
3.3.4. Multiple testing correction .................................................................... 49
3.3.5. Deviation from the null hypothesis ......................................................... 49
  3.3.5.1. Nominally significant groups of DNA words .................................... 50
3.3.6. Discussion ............................................................................................. 51

Chapter 4 DNA words near thyroid eQTLs identified by the Gene-Tissue Expression
(GTEx) Project ..................................................................................................... 52
4.1. Introduction: The GTEx Project ................................................................. 53
  4.1.1. GTEx data collection and processing ..................................................... 53
  4.1.2. GTEx methods for cis-eQTL analysis ................................................... 54
  4.1.3. Results of the GTEx project ................................................................ 55
4.2. Methods: DNA word discovery ................................................................. 55
  4.2.1. Pre-processing of GTEx files to select VOI .......................................... 55
  4.2.2. Primary analysis: Control variants matched for MAF and nucleotide frequency .... 56
    4.2.2.1. Selection of control variants .......................................................... 56
    4.2.2.2. Validating matching parameters ..................................................... 56
    4.2.2.3. Exploring nucleotide substitutions ............................................... 57
    4.2.2.4. SNPs with overlapping DNA sequences ........................................ 57
    4.2.2.5. Extracting DNA sequences ............................................................ 57
    4.2.2.6. Statistical analysis ........................................................................ 57
    4.2.2.7. Multiple testing correction ............................................................. 57
    4.2.2.8. Deviation from the null hypothesis ................................................ 58
  4.2.3. Secondary analysis: Control variants matched for MAF only ................. 59
  4.2.4. Tertiary analysis: Including LD proxies ............................................... 59
4.3. Results ........................................................................................................ 60
  4.3.1. Primary analysis: Control variants matched for MAF and nucleotide frequency .... 60
    4.3.1.1. Validating matching parameters ..................................................... 62
4.3.1.2. Exploring nucleotide substitutions .............................................. 63
4.3.1.3. Statistical findings .................................................................... 65
4.3.1.4. Multiple testing correction ....................................................... 67
4.3.1.5. Groups of DNA words with p-value equal to 1 .......................... 67
4.3.2. Secondary analysis: Control variants matched for MAF only .......... 69
   4.3.2.1. Validating matching parameter ............................................. 69
   4.3.2.2. Statistical findings .................................................................. 70
   4.3.2.3. Multiple Test Correction ....................................................... 72
   4.3.2.4. Nucleotide composition in SOI and control sequences .............. 73
4.3.3. Tertiary analysis: Including LD proxies for VOI and control variants 74
   4.3.3.1. VOI and control variants ...................................................... 74
   4.3.3.2. SNPs in high Linkage disequilibrium (LD) ............................ 74
   4.3.3.3. SNPs with overlapping DNA ................................................ 77
   4.3.3.4. Statistical findings ............................................................... 77
   4.3.3.5. Multiple Test correction ....................................................... 78
4.4. **Discussion** .................................................................................. 79
   4.4.1. Comparing the three analyses .................................................. 79
   4.4.2. Functional interpretation of significant DNA words around thyroid eQTLs 81
   4.4.3. The 1000 Genomes project: phase 1 vs. phase 3 data ................ 84
   4.4.4. Summary ................................................................................. 85

Chapter 5 General Discussion ................................................................ 86
5.1. **Comparison with known motif discovery algorithms** ................... 88
5.2. **DNA word discovery: AD GWAS hits versus thyroid eQTLs** .......... 89
5.3. **Data file formats and other computational challenges** ................... 90
5.4. **Limitations of my method** ......................................................... 91
   5.4.1. SNPs with overlapping DNA .................................................. 91
   5.4.2. Genetic Sequence characteristics ............................................. 92
   5.4.3. Multiple testing correction ...................................................... 93
5.5. **Future Directions** ....................................................................... 94
   5.5.1. Alternate data sources ............................................................. 94
     5.5.1.1. Variations in the GWAS catalog ....................................... 94
     5.5.1.2. Further exploration of the GTEx project .............................. 95
5.5.1.3. Whole blood eQTLs......................................................................................... 95

5.5.2. Alternate matching parameters........................................................................ 96
  5.5.2.1. SNPsnap parameters in eQTL data................................................................. 96
  5.5.2.2. The inter-marker distance between LD proxies............................................. 96

5.5.3. Search for different length of DNA words and regions analyzed............... 96

5.5.4. Identification of additional risk loci ............................................................... 97

References.................................................................................................................. 98

Contributions ............................................................................................................. 111
## List of Tables

Table 1.1. Type of variants in the NHGRI-EBI GWAS Catalogue for AD and related phenotypes. ........................................................................................................................................................................................................................................................................................................... 11

Table 3.1. Loci associated with AD from stage 1 or stage 2 or combined analysis Lambert et al. meta-analysis .............................................................................................................................................................................................................................................................................................................. 29

Table 3.2. Fisher’s exact test 2x2 contingency table. .............................................................................................................................................................................................................................................................................................................. 37

Table 3.3. Contingency table for Cochran Armitage Trend Test for the group of DNA words with most significant Fisher’s exact p-value (group: ATGAG / CTCAT) .............................................................................................................................................................................................................................................................................................. 38

Table 3.4. Control variants pairwise matched with VOI using SNPsnap. .............................................................................................................................................................................................................................................................................................................. 42

Table 3.5. Top 15 significant DNA words at Fisher’s exact test p-value threshold < 0.05 ........ 48

Table 4.1. 2x6 contingency tables to test difference in distribution of nucleotide composition between SOI and control sequences using Cochran Armitage trend test. .............................................................................................................................................................................................................................................................................................................. 63

Table 4.2. χ² test results for 12 nucleotide substitutions in VOI and control variants. .......... 64

Table 4.3. DNA words absent in both SOI and control sequences. ........................................ 65

Table 4.4. Groups of DNA words at Fisher’s exact test p-value < 0.05. ............................... 66

Table 4.5. DNA words absent in both SOI and control sequences. ........................................ 70

Table 4.6. Top 10 DNA words nominally significant at Fisher’s exact test p-value <0.05. .... 71

Table 4.7. 2x6 contingency tables to test difference in distribution of nucleotide composition between SOI and control sequences using Cochran Armitage trend test. .............................................................................................................................................................................................................................................................................................................. 73

Table 4.8. Pairwise matched VOI and control variants with large number of proxies (>1500)... 75

Table 4.9. Top 10 significant groups of DNA words from 512 Fisher’s Exact tests. .............. 78

Table 4.10. Flow table comparing data and results from three different GTEx analysis. ........ 80

Table 4.11. TFs matched with group 1 DNA words (GCGGG/CCCGC) and significant at FDR q-value < 0.05 .............................................................................................................................................................................................................................................................................................................. 83

Table 4.12. TFs matched with group 2 DNA words (CCGGG/CCCGG) and significant at nominal p-value < 0.05 .............................................................................................................................................................................................................................................................................................................. 84

Table 5.1. Comparison between analyses using AD GWAS hits and thyroid eQTLs. ............. 89
List of Figures

Figure 1.1. Manhattan plot of stage 1 GWAS analysis (Lambert et. al., 2013) .......................... 12
Figure 1.2. Typical illustration of a sequence logo representing PWM ................................. 20
Figure 3.1. Strategy of selecting SNPs with overlapping DNA sequences ......................... 35
Figure 3.2. Flowchart representing the steps for data pre-processing to selected variants of interest (VOI) and control variants. The data is extracted from the largest GWAS for Alzheimer’s disease (Lambert et al., 2013) .......................................................... 40
Figure 3.3. Distribution of MAF for VOI and control variants ............................................. 43
Figure 3.4. Distribution of LD proxies for VOI and control variants ................................. 43
Figure 3.5. Nucleotide substitution at VOI and control variants (Reference allele > alternate allele) .......................................................................................................................... 44
Figure 3.6. Number of LD proxies (blue + red bars) and number of SNPs dropped (red bars) across 20 VOI ......................................................................................................................... 45
Figure 3.7. Number of LD proxies (blue + red bars) and number of SNPs dropped (red bars) across 20 pairwise matched control variants .......................................................................................................................... 45
Figure 3.8. Chromosomal position of LD proxies (yellow) for rs6501579 (control variant; blue) in the human genome that has maximum number of drops (Machiela and Chanock, 2015) .... 46
Figure 3.9. Chromosomal position of LD proxies (yellow) for rs2019207 (Control variant; blue) in the human genome that has maximum number LD proxies (Machiela and Chanock, 2015) ... 46
Figure 3.10. P-value distribution for 512 Fisher’s exact tests (a) Histogram (b) QQ-plot ....... 47
Figure 3.11. Distribution of mean p-values from 1000 permutation runs ............................ 49
Figure 3.12. Distribution of mean p-values from permutation runs where rs10948363 and rs983392 (a) were tagged together either as VOI or control variants, (b) were tagged as VOI and control variant, respectively or vice versa .......................................................... 50
Figure 4.1. Flowchart representing the steps for data pre-processing to selected variants of interest (VOI) and control variants for primary and secondary analysis. The data is extracted from the GTEx project for thyroid tissue .......................................................... 61
Figure 4.2. Minor Allele Frequency (MAF) distribution for variants of Interest (VOI) (Black) and control variants (Grey) .......................................................................................................................... 62
Figure 4.3. Nucleotide substitution at VOI and Control variants (Reference allele > alternate allele) .......................................................................................................................... 64
Figure 4.4. P-value distribution for 509 Fisher’s exact tests. ................................................................. 65
Figure 4.5. Distribution of (a) mean p-values (b) Minimum p-values, from 1000 permutation runs. ............................................................................................................................................. 67
Figure 4.6. Distribution of proportion of groups of DNA words that matched SOI and control variants where (a) p-values = 1 (b) p-values < 1. ........................................................................................................................................... 68
Figure 4.7. Minor Allele Frequency (MAF) distribution for Variants of Interest (VOI) and Control variants.................................................................................................................................................. 69
Figure 4.8. P-value distribution for 542 Fisher exact tests (a) Histogram (b) QQ-plot................... 70
Figure 4.9. Distribution of (a) mean p-values (b) Minimum p-values, from 1000 permutation runs............................................................................................................................................... 72
Figure 4.10. Flowchart representing the steps for data pre-processing to selected variants of interest (VOI) and control variants for the tertiary analysis. The data is extracted from the GTEx project for thyroid tissue. ........................................................................................................................... 76
Figure 4.11. P-value distribution for 512 Fisher’s Exact tests (a) Histogram (b) QQ-plot. ........... 77
Figure 5.1. Distribution of length of 537 human TFBS extracted from JASPER database ......... 97
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>Alzheimer's Disease</td>
</tr>
<tr>
<td>ADGC</td>
<td>Alzheimer’s disease Genetic Consortium</td>
</tr>
<tr>
<td>AME</td>
<td>Analysis of Motif Enrichment</td>
</tr>
<tr>
<td>ANNOVAR</td>
<td>Annotate Variation Tool</td>
</tr>
<tr>
<td>APOE</td>
<td>Apo-lipoprotein E</td>
</tr>
<tr>
<td>APP</td>
<td>Amyloid Beta Precursor Protein</td>
</tr>
<tr>
<td>ASE</td>
<td>Allele Specific Gene Expression</td>
</tr>
<tr>
<td>Aβ</td>
<td>Beta amyloid</td>
</tr>
<tr>
<td>BIN1</td>
<td>Bridging Integrator 1</td>
</tr>
<tr>
<td>Bp</td>
<td>Base Pairs</td>
</tr>
<tr>
<td>CEU</td>
<td>Utah Residents (CEPH) with Northern and Western European Ancestry</td>
</tr>
<tr>
<td>CHARGE</td>
<td>Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium</td>
</tr>
<tr>
<td>CHB</td>
<td>Han Chinese in Beijing, China</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>Chromatin immunoprecipitation- Sequencing</td>
</tr>
<tr>
<td>Chr</td>
<td>Chromosome</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence Interval</td>
</tr>
<tr>
<td>CLOVER</td>
<td>Cis-eLement OVer-representation</td>
</tr>
<tr>
<td>CLU</td>
<td>Clusterin Gene</td>
</tr>
<tr>
<td>CNVs</td>
<td>Copy Number Variants</td>
</tr>
<tr>
<td>CRI</td>
<td>Complement receptor type 1</td>
</tr>
<tr>
<td>dbSNP</td>
<td>Single Nucleotide Polymorphism database</td>
</tr>
<tr>
<td>DHS</td>
<td>DNase I Hypersensitive Sites</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>DNase</td>
<td>Deoxyribonuclease</td>
</tr>
<tr>
<td>DREME</td>
<td>Discriminative Regular Expression Motif Elicitation</td>
</tr>
<tr>
<td>E2F2</td>
<td>E2F Transcription Factor 4</td>
</tr>
<tr>
<td>E2F1</td>
<td>E2F Transcription Factor 1</td>
</tr>
<tr>
<td>EADI</td>
<td>European Alzheimer’s Disease Initiative</td>
</tr>
<tr>
<td>ELF1</td>
<td>ETS-related transcription factor Elf-1</td>
</tr>
<tr>
<td>EM</td>
<td>Estimation Maximization</td>
</tr>
<tr>
<td>ENCODE</td>
<td>Encyclopaedia of DNA Elements</td>
</tr>
<tr>
<td>EPHA1</td>
<td>EPH receptor A1</td>
</tr>
<tr>
<td>eQTL</td>
<td>Expression Quantitative Trait Loci</td>
</tr>
<tr>
<td>EUR</td>
<td>European</td>
</tr>
<tr>
<td>FANTOM5</td>
<td>The functional annotation of the mammalian genome 5</td>
</tr>
<tr>
<td>FDR</td>
<td>False Discovery Rate</td>
</tr>
<tr>
<td>GCM2</td>
<td>Glial Cells Missing Homolog 2</td>
</tr>
<tr>
<td>GERAD</td>
<td>Genetic and Environmental Risk in Alzheimer’s disease</td>
</tr>
<tr>
<td>GLAM2</td>
<td>Gapped Local Alignment of Motifs</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>GoMo</td>
<td>Gene Ontology for Motifs</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>GPCOG</td>
<td>General Practitioner Assessment of Cognition</td>
</tr>
<tr>
<td>GTEx</td>
<td>Gene Tissue Expression Project</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome Wide Association Studies</td>
</tr>
<tr>
<td>HapMap</td>
<td>International Haplotype Map Project</td>
</tr>
<tr>
<td>HLA</td>
<td>Human Leukocyte Antigen</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>HOCOMOCO</td>
<td>HOmo sapiens COmprehensive MOdel Collection</td>
</tr>
<tr>
<td>I-GAP</td>
<td>International Genomics of Alzheimer's Project</td>
</tr>
<tr>
<td>IQCODE</td>
<td>Information questionnaire on Cognitive Decline in Elderly</td>
</tr>
<tr>
<td>IUPAC</td>
<td>International Union of Pure and Applied Chemistry</td>
</tr>
<tr>
<td>Kb</td>
<td>Kilo base Pairs</td>
</tr>
<tr>
<td>KLF1</td>
<td>Krueppel-like factor 1</td>
</tr>
<tr>
<td>KLF3</td>
<td>Krueppel-like factor 3</td>
</tr>
<tr>
<td>LCLs</td>
<td>Lymphoblastoid cell lines</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage Disequilibrium</td>
</tr>
<tr>
<td>LincRNA</td>
<td>Long intergenic noncoding RNA</td>
</tr>
<tr>
<td>MAF</td>
<td>Minor Allele Frequency</td>
</tr>
<tr>
<td>Mb</td>
<td>Mega Base Pairs</td>
</tr>
<tr>
<td>MECP2</td>
<td>Methyl-CpD-binding protein</td>
</tr>
<tr>
<td>MEME</td>
<td>Multiple EM for Motif Enrichment</td>
</tr>
<tr>
<td>MHC</td>
<td>Major Histocompatibility Complex</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>NHGRI-EBI</td>
<td>National Human Genome Research Institute- The European Bioinformatics Institute</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institute of Health</td>
</tr>
<tr>
<td>OR</td>
<td>Odds Ratio</td>
</tr>
<tr>
<td>PICALM</td>
<td>Phosphatidylinositol Binding Clathrin Assembly Protein</td>
</tr>
<tr>
<td>Polyphen2</td>
<td>Polymorphism Phenotyping v2</td>
</tr>
<tr>
<td>PSN1</td>
<td>Presenilin 1</td>
</tr>
<tr>
<td>PSN2</td>
<td>Presenilin 2</td>
</tr>
<tr>
<td>PSSM</td>
<td>Position specific scoring matrices</td>
</tr>
<tr>
<td>PSWM</td>
<td>Position-specific weight matrices</td>
</tr>
<tr>
<td>PWM</td>
<td>Position weight matrices</td>
</tr>
<tr>
<td>QQ plot</td>
<td>Quantile-Quantile Plot</td>
</tr>
<tr>
<td>ORMNDL3</td>
<td>orosomucoid like 3</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>RSAT</td>
<td>Regulatory Sequence Analysis Tools</td>
</tr>
<tr>
<td>SE</td>
<td>Stretch Enhancers</td>
</tr>
<tr>
<td>SIFT</td>
<td>Sorting Tolerant From Intolerant</td>
</tr>
<tr>
<td>SNiPA</td>
<td>Single Nucleotide Polymorphism Annotator</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>SORL1</td>
<td>Sortilin Related Receptor 1</td>
</tr>
<tr>
<td>SOI</td>
<td>Sequences of Interest</td>
</tr>
<tr>
<td>SPI</td>
<td>Specificity Protein 1</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>SP2</td>
<td>Specificity Protein 2</td>
</tr>
<tr>
<td>SpaMo</td>
<td>Spaced Motif</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription Factors</td>
</tr>
<tr>
<td>TFBS</td>
<td>Transcription Factors Binding Sites</td>
</tr>
<tr>
<td>TFDP1</td>
<td>Transcription Factor Dp-1</td>
</tr>
<tr>
<td>TOMM40</td>
<td>Translocase Of Outer Mitochondrial Membrane 40</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription Start Sites</td>
</tr>
<tr>
<td>UCSC</td>
<td>University of California</td>
</tr>
<tr>
<td>VEP</td>
<td>Variant effect predictor</td>
</tr>
<tr>
<td>VOI</td>
<td>Variants of interest</td>
</tr>
<tr>
<td>YRI</td>
<td>Yoruba in Ibadan, Nigeria</td>
</tr>
<tr>
<td>ZF64</td>
<td>Zinc finger protein 64 homolog</td>
</tr>
</tbody>
</table>
Chapter 1
Literature Review

Some sections of this chapter are modified from a dissertation submitted at Lancaster University, United Kingdom towards an academic course taken under an international student exchange program. No credits were awarded or transferred for this course.
1.1. Genes and gene expression

The functional parts of the deoxyribonucleic acid (DNA) that code for proteins are called genes, which consist of both coding and non-coding regions. The central dogma of molecular biology states that the flow of information is from DNA to mRNA (messenger ribonucleic acid) to protein. The mRNA consists of exons that are translated to amino acids through groups of three nucleotide base pairs called codons. The polymeric chains of these amino acids form proteins that are incorporated into the formation of tissues, organs and hence the entire human body. Some genes code for non-protein coding RNAs, for example, micro-RNA (involved in translation regulation), small nuclear RNAs (involved in splicing), long nuclear RNAs (involved in ribosomal RNA modification) and enhancer RNA (involved in alteration of 3D architecture of chromosomes) (Palazzo and Lee, 2015; Boland, 2017). The non-coding regions of a gene are called introns. The regions of the DNA between two genes are termed intergenic and are also categorized as non-coding.

Activation of different genes through transcription in different cells determines the type and function of a cell. The process by which a gene gets transcribed and translated into its functional product is called gene expression. The gene products are usually proteins. In case of genes that are non-protein coding, the gene product is functional RNA. Some genes are expressed in all cells to maintain the basic regulatory cellular functions under normal conditions and are called housekeeping genes. Enzymes like RNA polymerase and regulatory factors like transcription factors (TFs), play a major role in gene expression. TFs bind to the DNA in the promoter regions, at transcription start sites (TSS). In some cases, TFs binds to the enhancer region of a gene to regulate the gene expression level by stimulating or repressing transcription.

1.2. Variations in the genome

The observable characteristics of an organism are called its phenotype. Genetic variations have the potential to alter gene expression that may result in variations in a phenotype across a population. Heritability is the proportion of variation in a phenotype that is attributable to genetic variations. Genetic variations can be responsible for increasing or reducing one’s risk of developing a trait or disease.
1.2.1. Type of genetic variations

A genetic variation can be caused due to substitution, insertion or deletion (indels), or inversion of nucleotides. Different variants of nucleotide(s) at specific locations are called alleles. A variation in protein-coding DNA that does not change the amino acid sequence due to the redundancy in the genetic code is called synonymous. A non-synonymous variation causes alteration of the corresponding amino acid sequence and can further divided into multiple types. A missense change results in a protein that has a different amino acid sequence at one position, while a nonsense change results in a premature stop codon that terminates the translation of DNA into mRNA. Indels may cause a shift in three base pair reading frame and result into alteration of the polypeptide chain. Such variations are called frame-shift. The variations in the in intergenic and intronic regions are non-coding variants. Non-coding variants constitute the vast majority of variants in the human genome (The 1000 Genomes Project Consortium, 2015).

In this thesis, I have focused only on single nucleotide polymorphisms (SNPs) for which there is one-nucleotide change at the loci, i.e. bi-allelic (as opposed to tri-allelic or tetra-allelic). Within a population, the frequency of the less common allele is termed the minor allele frequency (MAF). SNPs with MAF <1% are considered rare variants (Panoutsopoulou et al., 2013).

Genotyping projects like the HapMap project (International HapMap Consortium, 2003; Thorisson et al., 2005) and sequencing projects like the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2012, 2015) have identified millions of SNPs in the human genome. The 1000 Genomes project has also identified indels, microsatellites, copy number variants (CNVs) and structural variants (Sudmant et al., 2015). dbSNP is the central repository for genetic variations in many organisms including humans (Sherry et al., 2001). The latest release of dbSNP is build 150, which was released in February 2017 and included over 300 million human genetic variations (~150 million new submissions compared to previous build).
1.2.2. Correlation between SNPs

During cell reproduction, alleles at nearby loci are often inherited together from either the paternal or maternal chromosomes into the new gametes. These set of alleles are called haplotypes and the non-random association between alleles at different loci is called linkage disequilibrium (LD).

**Linkage Disequilibrium (LD):**

Loci are said to be in LD when the joint distribution of alleles is different from what we expect in case of random/independent assortment (Slatkin, 2008). Alleles at two loci are presumed to be independent when the product of the frequency of allele A at one locus (\(P_A\)) and frequency of allele B at another locus (\(P_B\)) i.e., \(P_A P_B\) is equal to the frequency of occurrence of both alleles (A and B) together (\(P_{AB}\)). However, there is said to be linkage disequilibrium when \(P_A P_B\) differs from \(P_{AB}\). This is quantified by LD coefficient called D.

\[
D = P_{AB} - P_A P_B \quad \text{Equation 1}
\]

Where, the smallest possible value for D,

\[
D_{\text{min}} = \max \{-P_A P_B, -(1-P_A)(1-P_B)\} \quad \text{Equation 2}
\]

And the largest possible value for D,

\[
D_{\text{max}} = \min \{P_A(1-P_B), (1-P_A) P_B\} \quad \text{Equation 3}
\]

D is not easy to understand when the level of LD is being compared between pairs of SNPs with having different allele frequencies. \(D'\) and \(r^2\) are two commonly used measures for LD. Where,
equation 4 (Lewontin, 1964)

\[
D' = \begin{cases} 
  \frac{D}{\max\{P_A(1-P_A), (1-P_A) P_B\}} & \text{When, } D<0 \\
  \frac{D}{\min\{P_A(1-P_B), (1-P_B) P_B\}} & \text{When, } D>0 
\end{cases}
\]

and

\[
r^2 = \frac{D^2}{P_A(1-P_A) P_B(1-P_B)}
\]

equation 5 (Slatkin, 2008)

1.3. Association between variants and complex traits or diseases

Genome-wide association studies (GWAS) were introduced with primary goal to identify SNPs that are significantly associated with complex traits and diseases for example blood pressure, body-mass index, height, schizophrenia and Alzheimer’s disease (AD) (Bush and Moore, 2012). GWAS has enabled hypothesis-free examination of thousands to millions of variations across the entire human genome in contrast to previous candidate gene association studies.

1.3.1. GWAS study design

The most common experimental design for a disease GWAS is a case-control study that compares two groups of individuals (Bush and Moore, 2012). The participants are categorized into those affected by a specific disease (cases) and either controls without the specific disease or unscreened controls. GWAS for quantitative traits include traits such as blood pressure.

A subset of genetic variants is genotyped for all individuals. These variants represent other SNPs in high LD (“tag SNPs”) that are not genotyped (Christoforou et al., 2012). The HapMap (International HapMap Consortium, 2003) and 1000 Genomes project (The 1000 Genomes Project Consortium, 2012, 2015) have provided reference panels that contain information about frequencies of variants and LD patterns in various populations across the world. GWAS uses the
information on LD patterns of SNPs to impute variants that are not genotyped, thus allowing wide capture of variants in the genome. Next, variants are statistically tested for association with complex trait or diseases.

1.3.2. **Outcome and impact of GWAS**

In the past decade, GWAS has revolutionized association studies by providing insight into the genetic landscape of many diseases and traits. 55,857 SNP-trait associations from 3,245 publications are documented in the manually curated online repository called the NHGRI-EBI GWAS catalog (as of 01-Jan-2018) (MacArthur et al., 2017). All studies included in the catalog tested at least 100,000 variations. All the variations documented in the catalog were associated at a p-value threshold < 1.0e-5.

A successful GWAS reveals variations within the locus that are often “tag” SNPs. Given that the SNP being genotyped is a “tag” SNP in high linkage disequilibrium with one or more causal SNPs, there is a challenge to identify the actual candidate causal SNP(s). In addition to that, 90% (Hindorff et al., 2009; Maurano et al., 2012) of GWAS identified trait associated SNPs are in the non-coding regions of the genome and therefore require a biological explanation. This gap in knowledge has partly motivated my research project.

1.4. **Alzheimer’s disease (AD)**

AD is a chronic neurodegenerative disorder and the most common type of dementia that causes problems related to memory, thinking and behavior of an individual. AD is influenced by both genetic and environmental factors. GWAS has identified at least 19 loci that are associated with AD in addition to the well-known *APOE* region (Farrer et al., 1997; Lambert et al., 2013). AD is a disease with no cure, however some treatments are available that can temporarily alleviate symptoms. The incurability of the disease, challenging diagnostics and high impact on the population has motivated my research to investigate GWAS variants associated with AD (Lambert et al., 2013). This section includes general information about AD, its impact on the population followed by a brief discussion on GWAS performed for AD.
1.4.1. Clinical and pathological diagnosis of AD

AD can only be definitely confirmed by post-mortem examination (Ballard et al., 2011). The autopsies of the brains of patients diagnosed in life with AD demonstrate the accumulation of two main proteins. One protein is beta-amyloid (Aβ) that coalesces into plaques in the extracellular regions between neurons of the grey matter. These plaques can eventually interfere with the function of neurons. Many hypotheses have been developed for AD (Du et al., 2018), of which the most commonly known is the amyloid cascade hypothesis. The amyloid cascade hypothesis posits that extracellular Aβ plaques occur first and eventually leads to the intra-neuronal accumulation of the second protein, Tau, which exists in a hyper-phosphorylated state (Ballard et al., 2011). Hyper-phosphorylated Tau leads to the development of protein aggregates within neurons called neurofibrillary tangles and ultimately this results in neuronal death and neurodegeneration of brain circuits (Ballard et al., 2011).

While the gold standard for diagnosis of AD is via autopsy, a clinical diagnosis of AD is needed in order to inform patients and families about what is happening to them and to guide the use of symptomatic therapies. Probable clinical diagnosis of AD is performed by careful examination of the patient’s history, assessment of functional impairments, cognitive testing and by using updated diagnostic criteria (McKhann et al., 2011). Some examples of cognitive tests are the General Practitioner Assessment of Cognition (GPCOG) (Seeher and Brodaty, 2013), Mini-cog (screening for cognitive impairment in older adults) (Fage et al., 2015), Short Information Questionnaire on Cognitive Decline in Elderly (IQCODE) (Jorm, 2004) and Mini-Mental State Examination (MMSE) (Folstein et al., 1975).

1.4.2. Stages of AD

On the basis of clinical symptoms and signs, AD is categorized into three stages that include early, moderate and advanced. These stages are generally defined based on the cutoffs from cognitive tests. MMSE is a 30-point scale cognitive test and most extensively used in clinical or research settings (Folstein et al., 1975). In the early stage of AD, patients suffer from loss of newly learned memories and to some extent language problems i.e., shrinkage of vocabulary. Early stages of AD can also be marked by unnoticed effects on fine motor tasks, for instance,
An increase in speech problems, wandering, crying, illusions and irritability that lead to aggression, are the signs for the moderate stages of the disease. The final or the advanced stage is evident by biographical memory loss, semantic (memories of facts) memory loss and implicit (supporting daily activities e.g., eating) memory loss. At this stage, language is limited to only phrases or words and many people suffer from complete speech loss. Patients at this stage are entirely dependent upon caregivers.

1.4.3. Demographics and genetics of AD

Forty-seven million people in the world live with dementia of which 50-75% have AD (Van Cauwenberghe et al., 2016). Aging is the largest known demographic risk factor for AD. The symptoms of the disease gradually increase over a number of years. The clinical symptoms in the majority of AD cases develop after the age of 65 years (late-onset AD). On the other hand, about 2-10% of patients have early onset of disease (Van Cauwenberghe et al., 2016). Early onset AD is sometimes inherited in an autosomal dominant manner.

1.4.3.1. Causative genes for AD

The search for genetic influence has identified causative mutations in Amyloid Precursor Protein (APP) (Goate et al., 1991), Presenillin-1 (PSEN1) (Broeckhoven et al., 1992; George-Hyslop et al., 1992; Sherrington et al., 1995) and Presenillin-2 (PSEN2) (Sherrington et al., 1996) for early-onset AD (Van Cauwenberghe et al., 2016). The overall population prevalence of AD due to mutation in these genes is rare i.e., less than 1% (Van Cauwenberghe et al., 2016). Studies leading to the discovery of mutations in autosomal dominant family pedigrees have helped in understanding the pathogenesis of the disease. Two pathways breakdown APP: the constitutive (non-amyloidogenic) and amyloidogenic pathways. The amyloidogenic pathway that is associated with AD is enriched in the neurons and the proteolysis of APP is mainly facilitated by β- and γ-secretases resulting in a mixture of Aβ peptides of different lengths. Both the non-amyloidogenic and amyloidogenic pathways result in Aβ_{1-40} that is the common form of the Aβ peptide. The Aβ_{1-42} peptide is less common and derives only from the amyloidogenic pathway and is a predominant constituent of the amyloid plaque. Thirty-nine APP mutations in
93 families favour the breakdown of *APP* into Aβ_{1-42} peptides (O’Brien and Wong, 2011; Van Cauwenberghe et al., 2016).

A study of 1,795 Icelanders revealed that a predicted missense variant (rs63750847) in *APP* protects against AD (Jonsson et al., 2012). Alteration of nucleotide at this SNP results in an alanine to threonine (A673T) substitution. This substitution is more common in all three different population control groups compared to the cases of AD. The analysis with control group including individuals aged 85 year old or greater has an odds ratio (OR) = 5.29 (0.62% versus 0.13%) and P value = 4.78e-7. A673T substitution is located at 2\textsuperscript{nd} position in Aβ peptide and approximately reduced the formation of amyloidogenic peptides by approximately 40% in vitro.

*PSEN1* and *PSEN2* protein products are important components of the γ-secretase complex that help in the cleavage of *APP*. Mutations on *PSEN1* and *PSEN2* may restrict the cleavage of *APP* resulting in increased production of Aβ_{1-42} or decreased production of Aβ_{1-40}. Van Cauwenberghe et al. in 2016 reported that the mutations in *PSEN1* commonly result in the onset of AD at ages ranging from 25 to 65 years. The approximate age of onset of AD can vary from 39 to 83 years for individuals with mutations in *PSEN2* (Van Cauwenberghe et al., 2016).

### 1.4.3.2. Apolipoprotein E (*APOE*): The strongest genetic risk factor associated with late onset of AD

Apolipoprotein E (*APOE*) (Corder et al., 1993; Saunders et al., 1993) is a strong genetic risk factor for the late-onset form of AD. *APOE* codes for apolipoprotein E, that combines with lipids to form lipoproteins. The lipoproteins are responsible for cholesterol and other fat transportation in the bloodstream.

*APOE* has three major isoforms based on the combination of amino acids coded at two different sites. These two sites are at amino acid positions 112 and 158. The isoform with amino acid cysteine on both sites is called ε2 (haplotype: rs7412-T, rs429358-T). A combination of cysteine at residue 112 and arginine at 158 makes the ε3 isoform (haplotype: rs7412-C, rs429358-T). Arginine at both sites makes *APOE* ε4 (haplotype: rs7412-C, rs429358-C).
A clinical and autopsy-based study on Caucasian subjects reported the odds ratio (logistic regression controlling for age and sex) reflecting the association between the \textit{APOE} ε4/ε4 genotype and AD to be 14.9 (95% CI = 10.8-20.6) when compared to \textit{APOE} ε3/ε3 i.e., the reference genotype (Farrer et al., 1997). The odds ratios for association of ε2/ε4 and ε3/ε4 genotype with AD compared to the same reference were 2.6 (95% CI=1.6-4.0) and 3.2 (95% CI=2.8-3.8), respectively. The odds ratio is decreased for subjects with ε2/ε2 and ε2/ε3 genotypes to 0.6 (95% CI= 0.2-2.0) and 0.6 (95% CI=0.5-0.8), respectively. \textit{APOE} allele frequency variation was also investigated in other ethnic groups. The effect of \textit{APOE} ε4/ε4 compared to \textit{APOE} ε3/ε3 genotype is stronger in the Japanese population with an odds ratio of 33.1 (95% CI=13.6-80.5) (Farrer et al., 1997). This variation in results from two analyses using different populations is not significantly different as the confidence intervals are overlapping. However, this observation is likely due to different sample size and factors relating to ethnicity such as allele frequencies.

In the past, multiple candidate-gene association studies have been performed to identify additional loci that are associated with AD. However, researchers were only able to consistently replicate the association of specific variants in the sortilin-related receptor (\textit{SORL1}) gene with AD in independent datasets (Rogaeva et al., 2007; Tosto and Reitz, 2013). Possible reasons for the inconsistencies across various studies were the small sample sizes that lead to limited statistical power and heterogeneous nature of samples (in terms of LD patterns and allele frequency) (Tosto and Reitz, 2013).

1.4.4. GWAS for AD

NHGRI-EBI GWAS catalog (MacArthur et al., 2017) lists 702 SNPs from 28 different studies to be associated with AD and related phenotypes, at the p-value threshold of less than 1e-5 (keyword searched: “Alzheimer’s disease”; accessed on 22 October 2017). All studies listed in the GWAS catalogue analyzed a minimum of 100,000 SNPs in the initial stage. One hundred and thirty-four of 702 associations are significant at genome wide significant threshold (P < 5e-8). More than 50% of 702 SNPs in the GWAS catalogue are intron variants and 26% are intergenic (Table 1.1).
Table 1.1. Type of variants in the NHGRI-EBI GWAS Catalogue for AD and related phenotypes.

<table>
<thead>
<tr>
<th>Functional Classes</th>
<th>Number of Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intron variants</td>
<td>365</td>
</tr>
<tr>
<td>Intergenic variants</td>
<td>183</td>
</tr>
<tr>
<td>Downstream gene variants</td>
<td>41</td>
</tr>
<tr>
<td>Upstream gene variants</td>
<td>35</td>
</tr>
<tr>
<td>Non-coding transcript exon variants</td>
<td>21</td>
</tr>
<tr>
<td>3-prime UTR variants</td>
<td>18</td>
</tr>
<tr>
<td>Missense variants</td>
<td>17</td>
</tr>
<tr>
<td>Regulatory region variants</td>
<td>14</td>
</tr>
<tr>
<td>Synonymous variants</td>
<td>4</td>
</tr>
<tr>
<td>5-prime UTR variants</td>
<td>2</td>
</tr>
<tr>
<td>Splice region variants</td>
<td>2</td>
</tr>
</tbody>
</table>
In the catalogue, an intron variant rs2075650 is the most significant variant associated with AD (P = 1e-295) (Seshadri et al., 2010). rs2075650 is mapped to a nearby gene TOMM40 on chromosome 19. TOMM40 is present in the outer membrane of mitochondria and is responsible for the transportation of precursor protein molecules into the mitochondria. However, this variation is in high LD with APOE (Seshadri et al., 2010) and therefore the signal is probably driven by APOE.

In this thesis, I have focused on the largest GWAS conducted for AD that identified 21 associated loci (Lambert et al., 2013) in addition to APOE. This study was performed using the largest number of sample size compared to all other AD GWAS studies. The Manhattan plot of stage 1 analysis performed on 17,008 cases and 37,254 controls is shown in Figure 1.1 (further details of this study are discussed in Chapter 3).

![Manhattan plot of stage 1 GWAS analysis (Lambert et al., 2013).](image)

The red line represents the genome-wide p-value significant threshold (P < 5e-8). The genes names in black represent previously identified genes by GWAS whereas the ones in red are newly associated genes. The red diamond represents SNPs with the smallest p-values at each locus after stage 1 and stage 2 analyses. The p-values on the y-axis are capped. The most significant p-value (6.70e-536) in APOE is for rs429358 (Lambert et al., 2013-Stage 1 analysis). This figure is directly extracted from “Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer’s disease” (Lambert et al., 2013).

Permission to use image: Springer Nature License number- 4403231123626 [06-Aug-2018]
As mentioned earlier, most of the GWAS hits are non-coding. One way to attempt to understand their function will be to associate them with genes based on physical proximity. For example, AD associated SNPs identified by a number of GWAS, are located close to *APOE, CLU, BIN1, PICALM* and *CR1* genes (Harold et al., 2009; Lambert et al., 2009, 2013; Seshadri et al., 2010). Further exploration of functions of these genes indicated their possible influence on the biological mechanism linked to AD. For instance, similar to *APOE, CLU* (Clusterin), also known as apolipoprotein J, simulates lipid transportation and is hypothesized to influence the production and lyses of Aβ (DeMattos et al., 2004). *BIN1* (Bridging Integrator 1) participates in various cell functions including processes by which proteins and other macromolecules are transported within the cell and into the extracellular spaces (Tan et al., 2013). *BIN1* is associated with modulation of tau pathology involved in AD pathogenesis (Chapuis et al., 2013). Clathrin assembly protein is coded by *PICALM*. Clathrin plays a major role in the formation of coated vesicles that enable transportation within the cell. *PICALM* is involved in modulation of production, transportation and clearance of Aβ peptide (Xu et al., 2015). *CR1* encodes for a cell surface receptor involved in the complement system that is part of the immune system. *CR1* protein has binding sites for C3b complement factor that facilitate transportation of Aβ oligomer and is also involved in clearance of Aβ (Crehan et al., 2012).

In summary, a subset of the genes that are located near AD associated GWAS SNPs, are involved in biological pathways that include transportation of molecules within the cell and into the extracellular spaces, the immune system, lipid metabolism and APP processing. Although, investigating genes tagged by GWAS hits has helped develop hypotheses about the pathology of AD, the direct functionality of these variations or causal loci they tag remains unknown.

### 1.5. Expression Quantitative Trait Loci (eQTLs)

To further understand the importance and functionalities of genetic variations in human DNA, researchers are investigating the association with other biological phenomena like gene expression levels. Expression quantitative trait loci (eQTL) are genomic loci that are associated with variation in gene expression level (mRNA level). Loci associated with variation in expression level of a nearby gene are called cis-eQTLs. The GTEx Project defined cis-eQTLs
based on 1-mega base pairs (Mb) up- and downstream of the TSS. Trans-eQTLs are loci correlated with expression level of a distant gene (located outside of the cis-window).

One of the largest studies that aims to identify eQTLs in different human tissues is the Gene-Tissue Expression (GTEx) project (GTEx Consortium, 2017). GTEx also reported trans-eQTLs that are loci affecting the expression of distant gene (located outside of the cis-window). The latest version of the GTEx project is available at [www.gtexportal.org](http://www.gtexportal.org) (version name: V7) with documentation of the analysis performed on 10,294 samples collected from 48 human tissues and was released in October of 2017.

### 1.5.1. Identification of eQTLs

The identification of eQTLs starts by measuring the level of gene expression in different target tissues of multiple individuals. Microarrays were traditionally used to perform genome-wide profiling of gene expression levels. Advances in technology led to the introduction of RNA-seq that interrogates all RNA rather than a target subset (Majewski and Pastinen, 2011). The association between genotypes and gene expression levels is then tested.

Numerous genes have been documented to have cis-eQTLs by different studies and this number is likely to increase by analyzing larger sample sizes. Identification of trans-eQTLs is relatively difficult because investigating the whole genome for potential association is statistically and computationally challenging.

### 1.5.2. GWAS and eQTLs

Some studies have demonstrated the value of combining eQTLs analysis with GWAS findings to identify disease-associated genes. For example, a study testing over 300,000 SNPs for association with childhood asthma identified significant GWAS SNPs that are linked with expression levels of a nearby gene called *ORMDL3* (orosomucoid like 3) in chromosome 17 (P<10e-22) (Moffatt et al., 2007). Although no direct functional effects of this gene are linked to asthma in humans so far, some studies have indicated involvement of *ORMDL3* in cellular...
processes (like alteration of calcium homeostasis in lung epithelial cells) that may be relevant to the pathogenesis of asthma (Ono et al., 2013).

The GTEx project has identified 152,869 cis-eQTLs for over 19,000 genes from the single tissue analysis. A co-localization analysis performed by GTEx revealed that 52% of GWAS hits (from 21 traits) are co-localized with an eQTL in one or more tissues. Seven percent of these GWAS variants are co-localized with lincRNA and 93% are co-localized with protein coding eGenes, which implies a limited role of lincRNA in pathogenesis of GWAS traits. Further investigation revealed that all identified co-localized eGenes were shared across four tissues on average. Moreover, 26% of the co-localized loci were linked with more than one eGene. These results demonstrate that eQTLs may have important implications in functional characterization of disease-associated variations. Over 90% of the common variants analyzed within GTEx are associated with gene expression level in one or more tissues (P<0.05), and 50% of these variants survive multiple test correction (MTC) for number of tissues. This implies that GWAS significant variants are likely to be cis-eQTLs by chance.

The above mentioned studies and some others (Murphy et al., 2010; Nicolae et al., 2010; Knight et al., 2011; Croteau-Chonka et al., 2015) have helped provide insight into the possible role of some non-coding GWAS hits. However, until now the reason for association of many other variants with gene expression levels or complex traits remain unknown. In addition to the GWAS identified AD associated SNPs; I have also investigated cis-eQTLs in this thesis (Chapter 4).

1.6. Exploring GWAS hits and eQTLs

As stated earlier, many of the GWAS hits and eQTLs are in the non-coding region of the genome. Both, the biological role and reasons for association with traits of such non-coding variations is not completely known. Moreover, identified GWAS variant may not be the actual causal variant associated with the trait (refer to section 1.3.2). Some attempts that have been made to gain greater insight into these findings are listed below.
1.6.1. Fine Mapping

Fine mapping is performed to identify the variations (among the GWAS hits and all SNPs in high LD) that are truly functional and are driving the GWAS signals. To accurately fine map, information on all possible causal variants is required. Computer programs like IMPUTE2 (Howie et al., 2009) along with projects like the 1000 Genomes have enabled imputation of variations that are not genotyped. Additionally, custom genotyping arrays that contain around 200,000 variants, have been designed that provide dense genotyping of previously identified GWAS regions. Some of the examples include Metabochip, designed by Cardio-Metabochip Consortium (Voight et al., 2012) and Immunochip that is designed to fine map GWAS identified regions of 12 immune-mediated phenotypes (Trynka et al., 2011).

1.6.2. Prioritizing variations based on functional annotations

Some GWAS hits and eQTLs are located in the regulatory regions of the human genome. A previous study has reported the enrichment of GWAS hits in the 5-kb promoter regions (13.9% of 465 unique GWAS hits significant at P<5e-8) compared to randomly selected SNPs (6.7% of equal number of SNPs; 100 permutations) from genotyping arrays (Hindorff et al., 2009). Other studies have reported the concentration of GWAS hits in the DNase I hypersensitive sites (DHS) (Maurano et al., 2012) and NFκB transcription factor binding regions (Karczewski et al., 2013). The DNA is compactly packed in a cell to form macromolecules called chromatin. DHS are regions of chromatin that are sensitive to cleavage by DNase I enzyme to expose the DNA for transcription. On the other hand, NFκB family of TFs regulates the transcription of genes that are involved in apoptosis, immunity and inflammation (Oeckinghaus and Ghosh, 2009).

Similarly, further exploration of results from the GTEx project revealed the enrichment of eQTLs in the promoter and enhancer regions from the matched tissue (Wilcoxon rank sum test, P ≤ 9.3 × 10^-4) (GTEx Consortium, 2017). Another study applied bioinformatic approaches using position weight matrix (PWMs; refer to section 1.7.1.2) and statistical analysis to demonstrate importance of eQTL information in predicting TFBS in Arabidopsis thaliana (Von Rohr et al., 2007). The results from this study were validated by comparing sequences from AtcisDB database (Molina and Grotewold, 2005). These studies suggest that
GWAS loci and eQTLs linked to complex traits and gene expression levels tag regulatory regions of the DNA.

Computational tools can be used to incorporate additional biological information with statistical findings to better understand association of variation with different traits (Nishizaki and Boyle, 2017). Some other studies have also shown that combination of functional information (like methylation of DNA) and machine learning statistical algorithms (like random forest and support vector machines) can improve prediction of risk variations that are associated with complex diseases/traits (Gagliano et al., 2014; Kircher et al., 2014; Ritchie et al., 2014).

Online bioinformatic web tools like Ensembl’s variant effect predictor (VEP) (McLaren et al., 2010), HaploReg (Ward and Kellis, 2012) and ANNOVAR (Wang et al., 2010) can be incorporated into methods that combine statistical and biological information to prioritize genetic variation (Nishizaki and Boyle, 2017). Moreover, tools like polyphen2 (Adzhubei et al., 2010) and SIFT (Ng and Henikoff, 2003) use scoring algorithms to predict the deleteriousness of genetic variants and assess the severity of non-synonymous mutations on protein sequences, respectively.

Although the above-mentioned efforts have initiated advancement in characterizing regulatory regions, the ultimate value of GWAS and eQTL studies will come from complete understanding of biological mechanism of these signals. This will enable more accurate understanding of functional pathways that will lead to a better insight into disease biology.

1.7. Motif discovery: Potential method to analyze trait associated variants

Short and recurring patterns of DNA nucleotides that are presumed to have biological function in the human genome are called motifs (D’haeseleer, 2006). Short patterns of sequences are frequently observed at the DNA, RNA and protein levels and play biological roles. Proteins like nucleases and transcription factors (TFs) bind with some specific DNA motifs. This binding is a fundamental aspect of transcriptional regulation. In this thesis, I have explored the DNA regions
near variants to identify words (short sequence of nucleotides with no ambiguity at any position) that are associated with AD and gene expression level.

Computational modeling and experimental validation are essential to identify de-novo motifs and understand their functionalities. One of the major challenges in bioinformatics is to develop computational algorithms that can efficiently analyze biological sequences and identify patterns in DNA regions. Discovering patterns in DNA sequences, analyzing enrichment and developing databases are active research areas. Many bioinformatics tools have been developed for in silico identification of motifs. Computational methods to identify de novo motifs can be less time and money consuming than laboratory based techniques.

1.7.1. Motif analysis tools

Motif analysis web tools generally allow uploading of DNA or RNA or protein sequences that are put through a pipeline of already existing motif analysis algorithms. The users can customize the pipeline implementation strategies and parameters. MEME suite (Multiple EM for Motif Elicitation) is one of the most extensively used bioinformatic tools to analyze biological sequences and has marked a standard in this area of research (Bailey et al., 2009, 2015). Bailey and colleagues have described four major categories of motif analysis tools (listed below) based on the type of input requirements, algorithm performed and output they provide (Bailey et al., 2009). This scheme of classification may not be optimal because there is some overlap between the classes.

1. **Motif Discovery**: Analysis of two sets of user defined genetic sequences to identify de-novo motifs (referred to as positive and negative sequences).

2. **Motif Enrichment**: Analysis of two sets of user defined genetic sequences to report enrichment of user provided or known motifs (query motifs).

3. **Motif Scanning**: Scanning the user provided sequences or from online databases (like the human reference genome database) for matches with user provided motifs (query motifs).

4. **Motif Comparison**: Comparing the query motif with known motifs extracted from different motif databases (like JASPAR).
In this thesis, I have focused on developing a method to discover DNA words in biological sequences that are associated with AD and gene expression level. In this section, some extensively used motif discovery algorithms are discussed and later in section 5.1, a comparison of these tools with my method is presented.

### 1.7.1.1. Discriminative Regular Expression Motif Elicitation (DREME)

DREME is a motif discovery algorithm to identify short core DNA-binding motifs in eukaryotes (Bailey, 2011). DREME is one of the 14 services provided by MEME Suite (Bailey et al., 2009). DREME is typically designed to identify exact matches of motifs of length ranging from 4 to 8 nucleotides in the input sequences.

### 1.7.1.2. Position weight matrix

Position weight matrices (PWM) or position-specific weight matrices (PSWM) or position–specific scoring matrices (PSSM) are used to represent the probabilistic description of short genetic sequences (motifs) (Zhang, 2013). PWMs are commonly used to summarize and predict sequence motifs, and are also an important base component of various motif discovery algorithms like Gibbs sampler and some Monte Carlo methods.

PWMs are matrices created using an aligned set of genetic sequences (nucleotides or amino acid). PWMs have rows that may represent either 4 DNA nucleotides or 20 amino acids and columns that represent each position of aligned homologous sequences. The matrix contains scores (or log likelihood) that are calculated based on frequency of a nucleotide or amino acid at a specific position compared to a random background model. These matrices can be used to match query motif by giving higher weights to sequence positions that are conserved. PWM are often represented as sequence logos (Figure 1.2).
Figure 1.2. Typical illustration of a sequence logo representing PWM.
This sequence logo represents GCM2 transcription factor in human and extracted from JASPER open access database (Khan et al., 2017). GCM2 share core nucleotides with a motif associated with eQTLs in thyroid tissue (refer to chapter 4 for details).

1.7.1.3. Regulatory Sequence Analysis Tool (RSAT)

RSAT (Medina-Rivera et al., 2015) is another bioinformatics web service that combines 52 different tools together to perform investigation of sequence level data. The suite enables motif discovery, analyzing motifs by comparing and clustering to understand convergence and divergence of transcription factor binding sites and detection of regulatory variations. RSAT can also be used to extract DNA sequences from genome assemblies of various species including humans.

Earlier versions of RSAT were limited to analyze the yeast genome and incorporated computer programs that used previously developed algorithms to identify de-novo motifs. Two of such algorithms are Oligo-analysis (van Helden et al., 1998) and dyad-analysis (Helden et al., 2000). Oligo-analysis identifies DNA binding sites for transcription factors from families of genes that are co-regulated, for instance set of genes controlled by same regulators or stimulus. This method is efficient in detecting DNA motifs in organisms with smaller genomes due to the fact that they generally have shorter transcription factors binding sites that are physically located close to one another unlike eukaryotes. Oligo-analysis tends to return a large number of false positives when analyzing random sequences from a vertebrate genome due to different proportion of GC content in different regions of the genome (Thomas-Chollier et al., 2012). On the other hand, Dyad motifs are conserved words separated by regions of fixed size and variable
content and dyad analysis identifies such motifs in the upstream cis-regulatory regions of the yeast genome. The statistical significance of occurrence of a motif is calculated using a binomial expression based on the expected frequency calculated for each monad (word) in the input sequences and expected frequency of all dyads in the non-coding region of the yeast genome.

Peak motif is one of the latest computerized pipeline provided by RSAT to analyze high volume ChIP-seq data (Thomas-Chollier et al., 2012). Peak-motif uses position analysis (for heterogeneous positional distribution) and local-word analysis (for local over-representation of motifs) along with oligo and dyad-analysis, to identify de-novo motifs. It is computationally more efficient compared to programs using oligo and dyad analysis.

1.8. Summary and rationale

GWAS have reported multiple genetic variations that are associated with various human traits. EQTL studies have provided statistical evidence of association between genetic variants and tissue specific gene expression levels. Most of the GWAS variants and eQTLs reside in the non-coding region of the genome. This has led to questions on the biological role of such variants and reasons for their association with traits.

Some other studies (refer to section 1.6.2) have reported the relationship between GWAS hits and eQTLs with regulatory regions in the genome. Furthermore, bioinformatic motif discovery algorithms have shown potential in detecting regulatory biological sequences. The challenge now is to develop new approaches that can use findings from GWAS and eQTLs studies to identify patterns in DNA sequences. Identification of such DNA patterns may support future studies to better understand the reasons for association of variants with complex traits and gene expression levels. Moreover, such patterns may also help identify additional risk variants.
Chapter 2
Thesis Aims and Hypotheses
2.1. Aims and Hypotheses

The primary goal of my thesis is to develop a computational and statistical method that can be used to identify short DNA words near trait-associated variants in the human genome. This is achieved by comparing two groups of sequences that are referred to as sequences of interest (SOI) and control sequences. My method was developed using DNA regions around SNPs associated with Alzheimer’s disease (AD) that were identified by the largest GWAS for AD (Lambert et al., 2013). My method was then applied to another data set to identify DNA words around eQTLs in thyroid tissue (utilizing data from the GTEx project (GTEx Consortium, 2017)). Both the above analyses are independent of one another and the rationales behind selection of these datasets are discussed in Chapter 3 and Chapter 4.

The specific hypotheses I tested are as follows:

1. DNA words have different frequencies near GWAS hits for Alzheimer’s disease (AD), compared to matched regions from the rest of the genome.
2. DNA words have different frequencies near expression quantitative trait loci (eQTL) in thyroid tissue, compared to matched regions in the rest of the genome.
3. Appropriate selection of matching parameters like nucleotide composition, are important to identify significant DNA words linked with trait-associated loci.
Box 1. Important terms used in this thesis.

**Variants of interest (VOI):** Genetic variants that are associated with complex traits, diseases or gene expression levels.

**Sequences of interest (SOI):** DNA sequences near or centered at variants of interest.

**Control variants:** Genetic variants matched with variants of interest for parameters like minor allele frequency, nucleotide composition etc.

**Control sequences:** DNA sequences near or centered at control variants.

**DNA words:** Short sequences of DNA nucleotides with no ambiguity at any position.

**Groups of DNA words:** A pair of DNA word and its reverse complement sequence is called a “group of DNA words”.

2.2. **Structure of the Thesis**

My thesis is composed of two separate analyses. The first analysis was performed on data from the largest GWAS for AD (Lambert et al., 2013). To overcome power related challenges (refer to section 3.4), I performed the second analysis using top eQTLs reported by the GTEx project in thyroid tissue. Thyroid tissue was selected because it has the maximum number of genes with at least one eQTL among the 44-tested tissues in the GTEx project (refer to section 4.1).

The second analysis, using the GTEx data is further composed of three sub-analyses that differ based on the parameters used for the selection of variants of interest (VOI) and control variants (refer to sections 4.3.1, 4.3.2 and 4.3.3). I have then discussed the impact of varying parameters and data source on results from the three sub-analyses performed using GTEx data (refer to section 4.4.1). I have also briefly discussed the potential biological relevance of DNA words identified by my method (refer to section 4.4.2).

The final chapter of the thesis focuses on comparing my method with three known motif discovery methods, followed by limitations of my algorithm, overall conclusions from both analyses and possible directions for future studies.

Computer programs and data files used in this research project can be accessed at:

[https://github.com/kartikaychadha/IMS](https://github.com/kartikaychadha/IMS)
Chapter 3
DNA words near Alzheimer’s disease (AD) loci identified by GWAS
As stated earlier, GWAS has played an important role in understanding the genetics of various diseases and complex traits. The largest meta-study for AD has reported the association of 21 loci with the AD. In this thesis, I have investigated the frequency difference of “exact” DNA words in regions around these AD associated loci (VOI, stands for variants of interest).

3.1. Introduction: The largest meta-GWAS for AD

The largest meta-GWAS for AD was performed by Lambert and colleagues (Lambert et al., 2013). The meta-analysis was conducted on 74,046 individuals and over 7 million SNPs were tested for association with AD. The GWAS study (Lambert et al., 2013) was conducted in two stages. In this section, I have briefly explained the data, methods of analysis and results from this meta-GWAS.

Lambert’s GWAS stage 1 analysis

The data for analysis was collected by the International Genomics of Alzheimer's Project (I-GAP) from four different consortia including the Alzheimer’s Disease Genetic Consortium (ADGC) (10,273 cases and 10,892 controls), the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium (1,315 cases, 12,968 controls), the European Alzheimer’s Disease Initiative (EADI) (2,243 cases and 6,017 controls), and the Genetic and Environmental Risk in Alzheimer’s disease (GERAD) (3,177 cases and 7,277 controls). A total of 17,008 AD cases and 37,154 controls were studied in stage 1. The average age of onset for all cases was 74.7 (SD: 5.9) years and the average age of examination for controls was 68.5 (SD: 11.8) years.

The consortia used different technologies for genotyping in stage 1 including the HumanCNV370-Dup (Illumina), Illumina 370CNV Duo BeadChip system, Affymetrix GeneChip Human Mapping 500K Array, and the Affymetrix 50K Human Gene Focused Panel among others. Different numbers of SNPs were genotyped by each consortium and additional genotypes were imputed using IMPUTE2 (Marchini and Howie, 2010) and MaCH/Mnimac (Li et al., 2010). Each study imputed up to 11.8 million SNPs using the data from the 1000 Genomes project (The 1000 Genomes Project Consortium, 2012) (2010 Interim release based on
sequence data freeze from 4th August 2010 and phased haplotypes from December 2010) of the European super-population.

Each SNP had to be genotyped in at least 40% of the cases and 40% of the controls to be included in the final list of SNPs. 7,055,811 SNPs were included in the final list for stage 1. The association analysis was performed using a logistic regression model (for each study) including covariates for age, sex and principal components to adjust for possible population stratification.

The summary statistics were combined from all datasets. All the data was shared on the I-GAP website and the stage 1 meta-analysis was performed independently by all four consortium using METAL (Willer et al., 2010) and GWAMA (Mägi and Morris, 2010) software.

**Lambert’s GWAS analysis Stage 2**

SNPs surviving the p-value threshold P<5e-3 were included in stage 2 analysis. 19,532 SNPs met the criteria after excluding those in APOE flanking region (Chr19: 45,409,039–45,412,650) (GRCh37/Hg19). Illumina iSelect Microarray technology was used to genotype 14,445 of the SNPs for which assays could be designed. 11,632 SNPs passed the stage 2 quality control procedures and were tested for association in 8,572 AD cases and 11,312 controls from different countries.

The genotyped and imputed data were tested separately for association with AD. For the genotyped data form each country, a logistic regression using an additive genetic model adjusted for sex, age and principal components was applied using PLINK software (Purcell et al., 2007). For the imputed dataset, likelihood score test was performed to test association using SNPTEST (Marchini and Howie, 2010). The results from genotyped and imputed data were combined using an inverse variance approached implemented by METAL (Willer et al., 2010).

In stage 2, 2,562 SNPs reached the nominal level of significance (P <0.05). 116 SNPs with the same risk allele in stage 1 and 2 analyses were found to be significantly associated with AD after Bonferroni correction (P < 4.3e-6) for multiple testing in stage 2.
Table 3.1. Loci associated with AD from stage 1 or stage 2 or combined analysis Lambert et. al. meta-analysis.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr</th>
<th>Position</th>
<th>Major / Minor allele</th>
<th>MAF</th>
<th>Combined Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Odds Ratio (OR)</td>
</tr>
<tr>
<td>rs6656401a</td>
<td>1</td>
<td>207692049</td>
<td>G / A</td>
<td>0.20</td>
<td>1.18</td>
</tr>
<tr>
<td>rs6733839a</td>
<td>2</td>
<td>127892810</td>
<td>C / T</td>
<td>0.41</td>
<td>1.22</td>
</tr>
<tr>
<td>rs35349669a</td>
<td>2</td>
<td>234068476</td>
<td>C / T</td>
<td>0.49</td>
<td>1.08</td>
</tr>
<tr>
<td>rs190982</td>
<td>5</td>
<td>88223420</td>
<td>A / G</td>
<td>0.41</td>
<td>0.93</td>
</tr>
<tr>
<td>rs9271192c</td>
<td>6</td>
<td>32578530</td>
<td>A / C</td>
<td>0.28</td>
<td>1.11</td>
</tr>
<tr>
<td>rs10948363a</td>
<td>6</td>
<td>47487762</td>
<td>A / G</td>
<td>0.27</td>
<td>1.10</td>
</tr>
<tr>
<td>rs2718058</td>
<td>7</td>
<td>37841534</td>
<td>A / G</td>
<td>0.37</td>
<td>0.93</td>
</tr>
<tr>
<td>rs1476679a</td>
<td>7</td>
<td>100004446</td>
<td>T / C</td>
<td>0.29</td>
<td>0.91</td>
</tr>
<tr>
<td>rs11771145a</td>
<td>7</td>
<td>143110762</td>
<td>G / A</td>
<td>0.34</td>
<td>0.90</td>
</tr>
<tr>
<td>rs28834970a</td>
<td>8</td>
<td>27195121</td>
<td>T / C</td>
<td>0.37</td>
<td>1.10</td>
</tr>
<tr>
<td>rs9331896a</td>
<td>8</td>
<td>27467686</td>
<td>T / C</td>
<td>0.38</td>
<td>0.86</td>
</tr>
<tr>
<td>rs10838725a</td>
<td>11</td>
<td>47557871</td>
<td>T / C</td>
<td>0.32</td>
<td>1.08</td>
</tr>
<tr>
<td>rs983392a</td>
<td>11</td>
<td>59923508</td>
<td>A / G</td>
<td>0.40</td>
<td>0.90</td>
</tr>
<tr>
<td>rs10792832a</td>
<td>11</td>
<td>85867875</td>
<td>G / A</td>
<td>0.36</td>
<td>0.87</td>
</tr>
<tr>
<td>rs11218343a</td>
<td>11</td>
<td>121435587</td>
<td>T / C</td>
<td>0.04</td>
<td>0.77</td>
</tr>
<tr>
<td>rs17125944a</td>
<td>14</td>
<td>53400629</td>
<td>T / C</td>
<td>0.09</td>
<td>1.14</td>
</tr>
<tr>
<td>rs10498633a</td>
<td>14</td>
<td>92926952</td>
<td>G / T</td>
<td>0.22</td>
<td>0.91</td>
</tr>
<tr>
<td>rs8093731b</td>
<td>18</td>
<td>29088958</td>
<td>C / T</td>
<td>0.02</td>
<td>0.73</td>
</tr>
<tr>
<td>rs4147929a</td>
<td>19</td>
<td>1063443</td>
<td>G / A</td>
<td>0.19</td>
<td>1.15</td>
</tr>
<tr>
<td>rs3865444b</td>
<td>19</td>
<td>51727962</td>
<td>C / A</td>
<td>0.31</td>
<td>0.94</td>
</tr>
<tr>
<td>rs7274581</td>
<td>20</td>
<td>55018260</td>
<td>T / C</td>
<td>0.08</td>
<td>0.88</td>
</tr>
</tbody>
</table>

*a Previously GWAS defined association loci; b Not significant in Stage 2; c SNP in HLA region and therefore not included in my analysis, but is presented here; All positions are mapped to human genome Build GRCh37/Hg19.
Of these 116 SNPs, 80 survived genome wide significance threshold (P < 5e-8). Nineteen loci (out of 21 listed Table 3.1) from combined results of stage 1 and stage 2 were reported to be associated with AD at the genome wide significance threshold, of which 11 loci were newly reported. Two loci (rs8093731 and rs8093731) reached genome wide significance threshold in stage 1 analysis but were not replicated in the stage 2 or stage 1 and stage 2 combined analysis.

### 3.2. Methods: DNA word discovery

In this section, I explain my DNA word discovery method that was used to investigate the DNA sequences near SNPs associated with AD. The flowchart in Figure 3.2 represents steps involved in pre-processing and selection of the data.

#### 3.2.1. Data selection

##### 3.2.1.1. Variants of interest (VOI)

I selected 20 AD associated loci from results of the largest GWAS study for AD (Table 3.1), as my variants of interest (VOI) (Lambert et al., 2013). The APOE region (Chr19: 45,409,039-45,412,650; GrCh37/hg19) was excluded from my analysis for two reasons:

1. I developed my method using the final results from Lambert et al., 2013. SNPs in APOE locus were not included in stage 2 analysis of this meta-study.
2. The AD associated SNPs in APOE region are already known to have functional effect by changing the amino acid sequence (refer to section 1.4.3.2.)

In addition to APOE, SNPs in the HLA region (rs9271192) were excluded from the list of VOI. The HLA region encodes for major histocompatibility complex (MHC). This region is complex as it is characterized by high gene density, high polymorphism and high LD (Mungall et al., 2003).
3.2.1.2. Control variants

VOI were pairwise matched with an equal number of control variants using the SNPsnap web services (results extracted on 12 May 2016) (Pers et al., 2015). The matching was performed using the 1000 Genomes phase 3 data (The 1000 Genomes Project Consortium, 2015) (European super-population).

SNPsnap finds matches for a user provided query SNP based on four user-defined parameters (Box 2). Firstly, SNPsnap calculates the MAF, gene density based on the user-defined locus definition, number of LD proxies and distances from the nearest gene for all 20 VOI. The locus definition is provided by the user either as physical range in bp or as LD $r^2$ cut-off value. In the case where the LD $r^2$ threshold is selected, the physical distance is equal to the distance between the furthest SNPs from the index SNP with that $r^2$ threshold. Next, the same information is collected for all other SNPs in the genome. Finally, matching is performed within the user-defined range for all four parameters. SNPsnap can return multiple matches for each query SNP. Furthermore, it provides an option to exclude the HLA region (Chr6: 25,000,000-35,000,000; GrCh37/Hg19; non-editable range set by SNPsnap) from matching analysis.

3.2.1.3. SNPs in high Linkage disequilibrium (LD)

A VOI selected may be the causal SNP or a tag SNP in high LD with the causal SNP (refer to section 1.2.2). Therefore, all SNPs in high LD ($r^2 > 0.8$) with every VOI and control variants were also included in the analysis. LDLink (Machiela and Chanock, 2015) is a freely available online tool to extract LD proxies for any query SNPs. The input required for the tool is an rsID. The tool calculates $r^2$ values for all the query-tag SNP pairs based on 1000 Genomes (phase 3) data (European super-population). The output contains a list of pairs of SNPs with calculated D, D’ and $r^2$ values. The pairs are listed in descending order of their $r^2$ value. The earlier version of LDLink that I used in my analysis, did not allow batch submission or a parameter to trim the results based on $r^2$ cut-off value. However, LDLink3.0 (release data: 01 April 2017) now provides command line access that can be used to extract information from the database for multiple queries using a script.
Box 2. SNPsnap: parameters for matching.

1. **Minor Allele Frequency (MAF):** Minor allele frequency (MAF) is the frequency of the least common allele of a SNP in a given population. The deviation was set to ±1% i.e., if the MAF of the VOI is 10%, the matching window for control variant will be between 9% and 11%.

2. **Locus Definition:** The locus defined for a SNP is equal to the physical distance covering all the SNPs that are in high LD with the index SNP ($r^2 > 0.8$).

3. **Gene Density:** Gene density is the number of genes in the defined locus around the SNP. The deviation was set to ±50% i.e., if the gene density of the defined locus of the query SNP (VOI) is 10 genes, the range for the gene density in the control variant locus will be between 5 and 15.

4. **Distance from the nearest gene:** The distance of the SNP from the nearest 5’ start site of a gene. If the SNP is within the gene than the start site of that gene is used to calculate the distance. The deviation was set to ±50% i.e., if the query SNP (VOI) is 100 kb away from the nearest gene, the matching window for control variant will be between 50 and 150 kb.

5. **LD proxies:** The deviation was set to ±50% with $r^2$ cut-off 0.8 i.e., if the query SNP (Variants of interest - VOI) has 100 LD proxies at given $r^2$ threshold, the matching window for control variant will be between 50 and 150 proxies.

6. **Additional options:** The SNPsnap program was set to return 1 match per query SNP and the HLA region was excluded from the matching algorithm.

3.2.1.4. Validating matching parameters

After identifying all the VOI and control variants including the LD proxies, I performed Wilcoxon ranked sum test to compare MAF and number of LD proxies separately. Under the null hypothesis, no significant difference is expected between the distribution of MAF and LD proxies in VOI and control variants.
3.2.1.5. Exploring nucleotide substitutions

I investigated the difference between the distributions of nucleotide substitutions at each SNP in 20 VOI and pair-wise matched control variants (excluding LD proxies). A Fisher’s exact test was performed to test the difference between the distributions of nucleotide substitutions at VOI and control variants.

3.2.1.6. Summary of data selection

Twenty AD associated SNPs (VOI) and 20 pair-wise matched control variants were included in the analysis. All LD proxies ($r^2 > 0.8$) for each of VOI and control variants were also included in the final list of variants analyzed.

3.2.1.7. DNA sequences near VOI and control variants

My method can be used to investigate varying lengths of DNA sequences for different sizes of DNA words. However, I decided to analyze 201bp long nucleotide sequences centered at the VOI and control variants, for all possible 5-mer DNA words. A previous study performed to fine-map recombination rates and hotspots in the human genome investigated motifs ranging from 5bp to 9bp, for enrichment near hotspots, compared to matched regions in the genome (Myers et al., 2005). Another study demonstrated consistent patterns of 5-mer sequences in enriched 9bp long motifs located in the human promoter DNA sequences (FitzGerald et al., 2004).

3.2.2. SNPs with overlapping DNA sequences

Any two variants from the list of VOI and control variants that are physically less than 201 bp apart from one another will result in analyzing the same DNA sequence more than once. Such data will not be completely independent and could inflate type 1 error. Therefore, I developed a strategy to select only the SNPs that are at least more than 201bp apart from each other. I implemented my strategy of selection in R (v3.3.3) programming language, using Rstudio (v1.1.442).
3.2.2.1. Algorithm for selecting SNPs with non-overlapping DNA

The lists of VOI and control variants were investigated separately. My R scripts runs in two loops. The outer loop identifies 20 original VOI (or control variants) and creates 20 clusters of each SNP with its proxies. The inner loop calculates the inter-marker distance between SNPs within every cluster, followed by sorting SNPs into sub-clusters that are less than 201bp apart from each other.

A. Identification of Clusters and Sub-clusters

1. Each cluster includes one of the original VOI (or control variants) and all of its proxies.
2. All the SNPs within each cluster are sorted in ascending order of their chromosomal position and inter-marker distance is calculated.
3. Based on this distance, sub-clusters are identified. Each sub-cluster contains SNPs with inter-marker distance less than 201bp. Each sub-cluster is separated from all others by a distance of more than 201bp.
4. We include all the SNPs in the final analysis list that are more than 201bp apart from each other in each cluster or are not part of any sub-cluster.

B. Selection of SNPs within a Sub-cluster

1. A SNP chosen at random from within a sub-cluster and is included in the final analysis list.
2. All SNPs that are less than 201bp apart from the selected SNPs are excluded from the final list.
3. If there are any SNPs in the sub-cluster that are more than 201bp apart from the selected SNP, the iteration goes back to step 1 and the loop runs for those remaining SNPs.
4. Steps 1-3 are performed for all sub-clusters within a cluster.
3.2.2.2. Implementation of the algorithm to select SNPs with non-overlapping DNA

A step-by-step illustration of the above-explained strategy is given below with help of an example. A chromosome is represented in (Figure 3.1) with three VOI (S1, C2 and S3) and their proxies (S1-a, C2-a, C2-b, S3-a, S3-b, S3-c and S3-d).

![Figure 3.1. Strategy of selecting SNPs with overlapping DNA sequences.](image)

- **Figure 3.1. Strategy of selecting SNPs with overlapping DNA sequences.**
  The figure is a schematic representation of a chromosome (black solid horizontal line) with VOI (S1, C2 and S3) (black solid vertical lines) and associated proxies (S1-a, C2-a, C2-b, S3-a, S3-b, S3-c and S3-d) (black dotted vertical lines). Clusters represent a VOI and all associated tag SNPs. Sub-clusters includes all SNPs less than 201bp apart from each other.

1. Cluster 1, 2 and 3 are identified based on the VOI and tag SNP information.
2. All SNPs within the 3 clusters are arranged in ascending order of their chromosomal position and inter-marker distance is calculated.
3. Sub-Clusters are identified as: S1 and S1-a; S3-a and S3; and S3-b, S3-c and S3-d. SNPs S3-a and S3 are in sub-cluster 2 while S3-b, S3-c and S3-d are in sub-cluster 3 because S3 and S3-b are more than 201bp apart from each other. All SNPs within the sub-cluster 2 or 3 are less than 201bp apart from another.
4. C2-a, C2-b and C2 are directly added to the final SNPs list, as they are more than 201bp apart from one another and in the same cluster.
For sub-cluster 1:

i. S1 is selected at random and is included in the final list.
ii. S1-a is not included in the final list as it is less than 201bp away from S1.
iii. No SNP in this sub-cluster is more than 201bp away from S1. Therefore the iteration terminates.

For sub-cluster 2:

i. S3-a is selected at random and is included in the final list.
ii. S3 is not included in the final list as it less than 201bp away from S3-a.
iii. No SNP in this sub-cluster is more than 201bp away from S3-a. Therefore the iteration terminates.

For sub-cluster 3:

i. S3-b is selected at random and is included in the final list.
ii. S3-c is not included in the final list as it is less than 201bp away from S3-b.
iii. S3-d is the only SNP in sub-cluster that is more than 201bp apart from S3-b.
iv. The loop runs again and includes S3-d in the final list.
v. At the end of second iteration for sub-cluster 3, no SNP is more than 201bp away from S3-d. Therefore the iteration terminates.

The final list contains SNPs S1, C2-a, C2-b, C2, S3-a, S3-b and S3-c.

3.2.3. Calculating the frequency of DNA words

A pair of a DNA word and its reverse complement sequence is called a “group of DNA words” (the term “group” defined in Box 1; section 2.1). A DNA sequence and its reverse complement sequence are not independent of each other when only one strand of the double stranded DNA is included in the analysis. Therefore the counts of a DNA word and its reverse complement are added together. This is called the group count.
I extracted 100bp long nucleotide sequences on either side of VOI and control variants, from the human reference genome (Build GRCh37/hg19) (Lander et al., 2001) using the UCSC table browser tool (Karolchik et al., 2004). These 201bp long nucleotide sequences centered at VOI and control variants are called sequences of interest (SOI) and control sequences, respectively.

Next, I calculated counts for all possible 5-mer groups of DNA words in the SOI and control sequences. The 1024 possible 5-mer sequences were categorized to form 512 groups of DNA words.

### 3.2.4. Statistical analysis

#### 3.2.4.1. Fisher’s exact test

A Fisher’s exact test was used to test the null hypothesis that there is no significant frequency difference between the occurrence of a DNA group of words in SOI and control sequences. The test was performed for all 512 groups. The 2x2 contingency table is represented in (Table 3.2), where x is the number of SOI that contain the words of DNA from a group while n is the total number of SOI. Similarly, m is the total number of control sequences and y is the number of control sequences that contains the same group of DNA words.

<table>
<thead>
<tr>
<th></th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOI</td>
<td>x</td>
<td>n-x</td>
</tr>
<tr>
<td>Control sequences</td>
<td>y</td>
<td>m-y</td>
</tr>
</tbody>
</table>

A 2x2 contingency table used to perform a Fisher’s exact test for all 512 groups of DNA words. **SOI:** sequences of interest. **x:** Number of SOI that contain the DNA words group tested. **n:** Total number of SOI. **y:** Number of control sequences that contain the DNA words group tested. **m:** Total number of control sequences.

#### 3.2.4.2. Trend test

The top nominally significant group of DNA words from Fisher’s exact tests was further investigated. A Cochran Armitage trend test (Sasieni, 1997) was performed to test the
relationship between a variable with two categories (SOI and control sequences) and an ordinal variable (the number of times the word occurred in each DNA sequence). A single 5bp long DNA word cannot occur more than 197 times in a 201 bp long sequence. A subset of a 2x198 (197 possible number occurrence + number of sequences it is absent from) contingency table is shown in Table 3.3. This test was used to identify any trend in the distribution of words per 201bp long sequence.

Table 3.3. Contingency table for Cochran Armitage Trend Test for the group of DNA words with most significant Fisher’s exact p-value (group: ATGAG / CTCAT)

<table>
<thead>
<tr>
<th></th>
<th>Absent</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOI</td>
<td>248</td>
<td>84</td>
<td>19</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Control sequences</td>
<td>145</td>
<td>82</td>
<td>26</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Columns represented the number of sequences that contains the most significant group of DNA words from the Fisher’s exact test. This table is trimmed at column value 4 for clarity. This example group of DNA words occurred maximum of 4 times in a sequence. SOI- Sequences of interest.

3.2.5. Multiple test correction

In statistical testing, rejection of the null hypothesis will occur at a level of 5% by chance when a p-value of less than 0.05 is used to determine significance. A test rejecting a true null hypothesis may be an example of false positive or type 1 error. When performing more than one test, the chances of rejection of true null hypothesis increase if the p-values are not adjusted. Multiple test correction is required to control the type 1 error rate.

Bonferroni (Bonferroni, 1936; Bland and Altman, 1995) proposed a method for multiple test correction that assumes all tests to be independent of each other. According to this method the significant p-value threshold for k number of tests is equal to 0.05/k (9.77e-5 in this case, since k=512). However, DNA words are correlated to one another and therefore all groups of words are not completely independent.
Alternatively, the permutation approach of multiple test correction was used for all 512 groups of DNA words. Permutation allows the p-values to be calculated in relation to the true distribution of the test statistics followed by the data.

Twenty VOI and 20 control variants (excluding LD proxies) were shuffled and randomly tagged as VOI and control variants. All SNPs in high LD with the newly defined index VOI and control variants were extracted. After selecting SNPs that were more than 201bp apart of each other, groups of 5bp long DNA words were counted and the tests for all 512 groups were performed. All the above steps were performed 1000 times with different combinations of VOI and control variants.

The lowest p-value and the mean p-value were recorded for each permutation. An empirical p-value is calculated based on the distribution of test statistics obtained from the permuted data. The empirical p-value threshold is equal to the 5th percentile of the minimum p-values from all permutation runs.

The ratio of the number of times the lowest p-value from each run was less than and equal to the minimum p-value of the original 512 tests and the total number of permutation runs (1000) is equal to the empirical p-value.

Minimum empirical p-value = \[
\frac{\text{Counts of runs with min p-value} \leq \text{original min p-value}}{N+1}
\]

Equation 6

Where, \(N\) = number of permutation runs. One is added to the numerator and denominator of above equation to represent the observed dataset (North et al., 2002).

3.2.6. Deviation from the null hypothesis

Under the null hypothesis, the mean of the p-value distribution is expected to be 0.5. In other words, each bin of size 0.05 in the histogram is expected to contain equal number of groups of DNA words i.e., 26.
3.2.6.1. More than expected significant results

The groups of DNA words with a p-value less than 0.05 are nominally significant. I identified SNPs with largest number of proxies in the data and plotted the distribution of the mean p-values from all permutation runs that contain the identified SNPs (1) together as SOI or control sequences, and (2) in opposite sets of sequences i.e., one as SOI and the other as control sequence or vice versa. Wilcoxon ranked sum test was used to test if the two distributions are significantly different.

![Flowchart](image)

Figure 3.2. Flowchart representing the steps for data pre-processing to selected variants of interest (VOI) and control variants. The data is extracted from the largest GWAS for Alzheimer’s disease (Lambert et al., 2013).
3.3. **Results: DNA words frequency near AD loci**

3.3.1. **SNPs selected for analysis**

The 20 control variants pair-wise matched by SNPsnap with 20 AD associated VOI are shown in Table 3.4.

3.3.1.1. **Extracting LD proxies**

Twenty VOI and matched 20 control variants were submitted to LDLink (Machiela and Chanock, 2015) (LDProxy web services) and 409 VOI and 346 control variants were extracted.

3.3.1.2. **Validating matching parameters**

I performed a Wilcoxon rank sum test (using R v3.3.3) to statistically confirm the matching parameters after including all the SNPs in high LD ($r^2>0.8$). The distribution of MAF and number of LD proxies is not significantly different in VOI compared to control variants with p-values of 0.53 and 0.76, respectively. The distributions for both the parameters are represented in Figure 3.3 and Figure 3.4.
Table 3.4. Control variants pairwise matched with VOI using SNPsnap.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Position</th>
<th>rsID</th>
<th>Proxies</th>
<th>MAF</th>
<th>Chr</th>
<th>Position</th>
<th>rsID</th>
<th>Proxies</th>
<th>MAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>207692049</td>
<td>rs6656401</td>
<td>22/20</td>
<td>0.17</td>
<td>2</td>
<td>217389739</td>
<td>rs9789548</td>
<td>24/21</td>
<td>0.18</td>
</tr>
<tr>
<td>2</td>
<td>127892810</td>
<td>rs6733839</td>
<td>2/2</td>
<td>0.38</td>
<td>3</td>
<td>31175501</td>
<td>rs7625163</td>
<td>1/1</td>
<td>0.38</td>
</tr>
<tr>
<td>2</td>
<td>234068476</td>
<td>rs35349669</td>
<td>56/37</td>
<td>0.46</td>
<td>17</td>
<td>70908616</td>
<td>rs6501579</td>
<td>57/12</td>
<td>0.47</td>
</tr>
<tr>
<td>5</td>
<td>88223420</td>
<td>rs190982</td>
<td>2/2</td>
<td>0.37</td>
<td>8</td>
<td>11385118</td>
<td>rs2736364</td>
<td>1/1</td>
<td>0.38</td>
</tr>
<tr>
<td>6</td>
<td>47487762</td>
<td>rs10948363</td>
<td>98/88</td>
<td>0.25</td>
<td>14</td>
<td>76405298</td>
<td>rs2019207</td>
<td>67/52</td>
<td>0.25</td>
</tr>
<tr>
<td>7</td>
<td>37841534</td>
<td>rs2718058</td>
<td>2/2</td>
<td>0.37</td>
<td>3</td>
<td>29441020</td>
<td>rs7355803</td>
<td>2/2</td>
<td>0.38</td>
</tr>
<tr>
<td>7</td>
<td>100004446</td>
<td>rs1476679</td>
<td>10/10</td>
<td>0.30</td>
<td>14</td>
<td>106952009</td>
<td>rs61996482</td>
<td>13/12</td>
<td>0.30</td>
</tr>
<tr>
<td>7</td>
<td>143110762</td>
<td>rs11771145</td>
<td>1/1</td>
<td>0.36</td>
<td>12</td>
<td>25107985</td>
<td>rs7965671</td>
<td>1/1</td>
<td>0.37</td>
</tr>
<tr>
<td>8</td>
<td>27195121</td>
<td>rs28834970</td>
<td>6/6</td>
<td>0.34</td>
<td>17</td>
<td>10075497</td>
<td>rs2904908</td>
<td>6/5</td>
<td>0.34</td>
</tr>
<tr>
<td>8</td>
<td>27467686</td>
<td>rs9331896</td>
<td>11/8</td>
<td>0.40</td>
<td>10</td>
<td>12244612</td>
<td>rs2895504</td>
<td>10/6</td>
<td>0.39</td>
</tr>
<tr>
<td>11</td>
<td>47557871</td>
<td>rs10838725</td>
<td>43/41</td>
<td>0.28</td>
<td>15</td>
<td>83882783</td>
<td>rs12442572</td>
<td>38/34</td>
<td>0.27</td>
</tr>
<tr>
<td>11</td>
<td>59923508</td>
<td>rs983392</td>
<td>88/73</td>
<td>0.41</td>
<td>10</td>
<td>64497000</td>
<td>rs224056</td>
<td>63/52</td>
<td>0.42</td>
</tr>
<tr>
<td>11</td>
<td>85867875</td>
<td>rs10792832</td>
<td>2/2</td>
<td>0.37</td>
<td>1</td>
<td>12223451</td>
<td>rs590368</td>
<td>2/1</td>
<td>0.38</td>
</tr>
<tr>
<td>11</td>
<td>121435587</td>
<td>rs11218343</td>
<td>3/2</td>
<td>0.04</td>
<td>15</td>
<td>59776919</td>
<td>rs4468541</td>
<td>3/3</td>
<td>0.05</td>
</tr>
<tr>
<td>14</td>
<td>53400629</td>
<td>rs17125944</td>
<td>13/13</td>
<td>0.08</td>
<td>9</td>
<td>12795698</td>
<td>rs1041177</td>
<td>13/12</td>
<td>0.09</td>
</tr>
<tr>
<td>14</td>
<td>92926952</td>
<td>rs10498633</td>
<td>3/3</td>
<td>0.21</td>
<td>4</td>
<td>62223918</td>
<td>rs35872772</td>
<td>2/2</td>
<td>0.21</td>
</tr>
<tr>
<td>18</td>
<td>29088958</td>
<td>rs8093731</td>
<td>20/18</td>
<td>0.02</td>
<td>20</td>
<td>34885607</td>
<td>rs6121201</td>
<td>10/9</td>
<td>0.01</td>
</tr>
<tr>
<td>19</td>
<td>1063443</td>
<td>rs4147929</td>
<td>2/2</td>
<td>0.18</td>
<td>11</td>
<td>12007830</td>
<td>rs56073921</td>
<td>2/2</td>
<td>0.18</td>
</tr>
<tr>
<td>19</td>
<td>51727962</td>
<td>rs3865444</td>
<td>8/7</td>
<td>0.31</td>
<td>16</td>
<td>31440432</td>
<td>rs72789324</td>
<td>10/7</td>
<td>0.32</td>
</tr>
<tr>
<td>20</td>
<td>55018260</td>
<td>rs7274581</td>
<td>18/17</td>
<td>0.08</td>
<td>18</td>
<td>49183286</td>
<td>rs10502925</td>
<td>21/20</td>
<td>0.07</td>
</tr>
</tbody>
</table>

**Chr**- Chromosome number; **Position**- Physical location of the SNP (Human reference genome build GRCh37/Hg19); **Proxies**- Number of all Tag SNPs with LD ($r^2$>0.8) (Before/After: selecting SNPs at least 201bp apart from each other); **MAF**- Minor Allele Frequency (The 1000 Genomes project – Phase 1 (European super-population).
Figure 3.3. Distribution of MAF for VOI and control variants.
A histogram representing distribution of MAF for all VOI (Black) and control variants (Grey), including associated tag SNPs. VOI are represented using black bars while control variants are with grey bars. X-axis represents the range of MAF and Y-axis represents the number of SNPs from VOI or control variants.

Figure 3.4. Distribution of LD proxies for VOI and control variants.
A histogram representing distribution of LD proxies for 20 VOI (Black) and pairwise match equal number of control variants (Grey). X-axis represents the range of number of proxies and Y-axis represents the number of SNPs from VOI or control variants.
3.3.1.3. Exploring nucleotide substitutions

I tested the difference in distribution of nucleotide substitutions between VOI and control variants, using a Fisher’s exact test. The Fisher’s exact test p-value equal to 0.913 showed no significant difference between the two distributions (Figure 3.5).

![Nucleotide substitutions](image)

**Figure 3.5.** Nucleotide substitution at VOI and control variants (Reference allele > alternate allele)

3.3.2. SNPs with no overlapping DNA

Next, I analyzed VOI and control variants using my R script to assure that all SNPs were at least 201 bp apart from each other (refer section 3.2.2). After dropping 13.2% of VOI and 25.1% of control variants, 355 VOI and 259 control variants were included in the final list, respectively. The difference in percentage drop between VOI and control variants is significant with a chi-squared test of p-value 4.12e-5 ($\chi^2 = 16.821$, df=1).

The number of LD proxies per loci in VOI and control variants is shown in Figure 3.6, Figure 3.7 and Table 3.4. In VOI, rs10948363 and rs983392 have the two highest numbers of proxies after selecting SNPs that are at least 201bp apart from each other (98/88 and 88/73 LD proxies)
(before/after: selecting SNPs at least 201bp apart from each other), respectively. rs35349669 has the highest proportion of SNPs dropped (35%).

**Figure 3.6. Number of LD proxies (blue + red bars) and number of SNPs dropped (red bars) across 20 VOI.**

In the control variants, rs2019207, rs224056 and rs6501579 have the highest numbers of LD proxies (67, 63 and 57 LD proxies, respectively). rs6501579 has the highest drop percentage (81%).

**Figure 3.7. Number of LD proxies (blue + red bars) and number of SNPs dropped (red bars) across 20 pairwise matched control variants.**
Figure 3.8 shows the physical location of rs6501579 and variants in high LD ($r^2 > 0.8$) with it (1000 Genomes project; Phase 3; European super-population; using LDLink). Similarly, Figure 3.9 represents the chromosomal position of rs2019207 and its LD proxies. Comparing these two figures clearly show that although rs2019207 has the largest number of proxies among the matched control variants, rs6501579 has the largest number of drops because its proxies are located closer to one another on the chromosome. The x-axes in figures 3.8 and 3.9 are scaled to 1 unit = 500 bp and kept consistent to represent difference in number of proxies.

**Figure 3.8.** Chromosomal position of LD proxies (yellow) for rs6501579 (control variant; blue) in the human genome that has maximum number of drops (Machiela and Chanock, 2015).

**Figure 3.9.** Chromosomal position of LD proxies (yellow) for rs2019207 (Control variant; blue) in the human genome that has maximum number LD proxies (Machiela and Chanock, 2015).
3.3.3. Statistical findings

3.3.3.1. Fisher’s exact test

After counting the occurrences of all groups of DNA words (refer to section 3.2.3), 512 Fisher’s exact tests were performed to identify the groups with significant frequency difference between the SOI and control variants. Forty-two test results (Table 3.5) are significant at a nominal p-value threshold of less than 0.05. The distribution of p-values form 512 Fisher’s exact test is represented in Figure 3.10.

![Figure 3.10](image)

Figure 3.10. P-value distribution for 512 Fisher’s exact tests (a) Histogram (b) QQ-plot

3.3.3.2. Trend Test

The top nominally significant group of DNA words from this analysis was ATGAG/CTCAT with chi-squared nominal p-value equal to 5.6e-04 (Table 3.5). This group of DNA words was further investigated for trend in the data. The p-value from the Cochran Armitage trend test was 3.23e-4 (Table 3.3), which is not significant after Bonferroni multiple testing correction (threshold: P < 9.25e-5). This showed no trend in the distribution of counts for these two DNA words.
Table 3.5. Top 15 significant DNA words at Fisher’s exact test p-value threshold < 0.05

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Word 2</th>
<th>SOI Present</th>
<th>SOI Absent</th>
<th>Control Sequences Present</th>
<th>Control Sequences Absent</th>
<th>OR (CI 95%)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATGAG</td>
<td>CTCAT</td>
<td>107</td>
<td>248</td>
<td>114</td>
<td>145</td>
<td>0.55</td>
<td>(0.39-0.78)</td>
</tr>
<tr>
<td>ACAGT</td>
<td>ACTGT</td>
<td>150</td>
<td>205</td>
<td>74</td>
<td>185</td>
<td>1.83</td>
<td>(1.28-2.62)</td>
</tr>
<tr>
<td>AGCTG</td>
<td>CAGCT</td>
<td>128</td>
<td>227</td>
<td>128</td>
<td>131</td>
<td>0.58</td>
<td>(0.41-0.81)</td>
</tr>
<tr>
<td>TACTG</td>
<td>CAGTA</td>
<td>121</td>
<td>234</td>
<td>58</td>
<td>201</td>
<td>1.79</td>
<td>(1.23-2.63)</td>
</tr>
<tr>
<td>TGTGC</td>
<td>GCACA</td>
<td>109</td>
<td>246</td>
<td>111</td>
<td>148</td>
<td>0.59</td>
<td>(0.42-0.84)</td>
</tr>
<tr>
<td>GACCC</td>
<td>GGGTC</td>
<td>59</td>
<td>296</td>
<td>70</td>
<td>189</td>
<td>0.54</td>
<td>(0.36-0.81)</td>
</tr>
<tr>
<td>GTTGC</td>
<td>GCAAC</td>
<td>80</td>
<td>275</td>
<td>86</td>
<td>173</td>
<td>0.59</td>
<td>(0.40-0.85)</td>
</tr>
<tr>
<td>TGAGG</td>
<td>CCTCA</td>
<td>147</td>
<td>208</td>
<td>137</td>
<td>122</td>
<td>0.63</td>
<td>(0.45-0.88)</td>
</tr>
<tr>
<td>TTAGA</td>
<td>TCTAA</td>
<td>122</td>
<td>233</td>
<td>62</td>
<td>197</td>
<td>1.66</td>
<td>(1.14-2.43)</td>
</tr>
<tr>
<td>TTAGG</td>
<td>CCTAA</td>
<td>81</td>
<td>274</td>
<td>86</td>
<td>173</td>
<td>0.60</td>
<td>(0.41-0.86)</td>
</tr>
<tr>
<td>TGCCG</td>
<td>CGGCA</td>
<td>12</td>
<td>343</td>
<td>22</td>
<td>237</td>
<td>0.38</td>
<td>(0.17-0.81)</td>
</tr>
<tr>
<td>AACGA</td>
<td>TCGTT</td>
<td>28</td>
<td>327</td>
<td>7</td>
<td>252</td>
<td>3.08</td>
<td>(1.29-8.48)</td>
</tr>
<tr>
<td>CTTGG</td>
<td>CCAAG</td>
<td>112</td>
<td>243</td>
<td>109</td>
<td>150</td>
<td>0.63</td>
<td>(0.45-0.90)</td>
</tr>
<tr>
<td>TATAA</td>
<td>TTATA</td>
<td>134</td>
<td>221</td>
<td>72</td>
<td>187</td>
<td>1.57</td>
<td>(1.10-2.26)</td>
</tr>
<tr>
<td>TCTGC</td>
<td>GCAGA</td>
<td>124</td>
<td>231</td>
<td>117</td>
<td>142</td>
<td>0.65</td>
<td>(0.46-0.92)</td>
</tr>
</tbody>
</table>

**Word 1-** 5-mer words; **Word 2-** Reverse complement of word 1; **SOI present-** the proportion of SOI containing the DNA words; **Control sequences present-** the proportion of control sequences containing the DNA words; **SOI absent-** the proportion of SOI not containing the DNA words; **Control sequences absent-** the proportion of control sequence not containing the DNA words; **OR-** Odds Ratio; **CI-** Confidence Interval; **P-value-** Nominal p-value.
3.3.4. Multiple testing correction

None of the groups of DNA words remain significant after Bonferroni multiple test correction (significance threshold = 0.05/512 = 9.67e-5). One thousand permutation runs were performed by randomly distributing the 40 SNPs as new VOI and control variants (refer to section: 3.2.5). Under the null hypothesis we expect the mean of the distribution of mean p-values from permutation runs to be 0.5. The Figure 3.11 represents the distribution of means of p-values from all permutation run and has a mean of 0.46. The empirical p-value calculated for the most nominally significant group of DNA words was 0.577.

![Histogram of mean p-values from 1000 permutation runs.](image)

**Figure 3.11. Distribution of mean p-values from 1000 permutation runs.**

3.3.5. Deviation from the null hypothesis

The p-value distribution for the 512 Fisher’s exact tests is shown in Figure 3.10. The mean p-value of the distribution was 0.48, as opposed to the expectation under the null hypothesis (=0.5).
3.3.5.1. Nominally significant groups of DNA words

Unlike the expectation under the null (i.e., 26 tests), 42 Fisher’s exact tests (Table 3.5) have a p-value less than 0.05. Further investigation revealed that rs10948363 and rs983392 (VOI) are the two SNPs with large number of LD proxies compared to other SNPs in the analysis (rs10948363 has 98 LD proxies and rs983392 has 88 LD proxies). The distributions of the mean p-values from all permutation runs that tagged (1) these two SNPs together as either VOI or control variant (mean P=0.43), and (2) one SNP as VOI and the other as control variant (mean P=0.49), are shown in Figure 3.12(a) and Figure 3.12(b), respectively. The Wilcoxon ranked sum test confirmed a significant difference between these two distributions (P=4.85e-86), with higher values when the two SNPs were both in the same group as opposed different groups.

![Figure 3.12](image.png)

(a) Distribution of mean p-values from permutation runs where rs10948363 and rs983392 (a) were tagged together either as VOI or control variants, (b) were tagged as VOI and control variant, respectively or vice versa.
3.4. Discussion

Although, 42 out of 512 groups of DNA words have nominally significant p-values, none of them survived correction for multiple testing. In other words, none of the DNA words have significantly different frequency in 201bp long DNA regions centered at AD associated GWAS loci compared to matched control sequences.

Further investigation that is presented in section 3.3.5.1, revealed multiple reasons for deviation of the results from expectation under the null hypothesis. The two SNPs with the largest number of proxies were both VOI were rs10948363 and rs983392. rs10948363 has 98 proxies (LD $r^2>0.8$) of which 88 were selected to be analyzed after filtering the SNPs that were less than 201 bp apart from each other. Similarly, rs983392 has 88 proxies (LD $r^2>0.8$) and 73 were included in the analysis after applying the minimum inter-marker distance filter. Under the null hypothesis the distribution of the mean p-values from all permutation runs is expected to be normal with a peak at 0.5. However, Figure 3.12 (a) clearly shows a left skewed distribution. On the other hand, the distribution of mean p-values (Figure 3.12(b)) from permutation runs where these two SNPs were tagged as VOI and control variants, respectively (or vice versa), is closer to our expectation under the null hypothesis. This difference in distribution was also statistically confirmed by performing a Wilcoxon ranked sum test (refer to section 3.3.5.1). This implies that the SNPs with high count of proxies play a role in driving results of the analysis.

The results from this analysis suggested two possible conclusions: (1) The data is underpowered for some groups of DNA words with small effect size; (2) The underlying hypothesis is incorrect. In summary, the overall analysis results identified no DNA words that show relationship with AD associated loci in the genome. Important highlights from this analysis are:

• VOI or control variants with high number of proxies are associated with higher number of significant DNA words than expected under the null hypothesis (refer to section 3.3.5.1).

• No evidence of trend was found in the distribution of most significant DNA words around AD associated SNPs compared to the rest of the human genome.
Chapter 4
DNA words near thyroid eQTLs identified by the Gene-Tissue Expression (GTEx) Project
I implemented my method of DNA word discovery on the data from the GTEx project. I selected eQTLs in thyroid tissue as variants of interest (VOI) because it has 10,610 eGenes, the largest number reported among all other tissues.

4.1. Introduction: The GTEx Project

As mentioned earlier, the GTEx project is one of the largest studies that aimed to identify eQTLs in various human tissues. The unknown functionality of GWAS hits was one of the motivations for the GTEx project. My thesis focuses on results from single tissue cis-eQTL analysis documented in v6p version of the GTEx project that was performed on 7,051 samples collected from 455 human donors across 44 (42 distinct) tissues (GTEx Consortium, 2017). Hereafter all details refer to the cis-eQTL analysis, v6p release (the latest version is V7; refer to section 1.5). In this section, I will be discussing data, analysis methods and results from the GTEx project.

4.1.1. GTEx data collection and processing

The majority of the samples for the GTEx project (v6p) were collected from European donors (83.7%), with the next largest population being African Americans (15.1%). The autopsy samples were categorized based on the cause of death including traumatic injury, cerebrovascular, heart disease, liver or renal or respiratory and neurological. A total of 455 donors were recruited for the analysis of which 296 (65.1%) were male. After quality control steps, genotyping and RNA-sequencing was performed for 449 donors.

Thirty-one solid-organ tissues, 10 brain sub-regions (including duplicates of cortex and cerebellum), whole blood, and two cell lines that have at least 70 donors were included in the GTEx analysis. The gene expression was quantified from samples using RNA-seq technology that sequenced reads to a median depth of 78 million reads per tissue sample. The RNA-seq reads were aligned to human reference genome GRCh37/hg19 using Tophat v1.4.1 (Trapnell et al., 2009).
Whole blood samples were genotyped at the GTEx Laboratory Data Analysis and Coordination Center (LDACC) at the Broad Institute, using the Illumina 2.5M (2,378,075) and 5M (4,276,680) microarrays. Imputation was performed using the 1000 Genomes project data (phase 1 v3, multi-ethnic panel; 1000 Genomes Project Consortium, 2010) that generated 12.5 million variants for analysis of which 11,552,519 were autosomal and the remaining were located on X chromosome. Imputation was performed separately for the autosomal and sex chromosomes using IMPUTE2 (Howie et al., 2009). Annotation for physical position was based on human reference genome GRCh37/hg19 (International Human Genome Sequencing Consortium, 2001; Church et al., 2011). All raw data was made available on dbGaP (Mailman et al., 2007; Tryka et al., 2014) (accession phs000424.v6.p1).

4.1.2. GTEx methods for cis-eQTL analysis

The GTEx project defined 1 Mega base pairs (Mb) up- and downstream of the TSS as the cis-window for analysis. Genes with at least one eQTL are known as eGenes. All autosomal long intergenic noncoding RNA (lincRNA) and protein-coding genes were tested for association. The associated variants are called eVariants.

A linear regression model controlling for ancestry, sex, genotyping platform and latent factors (in expression data representing technical variables) was applied by GTEx (GTEx Consortium, 2017) to test the association between genotype and gene expression level. The alternate hypothesis was tested for deviation of model’s slope from 0, between the genotype and expression. FastQTL (Ongen et al., 2016) implemented permutation was used to calculate the empirical p-values to determine the most significant variant per gene. Next, the q-values (Storey, 2002) using false discovery rate (FDR) threshold of ≤0.05 were calculated to correct empirical p-values across all genes. The variants with nominal p-values less than the given gene-level thresholds were identified as the significant gene-variant pairs for the eGene (GTEx Consortium, 2017).
4.1.3. Results of the GTEx project

Single tissue analysis identified 152,869 cis-eQTLs for 19,725 eGenes. These eGenes represent 50.3% of known lincRNA and 86.1% protein-coding genes. The GTEx consortium reports the largest number of cis-eGenes in tibial nerve (8,087) (GTEx Consortium, 2017), however the earlier release reported the largest number of eGenes in thyroid (10,610) (Aguet et al., 2016). This discrepancy in results is likely due to the imputation that was performed using two different versions of the 1000 Genomes project. Cis-eQTLs show a distinct bimodal pattern of being shared across different tissues i.e. either are shared across all 44 tissues or are specific to a small subset of tissues. The cis-eQTLs discovered in tissues with large sample sizes were more likely to be tissue-specific as compared to the tissues with fewer samples.

4.2. Methods: DNA word discovery

As the result of an increase in data from 20 AD associated GWAS SNPs to 10,910 thyroid eQTLs, some steps of the method described in the previous chapter (section 3.2) were altered. I analyzed 5bp long nucleotide sequences centered at the VOI and control variants, as opposed to 201bp long sequences in the analysis of AD GWAS data. The change in length avoids possible correlation between adjacent DNA words in 201bp sequence that may inflate type 1 error. In this chapter, I have described a primary analysis along with secondary and tertiary analyses. The three analyses differ from each other based on (1) matching parameters; (2) inclusion or exclusion of LD proxies, (3) version of the data used from the 1000 Genomes project, and (4) steps performed for multiple test correction. In the upcoming sections of this chapter, I have discussed the common methodology used for all three analyses and the differences between them.

4.2.1. Pre-processing of GTEx files to select VOI

GTEx reported significant SNP-gene pairs based on the unique gene ids. The SNP-gene pairs with the minimum nominal p-value for each gene id were selected as VOI. In a case where more than one SNP had the same minimum p-value for the same gene, one SNP was randomly selected. The SNPs with minimum p-value for more than one gene were counted only once in
the final list. Only bi-allelic SNPs were used and the MAF was calculated based on the 1000 Genomes project European super-population data (phase 1 data was used for primary and secondary analysis and phase 3 data was used for tertiary analysis). The pre-calculated MAF for all bi-allelic SNPs, categorized by two versions of the 1000 Genomes project was downloaded from SNiPA (Arnold et al., 2015).

4.2.2. Primary analysis: Control variants matched for MAF and nucleotide frequency

In this section, I explain my DNA word discovery method used to investigate 5-mer DNA sequences centered at bi-allelic thyroid eQTLs. The flowchart in Figure 4.1 represents steps involved in pre-processing and selection of the data.

4.2.2.1. Selection of control variants

VOI were matched to control variants in a pool of non-eQTLs. All SNP-gene pairs that were tested for thyroid tissue and have a p-value more than 0.05 were identified. The unique SNPs excluding the VOI were included in the final list of non-eQTLs. The 1000 Genomes phase 1 European super-population (from SNiPA) data was used to obtain the MAF for all bi-allelic SNPs.

The control variants were then matched for MAF and nucleotide composition within the list of all non-eQTLs. Nucleotide composition is the proportion of each nucleotide in a given DNA sequence. The list of non-eQTLs was searched for SNPs with MAF within the range of ±1% of the VOI’s MAF. The filtered list was then searched for a SNP with exactly the same nucleotide composition (in 5-mer DNA centered at the SNP) as the VOI. In the case when more than one SNP met the above two parameters, one of the SNPs was randomly selected.

4.2.2.2. Validating matching parameters

The distributions of MAF in VOI and control variants were validated for difference using a Wilcoxon ranked sum test. Cochran Armitage trend tests were used for each nucleotide to
investigate potential differences between the distribution of nucleotide composition for SOI and control sequences.

4.2.2.3. Exploring nucleotide substitutions

The difference between the distributions of nucleotide substitutions at VOI and pair-wise matched control variants were investigated using a chi-squared test. Chi-squared tests for all 12 possible nucleotide substitutions were performed to identify the substitution(s) that most influence the global chi-squared test.

4.2.2.4. SNPs with overlapping DNA sequences

The length of the DNA sequence analyzed was 5-bp long centered at the index SNP. Therefore all SNP pairs those were at least more than 5bp apart from each other were identified using my strategy (explained in section 3.2.2) and included in the analysis.

4.2.2.5. Extracting DNA sequences

The 5-mer polynucleotide sequences were extracted from the human reference genome (Build GRCh37/hg19). I used RSAT (Medina-Rivera et al., 2015) to extract the DNA sequences centered at the VOI and control variants.

4.2.2.6. Statistical analysis

The proportion of SOI and control sequences that exactly matched the DNA words for all groups was recorded. Next, 512 Fisher’s exact tests were performed to test the null hypothesis that there is no significant difference in the frequency of a group (DNA words) between the SOI and control sequences.

4.2.2.7. Multiple testing correction

As previously discussed in section 3.2.5, rejection of the null hypothesis occurs at a level of 5% by chance when a p-value of 0.05 is used to determine significance. Similar to the analysis
described in chapter 3, one thousand permutation runs were performed to calculate the empirical p-value threshold and control multiple testing.

In the GTEx data set, for each permutation run, all 5-mer DNA sequences were randomly tagged into equal number of SOI and control sequences. The matches for all 512 possible groups of 5-mer DNA words were counted in the new SOI and control sequences. This was followed by Fisher’s exact test for each group of DNA word. The lowest p-value and the mean p-value were recorded for each permutation. The empirical p-value threshold was calculated that is equal to the 5th percentile of the minimum p-values from all permutation runs.

### 4.2.2.8. Deviation from the null hypothesis

Under the null hypothesis, the mean of the p-value distribution is expected to be 0.5. A deviation from the null hypothesis was further investigated by plotting the distribution for counts of groups of DNA words in tests with p-value equal to 1 and less than 1, respectively. The distribution of the two-plotted histogram was tested using a Wilcoxon ranked sum test for significant difference.
4.2.3. Secondary analysis: Control variants matched for MAF only

The secondary analysis was performed for same set of VOI as in primary analysis (refer to section 4.2.1), by matching the control variants only for MAF. Unlike the primary analysis, matching was not done for nucleotide composition in 5-mer sequences centered at VOI or control variants. The difference between the distribution of composition of each nucleotide in SOI and control sequences was tested using Cochran Armitage trend tests. This was the only difference between the methods of discovering DNA words in primary and secondary analysis.

4.2.4. Tertiary analysis: Including LD proxies

The parts of method used in the tertiary analysis that differ from the primary and secondary analyses are listed below:

1. The bi-allelic VOI and control variants were identified based on the 1000 Genomes phase 3 data (European super-population), as opposed to 1000 Genomes phase 1 data used in the primary and secondary analysis.

2. All the LD proxies (r²>0.8) for selected VOI and matched control variants were included in the analysis (similar to Chapter 3).
4.3. Results

GTEx tested 9,151,913 SNPs for association with the expression of 27,649 genes in thyroid tissue. About 1.6 million SNP-gene pairs in thyroid tissue (0.08% of tested SNP-gene pairs and 9.8% of SNPs tested; Figure 4.1) were significant based on the GTEx significance threshold. The SNP-gene pairs with the most significant p-value for all 10,610 genes were selected. Three hundred and twelve of these 10,610 eQTLs were shared by more than one gene. Therefore, I identified 10,298 unique SNPs that were selected as VOI.

4.3.1. Primary analysis: Control variants matched for MAF and nucleotide frequency

The 10,298 unique VOI were investigated for all bi-allelic SNPs using the 1000 Genomes phase 1 data European super-population, obtained from SNiPA. After applying this filter, 9,103 bi-allelic SNPs were listed as VOI. These VOI were then used to identify pair-wise matched control variants.

Out of the total SNP-gene pairs tested in thyroid tissue (179,693,530), 164,916,116 (91.7%) SNP-genes pairs (9,119,683 unique SNPs) have a nominal p-value more than 0.05 (Figure 4.1). Eighty seven percent of these unique SNPs, i.e., 7,942,756 SNPs are bi-allelic based on the data from 1000 Genomes phase 1– European super-population, hosted by SNiPA. After removing the duplicate entries of eQTLs that were shared by more than one gene, 7,933,653 SNPs were included in the final list of non-eQTLs. All 9,103 VOI were pairwise matched for control variants in the pool of non-eQTLs.
Figure 4.1. Flowchart representing the steps for data pre-processing to selected variants of interest (VOI) and control variants for primary and secondary analysis. The data is extracted from the GTEx project for thyroid tissue.
4.3.1.1. Validating matching parameters

The distributions of MAF of all 9,103 VOI and equal number of control variants are represented in Figure 4.2. The Wilcoxon ranked sum test confirmed no significant difference between the distribution of MAF in VOI and control variants ($P = 0.97$).

![Figure 4.2](image)

**Figure 4.2.** Minor Allele Frequency (MAF) distribution for variants of Interest (VOI) (Black) and control variants (Grey).

Next, Cochran Armitage trend tests were used to confirm that there was no significant difference in distribution of nucleotide composition (i.e., proportion of a nucleotide in a sequence) between SOI and control sequences for all four nucleotides (A, T, G and C). P-value equal to 1 (df=5) for each of the four tests (Table 4.1) confirmed no significant difference between the distributions of nucleotides.
Table 4.1. 2x6 contingency tables to test difference in distribution of nucleotide composition between SOI and control sequences using Cochran Armitage trend test.

<table>
<thead>
<tr>
<th>Nucleotide A count</th>
<th>Absent</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOI</td>
<td>2581</td>
<td>3509</td>
<td>2007</td>
<td>764</td>
<td>200</td>
<td>42</td>
</tr>
<tr>
<td>Control sequences</td>
<td>2581</td>
<td>3509</td>
<td>2007</td>
<td>764</td>
<td>200</td>
<td>42</td>
</tr>
</tbody>
</table>

P-value =1 (df=5)

<table>
<thead>
<tr>
<th>Nucleotide T count</th>
<th>Absent</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOI</td>
<td>2589</td>
<td>3411</td>
<td>2119</td>
<td>762</td>
<td>197</td>
<td>25</td>
</tr>
<tr>
<td>Control sequences</td>
<td>2589</td>
<td>3411</td>
<td>2119</td>
<td>762</td>
<td>197</td>
<td>25</td>
</tr>
</tbody>
</table>

P-value =1 (df=5)

<table>
<thead>
<tr>
<th>Nucleotide G count</th>
<th>Absent</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOI</td>
<td>2285</td>
<td>3328</td>
<td>2258</td>
<td>1000</td>
<td>222</td>
<td>9</td>
</tr>
<tr>
<td>Control sequences</td>
<td>2285</td>
<td>3328</td>
<td>2258</td>
<td>1000</td>
<td>222</td>
<td>9</td>
</tr>
</tbody>
</table>

P-value =1 (df=5)

<table>
<thead>
<tr>
<th>Nucleotide C count</th>
<th>Absent</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOI</td>
<td>2216</td>
<td>3268</td>
<td>2355</td>
<td>990</td>
<td>256</td>
<td>18</td>
</tr>
<tr>
<td>Control sequences</td>
<td>2216</td>
<td>3268</td>
<td>2355</td>
<td>990</td>
<td>256</td>
<td>18</td>
</tr>
</tbody>
</table>

P-value =1 (df=5)

4.3.1.2. Exploring nucleotide substitutions

The distribution of nucleotide substitution at each VOI and control variants, is represented in Figure 4.3. Chi-squared test p-value equal to 3.1e-4 (df=11, \(\chi^2\): 33.43) showed that the distribution of nucleotide substitution in VOI and control variants is significantly different.

The substitution with the most significant effect on the global chi-squared test result is A>G with a p-value equal to 2e-4 (Table 4.2). It is the only substitution that remains significant after Bonferroni multiple test correction (p-value cut-off = 0.05/12= 4e-03). The p-value from the global chi-squared test is equal to 2.5e-2 (df=10; \(\chi^2\): 20.45) without A>G substitution flavour, which is still nominally significant. A non-significant chi-squared test result (P=0.06; df=9; \(\chi^2\):
16.31) was obtained by performing the test without A>G and G>A (second most significant substitution) substitution flavours.

![Figure 4.3. Nucleotide substitution at VOI and Control variants (Reference allele > alternate allele)](image)

Table 4.2. $\chi^2$ test results for 12 nucleotide substitutions in VOI and control variants.

<table>
<thead>
<tr>
<th>Substitution type</th>
<th>VOI count</th>
<th>Control variants count</th>
<th>VOI Present (%)</th>
<th>Control variants Present (%)</th>
<th>$\chi^2$ Statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A&gt;C</td>
<td>347</td>
<td>296</td>
<td>4</td>
<td>3</td>
<td>4.03</td>
<td>4.47E-02</td>
</tr>
<tr>
<td>A&gt;G</td>
<td>1415</td>
<td>1237</td>
<td>16</td>
<td>14</td>
<td>13.83</td>
<td>2.00E-04</td>
</tr>
<tr>
<td>A&gt;T</td>
<td>246</td>
<td>280</td>
<td>3</td>
<td>3</td>
<td>2.13</td>
<td>0.144</td>
</tr>
<tr>
<td>C&gt;A</td>
<td>378</td>
<td>418</td>
<td>4</td>
<td>5</td>
<td>1.19</td>
<td>0.276</td>
</tr>
<tr>
<td>C&gt;G</td>
<td>408</td>
<td>440</td>
<td>4</td>
<td>5</td>
<td>1.13</td>
<td>0.289</td>
</tr>
<tr>
<td>C&gt;T</td>
<td>1770</td>
<td>1828</td>
<td>19</td>
<td>20</td>
<td>5.76</td>
<td>5.98E-03</td>
</tr>
<tr>
<td>G&gt;A</td>
<td>1767</td>
<td>1917</td>
<td>19</td>
<td>21</td>
<td>0.82</td>
<td>0.367</td>
</tr>
<tr>
<td>G&gt;C</td>
<td>449</td>
<td>422</td>
<td>5</td>
<td>5</td>
<td>1.21</td>
<td>0.271</td>
</tr>
<tr>
<td>G&gt;T</td>
<td>373</td>
<td>404</td>
<td>4</td>
<td>4</td>
<td>0.16</td>
<td>0.692</td>
</tr>
<tr>
<td>T&gt;A</td>
<td>260</td>
<td>270</td>
<td>3</td>
<td>3</td>
<td>1.59</td>
<td>0.208</td>
</tr>
<tr>
<td>T&gt;C</td>
<td>1360</td>
<td>1299</td>
<td>15</td>
<td>14</td>
<td>2.28</td>
<td>0.131</td>
</tr>
</tbody>
</table>

A- Adenine; T- Thymine; G- Guanine; C- Cytosine; VOI- Variants of interest.
4.3.1.3.  Statistical findings

All VOI and control variants were at least more than 5bp apart from each other and therefore all 18,206 VOI and control variants were included in the analysis. RSAT was used to extract 5-mer long sequences of interest (SOI) and control sequences centered at VOI and control variants, respectively. Three groups of DNA words (Table 4.3) had zero matches in both SOI and control sequences. Therefore the total number of Fisher’s exact tests performed was 509.

Table 4.3. DNA words absent in both SOI and control sequences.

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Word 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATCG</td>
<td>CGATA</td>
</tr>
<tr>
<td>GTACG</td>
<td>CGTAC</td>
</tr>
<tr>
<td>GTTCG</td>
<td>CGAAC</td>
</tr>
</tbody>
</table>

The p-value distribution for 509 tests is shown in Figure 4.4 with mean p-value 0.62 (SD=0.31), as opposed to expectation under the null i.e., 0.5. Fifteen groups of DNA words (Table 4.4) have nominally significant p-values at threshold less than 0.05. Under the null hypothesis and assuming that all tested groups of DNA words are independent, 26 groups are expected to have a p-value less than 0.05.

Figure 4.4. P-value distribution for 509 Fisher’s exact tests.
Table 4.4. Groups of DNA words at Fisher’s exact test p-value < 0.05.

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Word 2</th>
<th>SOI Present</th>
<th>SOI Absent</th>
<th>Control Sequence Present</th>
<th>Control Sequence Absent</th>
<th>OR (CI 95%)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCTGC</td>
<td>GCAGA</td>
<td>35</td>
<td>9068</td>
<td>12</td>
<td>9091</td>
<td>2.92</td>
<td>1.07E-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.48-6.19)</td>
<td></td>
</tr>
<tr>
<td>GACGC</td>
<td>GCGTC</td>
<td>37</td>
<td>9066</td>
<td>14</td>
<td>9089</td>
<td>2.65</td>
<td>1.74E-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.40-5.31)</td>
<td></td>
</tr>
<tr>
<td>AGGCT</td>
<td>AGCCT</td>
<td>25</td>
<td>9078</td>
<td>51</td>
<td>9052</td>
<td>0.49</td>
<td>3.76E-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.29-0.80)</td>
<td></td>
</tr>
<tr>
<td>AGTCC</td>
<td>GGACT</td>
<td>21</td>
<td>9082</td>
<td>6</td>
<td>9097</td>
<td>3.51</td>
<td>5.89E-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.37-10.62)</td>
<td></td>
</tr>
<tr>
<td>AGTTC</td>
<td>GAACT</td>
<td>8</td>
<td>9095</td>
<td>23</td>
<td>9080</td>
<td>0.35</td>
<td>1.06E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.13-0.80)</td>
<td></td>
</tr>
<tr>
<td>GAGCC</td>
<td>GGCTC</td>
<td>14</td>
<td>9089</td>
<td>31</td>
<td>9072</td>
<td>0.45</td>
<td>1.60E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.22-0.87)</td>
<td></td>
</tr>
<tr>
<td>TTTAG</td>
<td>CTAAA</td>
<td>25</td>
<td>9078</td>
<td>10</td>
<td>9093</td>
<td>2.50</td>
<td>1.66E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.16-5.84)</td>
<td></td>
</tr>
<tr>
<td>AGGAA</td>
<td>TTCCT</td>
<td>40</td>
<td>9063</td>
<td>21</td>
<td>9082</td>
<td>1.91</td>
<td>2.02E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.10-3.41)</td>
<td></td>
</tr>
<tr>
<td>ATAGC</td>
<td>GCTAT</td>
<td>3</td>
<td>9100</td>
<td>13</td>
<td>9090</td>
<td>0.23</td>
<td>2.12E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.04-0.84)</td>
<td></td>
</tr>
<tr>
<td>TAGAC</td>
<td>GTCTA</td>
<td>3</td>
<td>9100</td>
<td>13</td>
<td>9090</td>
<td>0.23</td>
<td>2.12E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.04-0.84)</td>
<td></td>
</tr>
<tr>
<td>GAATC</td>
<td>GATTC</td>
<td>13</td>
<td>9090</td>
<td>3</td>
<td>9100</td>
<td>4.34</td>
<td>2.12E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.19-23.74)</td>
<td></td>
</tr>
<tr>
<td>CTCCG</td>
<td>CGGAG</td>
<td>9</td>
<td>9094</td>
<td>1</td>
<td>9102</td>
<td>9.01</td>
<td>2.15E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.25-394.17)</td>
<td></td>
</tr>
<tr>
<td>AACCC</td>
<td>GGGTT</td>
<td>30</td>
<td>9073</td>
<td>15</td>
<td>9088</td>
<td>2.00</td>
<td>3.55E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.04-4.01)</td>
<td></td>
</tr>
<tr>
<td>AGGAG</td>
<td>CTCCT</td>
<td>24</td>
<td>9079</td>
<td>41</td>
<td>9062</td>
<td>0.58</td>
<td>4.60E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.34-0.99)</td>
<td></td>
</tr>
<tr>
<td>GAGGG</td>
<td>CCCTC</td>
<td>24</td>
<td>9079</td>
<td>41</td>
<td>9062</td>
<td>0.58</td>
<td>4.60E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.34-0.99)</td>
<td></td>
</tr>
</tbody>
</table>

**Word 1** - 5-mer words; **Word 2** - Reverse complement of word 1; **SOI present** - the proportion of SOI containing the DNA words; **Control sequences present** - the proportion of control sequences containing the DNA words; **SOI absent** - the proportion of SOI not containing the DNA words; **Control sequences absent** - the proportion of control sequence not containing the DNA words; **OR** - Odds Ratio; **CI** - Confidence Interval; **P-value** - Nominal p-value.
4.3.1.4. Multiple testing correction

None of the groups of DNA words remain significant after Bonferroni multiple test correction (significant threshold = 0.05/509= 9.82e-5). Figure 4.5 (a) represents the distribution of means of p-values from all permutation runs and has a mean of 0.61 (SD=0.01). Under the null distribution the mean is expected to be 0.5. The mean p-value of 509 originally tested groups of DNA words is 0.62 (SD=0.31) (Figure 4.4). The 5th percentile of the minimum p-value distribution is 3.1e-4 i.e., the empirical p-value threshold. None of the group of DNA words remained significant at this empirical p-value threshold. The distribution of minimum p-values from 1000 permutation runs is shown in Figure 4.5(b).

![Distribution of mean p-values and minimum p-values](image)

Figure 4.5. Distribution of (a) mean p-values (b) Minimum p-values, from 1000 permutation runs.

4.3.1.5. Groups of DNA words with p-value equal to 1

One hundred and twenty-one tests out of 509 have a p-value equal to 1. I used Wilcoxon ranked sum test to compare the counts of SOI and control sequences that matched the DNA words in groups with p-value equal to 1 and less than 1. A highly significant p-value of 5.73e-9 confirmed the difference in distribution (Figure 4.6). The mean of the sum of occurrence of DNA words as groups in SOI and control sequences together is 26.27 and median is equal to 23 in those with p-value is equal to 1, while in those with p-value less than 1 had mean and median of 38.43 and 35, respectively. This implies that groups of DNA words with p-value equal to 1 have lower counts on average.
Figure 4.6. Distribution of proportion of groups of DNA words that matched SOI and control variants where (a) p-values = 1 (b) p-values < 1.
4.3.2. Secondary analysis: Control variants matched for MAF only

The data used for this analysis was the same as for the primary analysis (refer to section 4.3.1), however the VOI were matched with control variants only for MAF (and not nucleotide composition). The results from this analysis identified two groups of DNA words that were significant after strict Bonferroni multiple test correction and three of these groups were significant at empirical p-value threshold calculated by permutation. The DNA words of these three groups have only G and C nucleotides.

4.3.2.1. Validating matching parameter

The difference between the distribution of MAF in VOI and pair-wise matched control variants was confirmed using a Wilcoxon ranked sum test. The p-value of 0.962 confirmed no significant difference between the distributions of MAF. The histogram of the distribution is shown in Figure 4.7.

![Figure 4.7. Minor Allele Frequency (MAF) distribution for Variants of Interest (VOI) and Control variants](image)
4.3.2.2. Statistical findings

All 5-mer DNA sequences centered at VOI and control variants were extracted using RSAT. Next, 512 groups of 5-mers DNA words were tested for significant frequency difference between the extracted SOI and control sequences. DNA words from one group (Table 4.5) were absent in both SOI and control sequences. The p-value distribution of remaining 511 tests is shown in Figure 4.8. Fifty-two groups of DNA words were significant at p-value threshold of less than 0.05, (top 10 tests: Table 4.6). The mean of the distribution is 0.51 (SD= 0.33).

**Table 4.5. DNA words absent in both SOI and control sequences.**

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Word 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTTCG</td>
<td>CGAAC</td>
</tr>
</tbody>
</table>

Figure 4.8. P-value distribution for 542 Fisher exact tests (a) Histogram (b) QQ-plot
Table 4.6. Top 10 DNA words nominally significant at Fisher’s exact test p-value <0.05.

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Word 2</th>
<th>SOI Present</th>
<th>SOI Absent</th>
<th>Control sequence Present</th>
<th>Control sequence Absent</th>
<th>OR (Cl 95%)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCGGG</td>
<td>CCCGC</td>
<td>67</td>
<td>9036</td>
<td>21</td>
<td>9082</td>
<td>3.21</td>
<td>8.68E-07</td>
</tr>
<tr>
<td>CCGGG</td>
<td>CCCGG</td>
<td>93</td>
<td>9010</td>
<td>45</td>
<td>9058</td>
<td>2.08</td>
<td>5.04E-05</td>
</tr>
<tr>
<td>GGGCG</td>
<td>CGCCC</td>
<td>15</td>
<td>9088</td>
<td>1</td>
<td>9102</td>
<td>15.02</td>
<td>5.16E-04</td>
</tr>
<tr>
<td>GCGAG</td>
<td>CTCGC</td>
<td>31</td>
<td>9072</td>
<td>9</td>
<td>9094</td>
<td>3.45</td>
<td>6.71E-04</td>
</tr>
<tr>
<td>GCCGG</td>
<td>CCGCC</td>
<td>80</td>
<td>9023</td>
<td>44</td>
<td>9059</td>
<td>1.83</td>
<td>1.51E-03</td>
</tr>
<tr>
<td>AATAT</td>
<td>ATATT</td>
<td>20</td>
<td>9083</td>
<td>46</td>
<td>9057</td>
<td>0.43</td>
<td>1.82E-03</td>
</tr>
<tr>
<td>CGAGG</td>
<td>CCTCG</td>
<td>10</td>
<td>9093</td>
<td>0</td>
<td>9103</td>
<td>inf</td>
<td>1.95E-03</td>
</tr>
<tr>
<td>AAGAA</td>
<td>TCTCT</td>
<td>35</td>
<td>9068</td>
<td>65</td>
<td>9038</td>
<td>0.54</td>
<td>3.43E-03</td>
</tr>
<tr>
<td>GCGGC</td>
<td>GCCGC</td>
<td>33</td>
<td>9070</td>
<td>13</td>
<td>9090</td>
<td>2.54</td>
<td>4.48E-03</td>
</tr>
<tr>
<td>TACGA</td>
<td>TCGTA</td>
<td>12</td>
<td>9091</td>
<td>31</td>
<td>9072</td>
<td>0.39</td>
<td>5.35E-03</td>
</tr>
</tbody>
</table>

**Word 1-** 5-mer words; **Word 2-** Reverse complement of word 1; **SOI present-** the proportion of SOI containing the DNA words; **Control sequences present-** the proportion of control sequences containing the DNA words; **SOI absent-** the proportion of SOI not containing the DNA words; **Control sequences absent-** the proportion of control sequence not containing the DNA words; **OR-** Odds Ratio; **Cl-** Confidence Interval; **P-value-** Nominal p-value.
4.3.2.3. **Multiple Test Correction**

The p-value threshold after strict Bonferroni multiple test correction for 511 Fisher’s exact test is 9.79e-05. Two groups of DNA words (group 1: GCGGG/CCCGC and group 2: CCGGG/CCCGG) were significant at this threshold.

The empirical p-value threshold calculated from 1000 permutation runs for this analysis was 2.68e-4. The distribution of minimum p-value from all permutation runs is shown in Figure 4.9 (b). The top two groups of DNA words with lowest p-value (Table 4.6) are significant at the empirical p-value threshold. The distribution of the mean p-values from all 1000 permutation runs is expected to be normal with a mean of 0.05. The observed distribution of mean p-values is shown in Figure 4.9 (a) and the mean is equal to 0.61.

Figure 4.9. Distribution of (a) mean p-values (b) Minimum p-values, from 1000 permutation runs.
4.3.2.4. Nucleotide composition in SOI and control sequences

The top 10 groups of DNA words have enriched G and C nucleotides (Table 4.6). To confirm the influence of nucleotide composition on the results, I tested the difference between the distributions of proportion of each nucleotide (in 5-mer sequences) in SOI and control variants. Four Cochran Armitage trend tests for each nucleotide were performed (Table 4.7) and highly significant p-values confirmed that the distribution of proportion of nucleotides is different between SOI and control sequences.

Table 4.7. 2x6 contingency tables to test difference in distribution of nucleotide composition between SOI and control sequences using Cochran Armitage trend test.

<table>
<thead>
<tr>
<th>Nucleotide A count</th>
<th>Absent</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOI</td>
<td>2581</td>
<td>3509</td>
<td>2007</td>
<td>764</td>
<td>200</td>
<td>42</td>
</tr>
<tr>
<td>Control sequences</td>
<td>2103</td>
<td>3325</td>
<td>2438</td>
<td>970</td>
<td>237</td>
<td>30</td>
</tr>
<tr>
<td>P-value:</td>
<td>9.38e-23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide T count</th>
<th>Absent</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOI</td>
<td>2589</td>
<td>3411</td>
<td>2119</td>
<td>762</td>
<td>197</td>
<td>25</td>
</tr>
<tr>
<td>Control sequences</td>
<td>2139</td>
<td>3288</td>
<td>2441</td>
<td>974</td>
<td>232</td>
<td>29</td>
</tr>
<tr>
<td>P-value:</td>
<td>6.22e-21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide G count</th>
<th>Absent</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOI</td>
<td>2285</td>
<td>3328</td>
<td>2258</td>
<td>1001</td>
<td>222</td>
<td>9</td>
</tr>
<tr>
<td>Control sequences</td>
<td>2697</td>
<td>3400</td>
<td>2085</td>
<td>767</td>
<td>147</td>
<td>7</td>
</tr>
<tr>
<td>P-value:</td>
<td>1.24e-20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide C count</th>
<th>Absent</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOI</td>
<td>2216</td>
<td>3268</td>
<td>2355</td>
<td>990</td>
<td>256</td>
<td>18</td>
</tr>
<tr>
<td>Control sequences</td>
<td>2616</td>
<td>3453</td>
<td>2079</td>
<td>794</td>
<td>151</td>
<td>10</td>
</tr>
<tr>
<td>P-value:</td>
<td>3.01e-24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Column header represents counts of nucleotide in the 5-mer long sequences of interest (SOI).
4.3.3. Tertiary analysis: Including LD proxies for VOI and control variants

This section documents the results from the tertiary analysis where the bi-allelic VOI were selected using the phase 3 data from the 1000 Genomes project (European super-population) and all the LD proxies were included in the analysis. The results from this analysis have inflated type 1 error. These signals were no longer significant when the LD proxies were not included in the analysis. The flowchart explaining data selection at each step is shown in Figure 4.10.

4.3.3.1. VOI and control variants

Out of the 10,298 unique VOI, 8,595 SNPs were identified as uniquely bi-allelic based on 1000 Genomes phase 3 data – European super-population, hosted in SNiPA. These 8,595 SNPs were selected as VOI for this analysis.

The VOI were pair-matched with control variants in the list of non-eQTLs for only MAF. To keep consistency in data used, all SNP-gene pairs tested in the GTEx project were searched for bi-allelic SNPs based on 1000 Genomes phase 3 data (European super-population, hosted in SNiPA). The new list of unique bi-allelic non-eQTLs includes 7,588,152 SNPs.

The list of non-eQTLs was filtered with SNPs having MAF within ±1% window size of the VOI being matched and only one SNP was randomly selected as a control variant from the filtered list.

4.3.3.2. SNPs in high Linkage disequilibrium (LD)

SNiPA was also used to identify all bi-allelic SNPs that were in high LD ($r^2>0.8$) with all VOI and control variants. 300,475 VOI and 360,033 control variants were included in the list at this step.

VOI: The mean of the number of LD proxies across all SNPs is 35.1 (SD= 65.7). Three SNPs (Table 4.8) have large number of proxies (>1500).
Control variants: The mean number of LD proxies across all control variants is 40.9 (SD=80.4). Five SNPs (Table 4.8) have large number of proxies (>1500).

Wilcoxon ranked sum test p-value of 3.84e-27 confirmed that there is a significant difference between the distribution of LD proxies of VOI and control variants, with larger number of proxies for control than VOI.

**Table 4.8.** Pairwise matched VOI and control variants with large number of proxies (>1500).

<table>
<thead>
<tr>
<th>rsID</th>
<th>MAF</th>
<th>Number of Proxies* Before</th>
<th>After</th>
<th>% dropped</th>
<th>rsID</th>
<th>MAF</th>
<th>Number of Proxies* Before</th>
<th>After</th>
<th>% dropped</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs184745911</td>
<td>0.24</td>
<td>1939</td>
<td>1878</td>
<td>3.15</td>
<td>rs8141671</td>
<td>0.23</td>
<td>83</td>
<td>83</td>
<td>0</td>
</tr>
<tr>
<td>rs62055660</td>
<td>0.24</td>
<td>2129</td>
<td>2063</td>
<td>3.10</td>
<td>rs113374742</td>
<td>0.24</td>
<td>23</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>rs543638572</td>
<td>0.24</td>
<td>2226</td>
<td>2151</td>
<td>3.34</td>
<td>rs2881611</td>
<td>0.24</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>rs7184653</td>
<td>0.24</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>rs62060853</td>
<td>0.24</td>
<td>2526</td>
<td>2442</td>
<td>3.44</td>
</tr>
<tr>
<td>rs1545503</td>
<td>0.26</td>
<td>124</td>
<td>120</td>
<td>3.33</td>
<td>rs9915547</td>
<td>0.25</td>
<td>2340</td>
<td>2260</td>
<td>3.42</td>
</tr>
<tr>
<td>rs7311078</td>
<td>0.24</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>rs17664048</td>
<td>0.24</td>
<td>2189</td>
<td>2117</td>
<td>3.29</td>
</tr>
<tr>
<td>rs9296857</td>
<td>0.25</td>
<td>46</td>
<td>46</td>
<td>0</td>
<td>rs62064364</td>
<td>0.24</td>
<td>1737</td>
<td>1682</td>
<td>3.17</td>
</tr>
<tr>
<td>rs12635828</td>
<td>0.09</td>
<td>46</td>
<td>45</td>
<td>2.17</td>
<td>rs61902020</td>
<td>0.09</td>
<td>1658</td>
<td>1601</td>
<td>3.44</td>
</tr>
</tbody>
</table>

This table contains 8 VOI and pairwise matched control variants that have a number of LD proxies more than 1500. *The number of proxies before and after performing the dropping algorithm is shown.
Figure 4.10. Flowchart representing the steps for data pre-processing to selected variants of interest (VOI) and control variants for the tertiary analysis. The data is extracted from the GTEx project for thyroid tissue.
4.3.3.3. **SNPs with overlapping DNA**

The SNPs in VOI or control variants that are at least 5bp apart from each other were included in the analysis. After implementing this parameter using my algorithm (explained in section 3.2.2), 295,090 VOI (1.8% dropped) and 352,935 control variants (2.0% dropped) were included in the analysis. The number of SNPs dropped at the loci with largest number of proxies is shown in Table 4.8.

4.3.3.4. **Statistical findings**

Out of the total 512 Fisher’s exact tests, 294 (57%) have a nominally significant p-value of less than 0.05 (Figure 4.11). Under the null distribution, 26 SNPs are expected to be significant at p-value threshold of less than 0.05. The mean of the p-value distribution deviates drastically from the expectation (0.5) to 0.19. The top ten most significant groups of DNA words are shown in Table 4.9. The top most groups of DNA words is same as in results from secondary analysis (Table 4.6).

![Figure 4.11. P-value distribution for 512 Fisher’s Exact tests (a) Histogram (b) QQ-plot.](Image)
Table 4.9. Top 10 significant groups of DNA words from 512 Fisher’s Exact tests.

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Word 2</th>
<th>SOI Present</th>
<th>SOI Absent</th>
<th>Control Sequences Present</th>
<th>Control Sequences Absent</th>
<th>OR (CI 95%)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCGGG</td>
<td>CCCGG</td>
<td>2097</td>
<td>292993</td>
<td>1574</td>
<td>351361</td>
<td>1.60</td>
<td>4.45E-45</td>
</tr>
<tr>
<td>GGCAG</td>
<td>CCCGC</td>
<td>1832</td>
<td>293258</td>
<td>1340</td>
<td>351595</td>
<td>1.64</td>
<td>2.19E-43</td>
</tr>
<tr>
<td>CCGGG</td>
<td>CCCGC</td>
<td>1377</td>
<td>293713</td>
<td>985</td>
<td>351950</td>
<td>1.68</td>
<td>1.53E-35</td>
</tr>
<tr>
<td>GCCGG</td>
<td>CCGGC</td>
<td>1372</td>
<td>293718</td>
<td>988</td>
<td>351947</td>
<td>1.66</td>
<td>1.16E-34</td>
</tr>
<tr>
<td>GCGTG</td>
<td>CACGC</td>
<td>1942</td>
<td>293148</td>
<td>1669</td>
<td>351266</td>
<td>1.39</td>
<td>2.77E-23</td>
</tr>
<tr>
<td>GGTGG</td>
<td>CCACC</td>
<td>1274</td>
<td>293816</td>
<td>1024</td>
<td>351911</td>
<td>1.49</td>
<td>1.83E-21</td>
</tr>
<tr>
<td>GCCGC</td>
<td>GCGCC</td>
<td>1185</td>
<td>293905</td>
<td>942</td>
<td>351993</td>
<td>1.51</td>
<td>5.07E-21</td>
</tr>
<tr>
<td>GGAGG</td>
<td>CCTGC</td>
<td>968</td>
<td>294122</td>
<td>736</td>
<td>352199</td>
<td>1.58</td>
<td>1.19E-20</td>
</tr>
<tr>
<td>ATATA</td>
<td>TATAT</td>
<td>1164</td>
<td>293926</td>
<td>1925</td>
<td>351010</td>
<td>0.72</td>
<td>9.88E-19</td>
</tr>
</tbody>
</table>

**Word 1** - 5-mers words; **Word 2** - Reverse complement of word 1; **SOI present** - the proportion of SOI containing the DNA words; **Control sequences present** - the proportion of control sequences containing the DNA words; **SOI absent** - the proportion of SOI not containing the DNA words; **Control sequences absent** - the proportion of control sequence not containing the DNA words; **OR** - Odds Ratio; **CI**: Confidence Interval; **P-value** - Nominal p-value.

4.3.3.5. **Multiple Test correction**

The p-value threshold after strict Bonferroni multiple test correction for 512 Fisher’s exact test is 9.77e-05. 125 out of 294 groups of DNA words remained significant after Bonferroni multiple test correction.
4.4. Discussion

4.4.1. Comparing the three analyses

As presented and discussed in chapter 3, the data form the AD GWAS study did not have enough power to identify any relationship between AD associated loci and DNA words. Therefore, I decided to increase the number of VOI selected from 20 AD associated loci to 10,610 eQTL in thyroid from the GTEx project. Out of these 10,610 eQTLs, 9,103 bi-allelic SNPs were selected for my analysis (explained in sections 4.3 and 4.3.1). SNPs in the HLA region (145 out of 9,103 VOI) were included in my analysis. However, none of these 145 SNPs were less than 5bp apart from each other and therefore no overlapping DNA was included in the analysis. The methods and results for three different analyses were presented earlier in this chapter and Table 4.10 contains a consolidated comparison between them.

The tertiary analysis was performed (refer to section 4.3.3) on 5bp long sequences centered at the index SNPs and their LD proxies. More than 50% of the groups of DNA words tested had highly significant nominal p-values, of which 125 remain significant after Bonferroni multiple test correction (refer to section 4.3.3.4). The number of significant groups of DNA words was 4 times more than the expectation under the null hypothesis (26 groups). A possible reason for the unexpected result can be the increase in data size, only by capturing the loci that are in high LD with each other, as opposed to identifying additional independent eQTLs.

To further understand the origin of the large number of significant results, secondary analysis was performed by repeating the method for all SOI and control sequences, but excluding the sequences around LD proxies (for any VOI or control variants) (refer to section 4.3.2). The control variants were matched based on same criteria i.e., MAF, however the source of data was switched from 1000 Genome phase 3 to phase 1 data (for European super-population). This decision was made to maintain consistency of the data source used in my method and the GTEx analysis (further discussed section 4.4.3). The secondary analysis identified two groups of DNA words that have significantly higher frequency of occurrence in SOI compared to control sequences, after Bonferroni correction for multiple testing. These DNA words have high GC content. Four Cochran Armitage trend tests confirmed the difference between the distribution of
nucleotide composition in VOI and control variants (Table 4.7), in the secondary analysis. A and T have higher frequency in controls sequence than in SOI, whereas G and C are more frequent in SOI. High GC content characterizes the promoter regions. The matched control variants were SNPs tested in the GTEx project and therefore are present in the ±1Mb window region of the TSS. When the control variants were matched for composition of nucleotides, these DNA words no longer met significance criteria (nominal p-values: group 1 (GCGGG/CCCGC): 0.724; group 2 (CCGGG/CCCGG): 0.717) (refer to section 4.3.1). This indicates towards the possibility of over matching the control variants. None of the DNA words was found to be associated with sequences of interest at empirical p-value cut-off threshold in the primary analysis. Moreover, it’s also important to note that I have performed a total of 1532 (509+511+512) tests, in all three analyses that may contribute towards correction for multiple testing. However, as the VOI were shared among all three analyses, it’s challenging to identify the actual number of independent tests performed (refer to section 5.4.3).

Table 4.10. Flow table comparing data and results from three different GTEx analysis.

<table>
<thead>
<tr>
<th>Steps</th>
<th>Primary analysis</th>
<th>Secondary analysis</th>
<th>Tertiary analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data selection</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VOI selected</td>
<td>9,103</td>
<td>9,103</td>
<td>8,595</td>
</tr>
<tr>
<td>Matching parameters for control variants</td>
<td>MAF+NC</td>
<td>MAF</td>
<td>MAF</td>
</tr>
<tr>
<td>Data source (bi-allelic SNPs)</td>
<td>1KG-P1</td>
<td>1KG-P1</td>
<td>1KG-P3</td>
</tr>
<tr>
<td>Control variants selected</td>
<td>9,103</td>
<td>9,103</td>
<td>8,595</td>
</tr>
<tr>
<td><strong>Methods</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extract all LD proxies for VOI and control variants</td>
<td>NA</td>
<td>NA</td>
<td>VOI: 300,475</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Control variants: 360,033</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drops SNPs within 5bp of each other</td>
<td>NA</td>
<td>NA</td>
<td>VOI: 295,090</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.018% dropped)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Control variants: 352,935</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.019% dropped)</td>
</tr>
<tr>
<td>Extract DNA sequence (SOI and control sequences)</td>
<td>5-mer centered at index SNP</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4.10. Continued. Flow table comparing data and results from three different GTEx analysis.

<table>
<thead>
<tr>
<th>Steps</th>
<th>Primary analysis</th>
<th>Secondary analysis</th>
<th>Tertiary analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA words counting (All possible 5-mers motifs)</td>
<td></td>
<td>512 groups of DNA words</td>
<td></td>
</tr>
<tr>
<td>Statistical Test</td>
<td>509 Fisher’s-exact</td>
<td>511 Fisher’s-exact</td>
<td>512 Fisher’s-exact</td>
</tr>
</tbody>
</table>

**Results**

<table>
<thead>
<tr>
<th>P-value &lt;0.05</th>
<th>15</th>
<th>52</th>
<th>294</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bonferroni P-value threshold:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;9.77e-5 (0.05/512)</td>
<td>NA</td>
<td>NA</td>
<td>126</td>
</tr>
<tr>
<td>&lt;9.79e-5 (0.05/511)</td>
<td>NA</td>
<td>2</td>
<td>NA</td>
</tr>
<tr>
<td>&lt;9.82e-5 (0.05/509)</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>P-value=1</td>
<td>121</td>
<td>82</td>
<td>6</td>
</tr>
<tr>
<td>Significant at Empirical p-value threshold</td>
<td>0</td>
<td>2</td>
<td>NA</td>
</tr>
</tbody>
</table>

For Table 4.10:

**Primary analysis**: 5-mer long SOI and control sequences matched (no proxies) for MAF and nucleotide composition using 1KG phase 1 data (EUR population).

**Secondary analysis**: 5-mer long SOI and control sequences matched (no proxies) for MAF using 1KG phase 1 data (EUR population).

**Tertiary analysis**: 200 bp long SOI and control sequences matched (with proxies) for MAF using 1KG phase 3 data (EUR population).

NA: Not applicable. P-value: Nominal p-value calculated from the Chi-square of Fishers Exact Test. **Empirical P-value**: Calculated based on permutation results (Section 4.2.2.7). MAF: Minor Allele Frequency. NC: Nucleotide Composition. 1KG-P1: 1000 Genomes Phase 1 (EUR). 1KG-P3: 1000 Genomes Phase 3 (EUR). SOI: Sequences of interest.

**4.4.2. Functional interpretation of significant DNA words around thyroid eQTLs**

DNA words that are enriched near thyroid eQTLs may have the potential to inform us about biological role of trait-associated variants. I briefly investigated significant DNA words from the secondary analysis (refer to section 4.3.2.2). The DNA words GCGGG/CCCGC (group 1; P=8.68e-07) and CCGGG/CCCGG (group 2; P=5.05e-05) were significant enriched in DNA regions around thyroid eQTL compared to matched control regions in the genome. These two
groups of DNA words were submitted to TOMTOM (Gupta et al., 2007) online service by MEME suite, to search for potential matches with known human transcription factors documented in different databases (http://meme-suite.org/db/motifs). This search was performed across 771 transcription factors listed in the HOMo sapiens COMprehensive MOdel Collection (HOCOMOCO) database (v11 FULL), which contains all human transcription factors models produced by ChipMunk (motif discovery tool) (Kulakovskiy et al., 2010).

TOMTOM matched group 1 and group 2 with core sequences of 29 and 11 transcription factors, respectively. The top 7 out of 29 matches for group 1 remained significant at FDR correction q-value threshold < 0.05. However, none of the results survive Bonferroni p-value cut-off (P<6.5e-5) for multiple test correction. The top matches for group 1, which are significant at FDR q-value < 0.05, are shown in Table 4.11 (SP1 and its isoform were matched separately and are included only once in the table). On the other hand, none of the 11 matches for group 2 were significant after FDR or Bonferroni MTC. However, nominally significant matches for group 2 are shown in Table 4.12.

Next, I explored biological functions of these TFs using the Uniport database (The UniProt Consortium, 2017). These functions are listed in Table 4.11 and Table 4.12. The findings from this exploration of significant DNA words near thyroid eQTLs are preliminary and require more extensive evaluation. However, this exploration has demonstrated a prospective trajectory to understand the reason and biological role of association between genetic variants and gene expression levels.
Table 4.11. TFs matched with group 1 DNA words (GCGGG/CCCCG) and significant at FDR q-value < 0.05.

<table>
<thead>
<tr>
<th>Transcription factor</th>
<th>p-value</th>
<th>q-value</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCM2</td>
<td>8.79e-05</td>
<td>3.85e-02</td>
<td>An activator of transcription. Role in parathyroid gland development. Mutations on GCM2 are involved in hyperparathyroidism (Guan et al., 2016).</td>
</tr>
<tr>
<td>E2F4</td>
<td>1.13e-04</td>
<td>3.85e-02</td>
<td>Transcription activator involved in cell regulation and DNA replication (Attwooll et al., 2004).</td>
</tr>
<tr>
<td>E2F1</td>
<td>1.26e-04</td>
<td>3.85e-02</td>
<td>Activator involved in cell cycle regulation and DNA replication (Attwooll et al., 2004).</td>
</tr>
<tr>
<td>SP1</td>
<td>1.26e-04</td>
<td>3.85e-02</td>
<td>Activator or repressor for physiological and pathological stimuli. Influence the regulations of many genes involved in cell growth, apoptosis, differentiation and immune responses. Highly expressed in cancer and tumour cells (Safe et al., 2014).</td>
</tr>
<tr>
<td>TFDP1</td>
<td>1.26e-04</td>
<td>3.85e-02</td>
<td>Can stimulate E2 dependent transcriptions and hence also involved in cell cycle regulation and DNA replication.</td>
</tr>
<tr>
<td>SP2</td>
<td>2.26e-04</td>
<td>4.94e-02</td>
<td>Involved in selective activation of mRNA synthesis by binding to GC box promoter elements.</td>
</tr>
</tbody>
</table>

**p-value** - The nominal probability that the match has occurred by chance under the null hypothesis.; **q-value**: The minimum false discovery rate required to include the match; **Functions**: Reported functions for the transcription factors on Uniport database (http://www.uniprot.org/uniprot). (The UniProt Consortium, 2017) or identified based in PubMed literature search. **GCM2**- Glial Cells Missing Homolog 2; **E2F4**- E2F Transcription Factor 4; **E2F1**- E2F Transcription Factor 4; **SP1**- Specificity Protein 1; **TFDP1**- Transcription Factor Dp-1; **SP2**- Specificity Protein 2.

TOMTOM performed search in a list of 771 TFs. (Bonferroni P-value threshold: 0.05/771= 6.5e-5): none TF significant at this threshold.
Table 4.12. TFs matched with group 2 DNA words (CCGGG/CCCCG) and significant at nominal p-value < 0.05.

<table>
<thead>
<tr>
<th>Transcription factor</th>
<th>p-value</th>
<th>q-value</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZF64A</td>
<td>1.51e-04</td>
<td>0.16</td>
<td>Involved in transcriptional repression.</td>
</tr>
<tr>
<td>MECP2</td>
<td>1.01e-03</td>
<td>0.51</td>
<td>The protein binds specifically with a single methyl-CpG pair and mediates transcription through interaction with histone de-acyetylase and specific co-repressors (Wood and Zhou, 2016).</td>
</tr>
<tr>
<td>KLF1</td>
<td>3.35e-03</td>
<td>0.96</td>
<td>Involved in maintaining foetal haemoglobin levels by activation and repression of specific genes (Huang et al., 2015:1)</td>
</tr>
<tr>
<td>ELF1</td>
<td>3.35e-03</td>
<td>0.97</td>
<td>Partly responsible for activation of some promoters and gene expression.</td>
</tr>
<tr>
<td>KLF3</td>
<td>6.41e-03</td>
<td>1</td>
<td>Involved in hematopoiesis and also binds with CACCC box of erythroid cell-expressed genes (Dewi et al., 2015).</td>
</tr>
<tr>
<td>SP1</td>
<td>6.1e-03</td>
<td>1</td>
<td>Activator or repressor for physiological and pathological stimuli. Influence the regulations of many genes involved in cell growth, apoptosis, differentiation and immune responses (Safe et al., 2014).</td>
</tr>
</tbody>
</table>

p-value: The nominal probability that the match has occurred by chance under the null hypothesis.; q-value: The minimum false discovery rate required to include the match; Functions: Reported functions for the transcription factors on Uniport database (http://www.uniprot.org/uniprot). (The UniProt Consortium, 2017) or identified based in PubMed literature search. ZF64: Zinc finger protein 64 homolog; MECP2: Methyl-CpG-binding protein; KLF1: Krueppel-like factor 1; ELF1: ETS-related transcription factor Elf-1; TFDP1: Transcription Factor Dp-1; KLF3: Krueppel-like factor 3.

TOMTOM performed search in a list of 771 TFs. (Bonferroni P-value threshold: 0.05/771= 6.5e-5): none TF significant at this threshold.

### 4.4.3. The 1000 Genomes project: phase 1 vs. phase 3 data

About 88 million variants from 2,502 individuals across 26 populations are listed in the 1000 Genomes phase 3 data and over 38 million variants from 1092 individuals across 14 populations were included in phase 1 data (The 1000 Genomes Project Consortium, 2015). Some of the samples from phase 1 data were not included in the phase 3 analyses, but many more are in phase 3 than phase 1. Approximately 2.3 million sites are not present in phase 3 compared to phase 1, of which 1.92 million are either indels or structural variants. A mixture of Illumina, ABI SOLiD and 454 sequencing platforms were used in phase 1, as opposed to only Illumina sequencing platform in phase 3. Moreover, a mixture of 36bp and >100bp reads were used in phase 1, whereas only >70bp long reads were used in phase 3. In addition to that, different
variant and genotyping-calling pipelines were used between phase 1 and 3. The genotype calling in phase 3 was performed using ShapeIT2 and MVNcall that allowed integration of multi-allelic variants. These are possible reasons that some multi-allelic SNPs in phase 3 are reported bi-allelic in phase 1.

In chapter 4, one of the differences between the three analyses was switching the data sources from phase 3 to phase 1 data, to identify bi-allelic SNPs and calculate MAF. The number of bi-allelic eQTLs based on phase 1 data is equal to the original number of significant bi-allelic SNPs that are identified by the GTEx project. Among 10,610 thyroid eQTLs, 9,103 and 8,595 SNPs were bi-allelic SNPs as per the 1000 Genomes project phase 1 and phase 3 data, respectively (European super-population) (Table 4.10). Therefore, phase 1 was used in the primary analysis with larger numbers of eQTLs.

4.4.4. Summary

The results from all three analyses generate three possible conclusions: (1) The SOI are still underpowered to capture the frequency differences for some of the DNA words (refer to section 4.4); (2) The matching parameters are limiting selection of effective control sequences; (3) The hypothesis underlying this study may be incorrect. Some other highlights from the analyses are:

1. Including DNA sequences centered at all SNPs in high LD with VOI and control variants highly increases the number of false positive results.
2. G and C enriched DNA words were found to have significantly higher frequency of occurrence in the SOI compared to control sequences. These signals disappeared when the match was performed for nucleotide composition, but this is likely due to over-matching.
Chapter 5
General Discussion
My research project has investigated the frequency difference of 5bp long DNA words between sequences of varying length (201bp or 5bp sequences centered at VOI) centered at Alzheimer’s disease and thyroid gene expression associated loci, respectively (separately), and matched control sequences that represent the rest of the genome.

The first analysis was performed on 355 DNA sequences that were 201bp long and centered on 20 GWAS SNPs associated with AD (and all SNPs in high LD). The 20 AD associated loci identified by the largest meta-study were pairwise matched with an equal number of SNPs in the rest of the genome based on various parameters using SNPsnap (refer to section 3.2.1.1). We showed that the results had more than expected (under the null) DNA words that differed in frequency, but no words survived correction for multiple testing (refer to section 3.3.3). Further investigation suggested that the data was underpowered to identify any associated DNA words (refer to section 3.3.5).

The second data analyzed contained 5bp long DNA sequences, centered on 9,103 eQTLs from the GTEx project (refer to Chapter 4) in thyroid tissue. All top eQTLs for each eGene were selected as VOI. Different results were obtained by changing the matching parameters. When control variants were matched with VOI only for MAF, two groups of DNA words showed significant differences in frequency of occurrence (between 5 bp long SOI and control sequences; more frequent in SOI). Preliminary analysis showed that these DNA words were matched with core regions of known transcription factors using TOMTOM (refer to section 4.4.2). These signals were statistically not strong enough to make concrete conclusions. However, these results indicate that my method has the potential to identify words associated with non-coding disease or gene expression associated loci. I also showed that the two significant groups of DNA words were no longer significant when the control variants were matched for nucleotide composition in addition to MAF (refer to section 4.3.1).

With regard to my initial hypotheses in section 2.1,

1. None of the DNA words have significantly different frequencies of occurrence around top GWAS hits for Alzheimer’s disease and matched control sequences.
2. Two groups of DNA words showed significantly frequent occurrences in DNA sequences around top eQTLs in thyroid tissues when compared to DNA sequences around non-eQTLs that were matched only for MAF.

3. These two groups of DNA words were no longer significant when the control variants were also matched for nucleotide composition along with MAF. Moreover, inclusion of LD proxies for eQTL data resulted into highly inflated type 1 error and this indicates towards need for improvement in data selection parameters.

5.1. Comparison with known motif discovery algorithms

Already existing bioinformatics tools to perform motif discovery like DREME, PWMs and RSAT are potential alternate methods to investigate biological sequences. My method is a different way of approaching the research question than already known methods because it incorporates a straightforward approach of “exact” DNA word counting (no ambiguity allowed) of a user-defined length. Some studies have shown the importance of exact word analysis in characterization of genomic words. For example, a study has shown that the distribution of distances between 6-mer and 7-mer DNA words and the nearest reverse complements in the human genome, is significantly different from expected distribution (obtained under k-order Markov model) (Tavares et al., 2017). The frequency of DNA words around GWAS hits or eQTLs have not been extensively evaluated, which motivated my project.

DREME is a well-known online motif discovery online tool that has marked a standard in analyzing high volume ChIP-seq experiment data. DREME searches exact-matches for nucleotide symbols defined by International Union of Pure and Applied Chemistry (IUPAC) (Cornish-Bowden, 1985; Johnson, 2010) in user provided biological sequences. IUPAC nomenclature allows ambiguity at any position of the biological sequences analyzed, unlike my method. Moreover, DREME uses a different methodology pipeline to identify regular expression motifs compared to my method. Position weight matrix (PWM) is another approach that can be used to quantify the proportion of nucleotide composition at every position of the analyzed DNA sequence, thus allowing ambiguity. PWM are also used by RSAT to analyze biological sequences in the human genome. Oligo- and Dyad- motif discovery algorithms,
unlike my method used sequence characteristics (for example overall nucleotide frequency in
the genome; in yeast genome $f_A=f_T=0.31; f_G=f_C=0.19$) when calculating the expected
background frequency.

5.2. DNA word discovery: AD GWAS hits versus thyroid eQTLs

In this section, I explain the difference between the analyses performed to identify DNA words
near AD associated loci (chapter 3) and GTEx identified thyroid eQTLs (chapter 4). I chose to
investigate eQTLs around thyroid to increase my data size from 20 AD associated loci (and all
SNPs in high LD) to 9,103 bi-allelic eQTLs. Adjacent DNA words in 201 bp long sequences are
correlated and therefore I analyzed only 5bp long sequences when investigating DNA around
eQTLs. Five bp long sequences centered at AD associated loci are likely to have low frequency
for any possible 5-mer DNA word. Therefore, I did not analyze only 5bp sequences around AD
associated loci. The results with inflated type 1 error from the analysis where all LD proxies for
eQTLs were also investigated are shown in section 4.3.3. To control the influence of LD
proxies, the primary eQTL data analysis did not include any tag SNPs. The major differences
between the analyses performed on two data sets are shown in Table 5.1.

Table 5.1. Comparison between analyses using AD GWAS hits and thyroid eQTLs.

<table>
<thead>
<tr>
<th></th>
<th>Largest AD GWAS data</th>
<th>GTEx Thyroid eQTL data (Primary analysis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOI</td>
<td>20</td>
<td>9,103</td>
</tr>
<tr>
<td>LD proxies</td>
<td>Included</td>
<td>Not included</td>
</tr>
<tr>
<td>Matching Parameters</td>
<td>SNPsnap parameters:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• MAF</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Gene density (locus</td>
<td></td>
</tr>
<tr>
<td></td>
<td>definition)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Distance from the</td>
<td></td>
</tr>
<tr>
<td></td>
<td>nearest gene</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Number of Proxies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(LD buddies)</td>
<td></td>
</tr>
<tr>
<td>Sequence length analyzed</td>
<td>201bp</td>
<td>5bp</td>
</tr>
</tbody>
</table>

VOI- Variants of Interest; LD- Linkage disequilibrium; AD- Alzheimer’s disease; GWAS- Genome-wide
association analysis; eQTL- expression quantitative trait loci; MAF- Minor allele frequency
5.3. Data file formats and other computational challenges

The data files extracted from different sources often had inconsistent nomenclature. Therefore, annotation and merging of files was required to perform my analysis. MAF was required to perform matching between VOI and control variants. Although, the GTEx project files for all significant SNP-gene pairs contain MAF for SNPs, the files listing all non-significant SNP-gene pairs (non-eQTLs) do not provide MAF information. This information was extracted from the SNiPA database that contains MAF for all bi-allelic SNPs calculated based on the 1000 Genomes project phase 1 and phase 3 data for European super-population. The samples from the GTEx project were distributed across different populations with the majority being Europeans (83.7%; refer to section 4.1.1). Therefore, I used the 1000 Genomes project data from European super-population, which represents the best single match for the GTEx data.

SNiPA requires rsID identifiers, however the GTEx files contain only the physical coordinates of variants. The rsIDs and allele information were extracted from the dbSNP build 147 (Sherry et al., 2001).

In addition, other online tools that were used in this project, like UCSC genome table browser, SNPSnap, and LDLink, have different input requirements and limitations. For example, UCSC browser table has an upper limit of queries (=1000 queries) that can be processed at one time and therefore I replaced the browser table with RSAT for the analysis performed using GTEx eQTLs (refer to section 4.2.2.5). Similarly, the initial version of LDLink used in this project allowed only a single query search and the programmatic access to perform multiple query searches was only made available in early 2017. SNPSnap was unable to perform matching for 9,103 eQTLs due to high computational data volume and therefore an in-house R script was developed to perform the matching. Programmatic access to SNPSnap to perform multiple query searches is not available.

Using various tools within a single automated pipeline is difficult and therefore I developed R functions to perform each step of the algorithm separately. Separate R programs require additional manual supervision and therefore computational efficiency is compromised.
However, computer scripts developed in this research project were individually efficient in terms to analyzing over 300,000 sequences of 5bp length (refer to section 4.3.3). The efficiency of my R scripts may decrease with expansion of sample size or DNA word length.

5.4. Limitations of my method

There are some limitations to my method, most of which fall into three broad categories: (1) SOI and control sequences with overlapping DNA, (2) Genetic sequence characteristics, and (3) Multiple testing correction.

5.4.1. SNPs with overlapping DNA

I developed a strategy to select SNPs that are sufficiently distant from one another to avoid including the same DNA sequences in the analyses (refer to section 3.2.2). Another strategy that may be explored in future is to merge the overlapping DNA sequences centered at two close by SNPs. However, either way the algorithm works with only one index SNP (VOI or control variant) and its tag SNPs when identifying clusters (refer to section 3.2.2). This approach does not address the possibility of having two or more index SNPs close enough to each other to have overlaps in flanking DNA or in the DNA flanking their respective tag SNPs. To confirm that this limitation did not influence my results from analysis on AD GWAS data, one pair of SNPs from the list of AD associated VOI (rs28834970 and rs9331896) was further investigated. These two VOI are the only two SNPs included in the analysis that are located on the same chromosome and are physically close (272,565 bp) to one another. All the LD proxies for both the SNPs are located within a range of 25,189 bp and 12,250 bp, respectively (LD \( r^2 > 0.8 \); distance between the farthest SNPs in LD with the index SNP). I also calculated the inter-marker distance between all 9,103 bi-allelic thyroid eQTLs (separately for all chromosomes) and identified only 1 SNP pair on chromosome 1 that is 2bp apart from each other (rs6541292 and rs10864704; only 5bp flanking regions are included in this analysis). Therefore, I concluded that my analysis was not influenced by inclusion of SNPs with overlapping flanking DNA sequences. However, this limitation can influence the results for other larger data sets and therefore, future work to improve on this is required.
5.4.2. Genetic Sequence characteristics

As mentioned earlier, over 300 million variations are reported in the human genome. Based on this information and considering the total length of the human genome (~3 billion nucleotides), at least 10 variations are expected to occur every 100bp in a DNA sequence across the population. My method of DNA words discovery does not include the possibility of having another variation(s) in the flanking DNA sequence(s) centered at the index SNP. This can also influence the results, as the allele (reference or alternate) included in this analysis at that position will alter the “exact” count of DNA words. In addition to that, my method was developed to analyze DNA sequences containing only the reference allele at the index position (VOI or control variants). The results may differ when the alternate alleles are used instead. One approach to incorporate this information is by using the IUPAC nomenclature for nucleotides (Cornish-Bowden, 1985; Johnson, 2010).

Higher number of variations present in the sequence leads to more possibilities of combination of alleles and hence, more number of possible sequences that can be analyzed. This complexity increases in the case of multi-allelic SNPs. It is possible to analyze multi-allelic SNPs using my method but only when the reference allele is analyzed.

My method cannot analyze sequences around indels because they may alter the length of DNA sequences. As mentioned earlier (refer to section 4.2), longer DNA sequences have sharing of nucleotides between adjacent DNA words. On the other hand, a smaller sequence will not exactly match with query DNA word. This limitation may be overcome by using PWMs (refer to section 1.7.1.2) that quantifies the probability of nucleotide occurrence at every position by aligning sequences with one another.
5.4.3. Multiple testing correction

Some of the statistical tests are performed on data that may not be completely independent of one another. This is because consecutive DNA words share nucleotides (letters). In other words, while counting all possible 197 DNA words (5bp long) in a 201bp long sequence (moving by 1 window), two consecutive words will differ only at 1 position (Tavares et al., 2017). Therefore, DNA words are correlated to one another. The Bonferroni approach of MTC assumes that all tests are independent of each other and therefore it is conservative for the method described in this project.

False Discovery Rate (FDR) is another widely used approach to perform MTC. FDR was not used in my method because it does not incorporate the correlated characteristics of DNA words as explained above. The Nyholt method (Nyholt, 2004) for MTC is performed by estimating the actual number of independent tests based on the correlation within the data. Nyholt used LD as a factor to incorporate the co-relation between tested SNPs and determine the actual number of independent tests for association studies. Although the concept used by Nyholt is closer to the requirements of my analysis, it is difficult to quantify the correlation between genetic sequences as compared to SNPs. This also limits the implementation of Li and Ji method for MTC, which uses eigenvalues of a correlation matrix to adjust multiple testing in multiple locus analyses (Li and Ji, 2005).

Permutation was performed to correct probability values for multiple testing as discussed in this thesis. Although this approach controlled the proportion of false positives, it increased the computational complexity. This can be a potential limiting factor when analyzing data sets with larger number of sequences.
5.5. Future Directions

The results from this research study demonstrate that there is a possible relationship between DNA words and trait associated genetic variations. These DNA words may have biological functionality that may help understanding the role of non-coding variations. Moreover, these variations may be used to identify additional trait-associated loci.

5.5.1. Alternate data sources

An important next step will be to implement the method on different data sources. Some of the possible interesting data sources that can be used are listed below.

5.5.1.1. Variations in the GWAS catalog

I chose to explore the genetic variations associated with Alzheimer’s disease due to the high impact of the disease on population (discussed in section 1.4). The number of associated variations for AD is low. This limits the method in terms of statistical power to discover associated DNA words. As introduced in section 1.4.4, NHGRI-EBI GWAS catalog search for keyword “Alzheimers disease” revealed 702 significant associated loci at p-value threshold less than 5e-04, of which 134 are genome wide significant (p-value < 5e-8). These variants are either associated with AD or related phenotypes and are likely to overlap with the VOI in my analysis because of sequential meta-analysis. Taking this information into account, NHGRI-EBI GWAS catalog can be used to identify other relevant variants of interest and investigation of DNA words around these additional loci may improve power. Moreover, the GWAS catalog also contains genetic variations associated with other complex diseases and body traits that can be a potential source of data, for example, Breast Cancer (1,571 associations listed in catalogue) and Body-Mass index (4,073 associations listed in catalogue). However, it is important to note that some studies included in the catalog may have overlapping samples that may be an additional potential limiting factor.
5.5.1.2. Further exploration of the GTEx project

In addition to cis-eQTL analysis, GTEx consortium has identified trans loci associated with the gene expression level across various tissues. Trans-eQTLs analysis was performed to identify association between all autosomal variations and all protein coding or lincRNA genes (GTEx Consortium, 2017). One of the major challenges faced by the trans-eQTL analysis is the correction for multiple testing across the genome. Six hundred and seventy three trans eQTLs associated with 93 unique genes were identified by the GTEx project across 16 different tissues (GTEx Consortium, 2017). DNA sequences centered at these 673 eQTLs may have significantly different frequency of words that perhaps may help identify additional trans-eQTLs. However, it is important to note that some trans-eQTLs may overlap with cis-loci.

DNA word analysis across all 44 tissues in the GTEx project could be another area to explore in future. GTEx analysis demonstrates sharing of cis-eGene across tissues (refer to section 4.1.3) that may enforce findings similar my results that are presented in this thesis. However, analyzing the non-shared cis-eQTLs may be one approach to reduce the correlation between cis-eGenes across tissues.

5.5.1.3. Whole blood eQTLs

The largest number of samples (338) was collected from whole blood tissue in the GTEx project and over 5,000 eQTLs were identified. One of the prospective future studies will be to perform the analysis on DNA sequences around eQTLs identified from the whole blood samples in the GTEx and then try to replicate findings in the independent study (Battle et al., 2014). One prospective data source that can be used for replication of results is from an independent eQTL study that analyzed whole blood tissue samples from 922 individuals (Battle et al., 2014). Each individual was genotyped for 720,591 autosomal SNPs of which 13,857 were locally (cis) associated with the gene expression of 10,914 genes.
5.5.2. Alternate matching parameters

The results shown in this study demonstrate that the data analyzed and matching parameters play an important role in influencing meaningful analysis. Here, I discuss different parameters that can be used in future studies to improve this accuracy of results.

5.5.2.1. SNPsnap parameters in eQTL data

SNPsnap web-service was used to match control variants for 20 AD associated variants of interest but this tool was not suitable for the eQTL data. Therefore, I developed a computer program to match control variants for MAF and nucleotide composition. In addition to MAF, SNPsnap also matched the control variants with VOI for additional parameter like distance from the nearest gene, gene density and number of proxies per SNP, which I did not implement in my program. The effect of incorporation of different matching parameters on results from eQTL data can be explored in future studies.

5.5.2.2. The inter-marker distance between LD proxies

As discussed in section 3.2.6.1, two SNPs with large number of proxies influence the results. Although, the VOI were pairwise matched to have similar number of LD proxies, the inter-marker distance between the proxies influences the results. SNPs with proxies closer to one another are likely to include SNPs in the analysis that have overlapping DNA sequences, which will subsequently be dropped. Therefore, an additional parameter controlling for inter-marker distances between the proxies of VOI or control variant may be used for better selection of control sequences.

5.5.3. Search for different length of DNA words and regions analyzed

In this thesis, I have performed analysis only for 5-mer DNA words, in 201 bp or 5bp long DNA regions around trait-associated loci. However, as mentioned earlier, my method is independent of DNA word length and investigated region’s length. Therefore, the immediate next steps of this research project could be to repeat the analysis and investigate the frequency of longer length of DNA words. One possible role of significant DNA words can be to act as binding sites
for transcription factors. I downloaded 537 human TFBS sequences from the JASPER database and plotted the distribution of length of these sequences (Figure 5.1). The length ranges from 5bp to 21bp with a mean length of 11.9.

Figure 5.1. Distribution of length of 537 human TFBS extracted from JASPER database

5.5.4. Identification of additional risk loci

The significant DNA words identified by analyzing GTEx thyroid eQTL data are enriched around eQTLs compared to matched controls (refer to section 4.3.2). Significant DNA words that are associated with GWAS variants or eQTLs can be searched for occurrence in other part of the genome. This information can be used to create a map that may enable identification of additional eQTLs in future studies.


Contributions

Some sections of chapter 1 are modified from a dissertation submitted at Lancaster University, United Kingdom towards an academic course taken under an international student exchange program. Along with primary supervisors, Dr. Frank Dondelinger at the Lancaster University evaluated and provided valuable feedback. No credits were awarded or transferred for this course.

Mr. Jon Pipitone, Mr. Andy Wang and Mr. David Rotenberg, staff members of the scientific computing at CAMH, provided technical and computational support throughout this project.