University of Plymouth

PEARL **https://pearl.plymouth.ac.uk**

04 University of Plymouth Research Theses 01 Research Theses Main Collection

2023

Investigating the genomic distribution and potential contribution of retrotransposable elements in relation to their potential impact on genome function and predisposition to human diseases.

Ali, Randa Aweis

http://hdl.handle.net/10026.1/20184

http://dx.doi.org/10.24382/1020
University of Plymouth

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's prior consent.

INVESTIGATING THE GENOMIC DISTRIBUTION AND POTENTIAL CONTRIBUTION OF RETROTRANSPOSABLE ELEMENTS IN RELATION TO THEIR POTENTIAL IMPACT ON GENOME FUNCTION AND PREDISPOSITION TO HUMAN DISEASES.

By

RANDA AWEIS ALI BSc (Hons)

A thesis submitted to the University of Plymouth in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Biomedical Sciences

October 2021

Acknowledgements

Praise is to Allah by whose grace good deeds are completed. First and foremost, I would like to extend my greatest gratitude and appreciation to my esteemed supervisor Dr. Elaine Green, whom I would not have gotten this far in my academic journey without her continuous support and encouragement. Thank you for always believing in me and motivating me to carry on, especially at the times when I was doubting myself. You have truly been an excellent mentor. I would also like to extend my deepest appreciation to Dr. Vasilis Lenis for being the best big brother I could have asked for. I cannot thank you enough for always coming to my aid no matter how busy you were and for your continued support, whether with coding issues or life in general. With special thanks to Roxane Dunbar, for helping with this research project and for putting up with my occasionally distracting behaviours in the office. I will always cherish the memories and the fun times we shared during what has been some of the most challenging years of our life so far.

Thank you to all my friends at Plymouth University, with special appreciation to Dr. Doaa Althalathini, Dr. Zainab Bu Sinnah, and Dr. Suhailah Ali, for their friendship, invaluable advice, and continuous support. I am eternally grateful for all the meals you have cooked for me, and I hope our friendship lasts a lifetime, no matter where each of us is around the world.

Hats off to my childhood friends, with special appreciation to Camilla Scanlan and Muna Rahman, who made my transition back to Plymouth so much easier than I could have ever imagined. I do not know how I would have survived without you guys keeping me sane! I do not have the words to express how much I cherish your friendship. All I can say is that I am truly blessed for having you in my life. With thanks from the bottom of my heart to my friend Jeannie Campbell for taking the time to patiently check through sections of my thesis, even the very long one you called “the beast”! I owe you big time!

A huge thank you to my family for being so incredibly kind and supportive, and for always cheering me on and believing in me. I am truly blessed with the best family I could have asked for. With special gratitude to my darling sister Nihad, for also being my best friend. Thank you for always knowing what to say, and for being a source of comfort and joy in what has been a difficult time in my life.

With special thanks to my husband Zaid Al-Hamdi, for being the best life partner, and for putting up with me with patience, kindness, and love. I will always appreciate everything you have done and continue to do for me.

Saving the best for last, with the biggest and most special appreciation to my parents, for I would not be the person I am today without them. To my mum, Zahra Ali Bafadhel, the dedicated housewife who always put us before herself. And to my Dad, Ali Aweis Bafadhel, for putting up with so much to make sure we always had everything we needed. I truly appreciate all the sacrifices you both have made for us so that we could enjoy a better life.

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee.

The database submitted for this research degree has formed part of Roxane Dunbar's doctor of philosophy degree at the University of Plymouth.

A programme of advanced study was undertaken, which included: BIOM5001 Molecular Biology: Genomics, Transcriptomics, and Proteomics.

Presentations at conferences:

Poster presentation: Plymouth university annual research event 2017

Poster presentation: Plymouth university annual research event 2018

Poster presentation: CSHL Transposable Elements, New York 2018

Poster presentation: Plymouth university annual research event 2019

Word count of main body of thesis: **32,874**

Signed:

Date: 11/01/2023

Abstract

Investigating the genomic distribution and potential contribution of retrotransposable elements in relation to their potential impact on genome function and predisposition to human diseases.

Randa Aweis Ali

Active retrotransposable elements (RTEs) provide a continuous source of genomic diversity in humans. The potential impact of a novel RTE insertion depends on its genomic location. Previous studies investigating the landscape of polymorphic RTEs report their higher fraction in functional regions compared with fixed elements, thereby highlighting the potential impact of RTE activity on genome function (GF). However, studies have only recently begun incorporating RTE variants (RTEV) in association with complex human diseases. This study aimed to investigate the impact of RTEs activity on GF and the potential association of RTEV with disease susceptibility. A comprehensive database of all non-reference L1s, Alus, and SVA insertions reported in the literature to April 2019 was curated (n=39,798 RTEs). The curated database includes numerous singleton and rare RTE insertions. Such insertions potentially faced fewer selection pressures compared with common RTEs, thus are likely more representative of RTEs preferred integration site. The genomic distribution of the curated RTEs was compared with the distribution of ancient RTEs that are fixed in the human genome to hypothesise the likely effect of new RTE insertions. Non-reference insertions were found at higher frequencies in functional regions and had a more even genomic distribution than fixed RTEs, suggesting their ability to impact GF. The positional overlap between RTEs and trait-associated SNPs

(TASs) was investigated to determine the potential of RTEs as causal variants in GWAS risk loci. L1s, Alus, and SVA elements were significantly enriched in GWAS risk regions, suggesting the potential impact of RTEV on human health. Next, 354 novel RTE-TAS associations were identified via linkage analysis between RTEVs and genome-wide significant SNPs identified in European populations. Finally, SVA elements likely impose the highest impact on GF and human health based on their genomic accumulation in functional regions and their higher proportion in GWAS risk regions. Collectively, the results of this thesis depict the functional impact of RTEs on GF and human health, which have proven to be invaluable for future association studies to further the current knowledge regarding the aetiology of complex traits and disorders.

Keywords: Transposable elements; Polymorphic retrotransposable elements; Structural variants; Causative variant; GWAS; Genome function; Human health

Table of Contents

Acknowledgements III
Author's Declaration IV
Abstract V
List of Tables: XI
List of Figures: XIV
List of Abbreviations: XVI
1. Introduction 1
1.1. Transposable elements (TEs): 1
1. 2. Retrotransposable elements (RTEs) 3
1.2.1. Classes of retrotransposable elements (RTEs) 3
1.2.2. Structural organisation of non-LTR RTEs 4
1.2.2.1. LINE 4
1.2.2.2. SINE: Alu 5
1.2.2.3. SVA 5
1.2.3. Mechanism of L1 retrotransposition 8
1.2.4. Alu and SVA elements hijack the L1 machinery 9
1.2.5. Effect of retrotransposition on genome function and integrity 10
1.2.6. Mechanisms of retrotransposition silencing 12
1.3. Functional role of TE activity in genome evolution 13
1.4. RTEs as structural variants 15
1.3.1. RTE variants and their detection in the human genome 15
1.4.2. The effect of RTE variants on human health and disease 17
1.5. Summary of current study 18
1.6. Aims and Objectives 19
1.6.1. Research Aims: 19
1.6.2. Main Objectives: 19
2. Database curation 20
2.1. Introduction: 20
2.1.1. Early methods of RTE discovery: 21
2.1.2. Current methods of RTE discovery: 22
2.1.3. Limitations of RTE discovery using NGS data: 25
2.1.4. Online databases for non-reference RTE insertions: 25

2.1.5. Study aims: 26
2.2. Methods: 27
2.2.1. Criteria of RTE database curation: 27
2.2.2. Study selection: 28
2.2.3. Data curation: 32
2.2.4. Quality control: 34
2.2.5. Addressing duplicate RTE profiles from the database by individual: 35
2.3. Results: 37
2.3.1. Study selection: 37
2.3.2. Database structure/content: 38
2.3.3. General database: 39
2.3.4. Databases by individual for L1Hs, AluY, and SVA_E/F: 40
2.4. Discussion: 45
2.4.1. Study database vs. existing online databases 45
2.4.2. Issues with current methods of RTE detection 46
2.4.3. Correlations between population growth rate and allele frequency spectrum 47
2.4.4. Correlations between population growth rate and efficiency of natural selection 48
2.4.5. Study overview 49
3. Genomic Distribution of non-LTR RTEs 50
3.1. Introduction 50
3.1.1. Retrotransposition of L1s, Alus, and SVAs 50
3.1.2. Distribution of endogenous RTEs 51
3.1.3. Effects of RTE activity on genome function 52
3.1.4. Effects of RTE activity on human health 52
3.1.4.1. RTEs and monogenic diseases 52
3.1.4.2. RTEs and complex diseases 53
3.1.5. Genomic distribution of recent RTE insertions 54
3.2. Methods 61
3.2.1. RTE Datasets 61
3.2.1.1. Non-reference database: 61
3.2.1.2. Reference database: 61
3.2.2. Genomic Distribution analyses 63
3.2.2.1. Chromosomal distribution 63
3.2.2.2. Local GC content 64

3.2.2.3. Distribution in functional regions 65
3.2.2.4. Local recombination rate 67
3.2.2.5. RTEs Enrichment in functional regions and accessible chromatin domains 69
3.3. Results 72
3.3.1. Chromosomal distribution 72
3.3.1.1. Linear regression analysis of RTE and chromosome size or gene density 73
3.3.2. Local GC content 80
3.3.3. Distribution in functional regions 85
3.3.4. Local recombination rate 92
3.3.5. Local chromatin accessibility 98
3.4. Discussion 104
3.4.1. Chromosomal distribution 105
3.4.2. Local GC content 108
3.4.3. Distribution in genomic regions of functional relevance 109
3.4.4. Local recombination rate 111
3.4.5. Local chromatin accessibility 112
3.4.6. Study overview and concluding remarks 113
4. RTEs as potential variants of disease 115
4.1. Introduction 115
4.1.1. The missing heritability of complex traits: 115
4.1.2. Structural variants and complex traits: 117
4.1.3. RTE insertions as SV: 117
4.1.4. RTE-mediated SVs and Complex traits: 119
4.1.5. Study overview: 120
4.2. Methods 121
4.2.1. Overview of methods: 121
4.2.2. Datasets: 122
4.2.2.1. Polymorphic retrotransposable element (RTE) insertions: 122
4.2.2.2. Trait associated SNPs: 123
4.2.2.3. SNP and RTE genotype files from the 1000 genome project: 123
4.2.2.4. Functional regions file: 125
4.2.3. Method of data analyses 125
4.2.3.1. Overlapping RTEs with TAS LD-blocks 125
4.2.3.2. Enrichment of RTE variants in GWAS risk loci 126

4.2.3.3. LD between RTEs and TAS: 127
4.3. Results 128
4.3.1. TAS Linkage disequilibrium blocks (LD-blocks): 128
4.3.2. Overlapping RTEs with TAS LD-blocks: 129
4.3.3. Enrichment of RTEs in TAS LD-blocks: 130
4.3.4. LD between RTEs and TAS: 133
4.3.5. The distribution of RTEs in LD with TAS in functional genomic regions: 136
4.4. Discussion 140
4.4.1. RTEs overlap and enrichment in GWAS risk loci 140
4.4.2. LD analysis in comparison to previous reports in the literature 141
4.4.3. Distribution of RTEs in LD with TAS: 142
4.4.4. RTEs as causative variants affecting gene expression: 143
4.4.5. Study overview and concluding remarks 144
5. General discussion and future directions 145
5.1. Database curation 146
5.1.1. Reflections and limitations 148
5.2. Genomic distribution of RTEs 149
5.2.1. Study limitations 152
5.3. Correlation between RTEs and TASs 153
5.3.1. Study limitations 155
5.4. Summary of key findings 156
5.5. Future directions 157
References 159
Appendix 1: Additional information about the 45 studies included in the curated database of this study 182
Appendix 2: Full list of RTEs in LD with TAS 184
Appendix 3: RTE-TAS associations in the literature compared to the results of this study. 194

List of Tables:

Table 1: Summary of the inclusion and exclusion criteria for curating a comprehensive database of retrotransposable element insertions that contribute to the inter-individual genomic diversity.....28

Table 2: Search terms used for extract studies identifying retrotransposable element insertions in the human genome from PubMed…….……..29

Table 3: Principle investigators queried in PubMed to extract potentially non-indexed recent studies identifying retrotransposable element insertions in the human genome……..…..30

Table 4: Studies included in the curated non-reference retrotransposable element (RTE) databases, including the database by individual (db-individual), along with the total number of samples and RTEs pre- and post- quality control (QC) steps…….……....37

Table 5: Counts of retrotransposable elements (RTE) curated from 45 studies……..….…39

Table 6: Counts of duplicate RTE profiles in the database-by-individual (db-individual)…….……..……..….41

Table 7: Count and frequency of singleton RTE insertions identified in each of the ethnic groups within the curated database……...44

Table 8: Summary of the genomic distribution of old RTEs that are fixed in the human genome, in comparison to the distribution of polymorphic RTE elements……..…….....57

Table 9: Count of Reference and Non-Reference LINE 1, Alu, and SVAs per chromosome displaying percentage of the whole genome taken from the curated RTE databases……..……72

Table 10: Counts and percentages of RTEs located within intergenic and intragenic regions…….87

Table 11: Counts and percentages of RTEs located within enhancer regions of the GeneHancer database…….…....87

Table 12: Z-test statistics for the distribution of RTEs located within intergenic and intragenic regions of the NCBI RefSeq genes in comparison with a random database including 1,000x iterations…….……88

Table 13: Counts and percentages of RTEs located within recombination regions using the standardised sex-averaged (female and male) recombination map described by Kong et al., (2010)…….…….93

Table 14: Counts and percentages of RTEs located within recombination regions using the standardised sex-averaged (female and male) recombination map described by Kong et al., (2010)…….….....94

Table 15: Empirical P values for the enrichment of RTEs located within euchromatin domains using the H3K4me3 profiles of the Roadmap project. RTEs are polymorphic insertions curated in-house from published studies……100

Table 26: The overlap between polymorphic RTEs and the LD-blocks of genome-wide significance ($P \leq 5x10^{-8}$) TASs. LD blocks were generated using tagging SNPs ($r^2 \geq 0.8$) in PLINK…….....129

Table 17: Polymorphic RTEs in LD ($r^2 \geq 0.8$) with GWAS TAS……..134

Table 18: A list of the overlap between gene regions and RTEs in LD ($r^2 > 0.6$) with TAS……..137

Table 19: List of RTEs in LD with TAS that overlap with enhancer regions……..…139

Table 20: Comparing the list of polymorphic RTEs in LD ($r^2 > 0.6$) with GWAS TAS identified by previous similar studies with the list of RTE-TAS associations identified by the current study……..142

Table 21: Studies included in the curated non-reference retrotransposable element (RTE) databases, including name of the detection tool used by each study, information about sensitivity/validation, and source file from which the insertions in the database got extracted from…….……..…159

Table 22: Full list of RTEs in LD with TAS……..…161

Table 23: List of RTEs in LD ($r^2 > 0.6$) with TAS identified by previous studies in the literature and how it compares to RTE-TAS associations identified by this study…….....171

List of Figures:

Figure 1: Components of the human genome. 2

Figure 2: Schematic representation for the structural organization of (A) LINE1, (B) Alu, and (C) SVA elements and multiple sequence alignment from various subfamilies of each element. …….....7

Figure 3: L1 retrotransposition cycle…….....8

Figure 4: Schematic representation of polymorphic retrotransposable elements (RTE) detection using next-generation sequencing data. …….....24

Figure 5: Overview of study selection process for curating a comprehensive database of retrotransposable elements (RTE) insertions…….....32

Figure 6: An overview of the procedure applied for selecting one retrotransposable element (RTE) profile for inclusion in the database by individual when duplicate RTE profiles for the same individual (individual x) were produced by two studies (study A and B)…….…….…….....36

Figure 7: Genomic Distribution workflow…….……62

Figure 8: Schematic representation of RTEs enrichment in functional and accessible chromatin domains……...……..70

Figure 9: Scatter plots of the distribution of L1 elements across chromosomes and gene density…….……75

Figure 10: Scatter plots of the distribution of Alu elements across chromosomes and gene density……..……77

Figure 11: Scatter plots of the distribution of SVA elements across chromosomes and gene density…….…….....79

Figure 12: Frequency distribution of L1 elements in different GC fractions of the human genome….……...….84

Figure 13: Frequency distribution of Alu elements in different GC fractions of the human genome……..……..84

Figure 14: Frequency distribution of SVA elements in different GC fractions of the human genome…….……...….85

Figure 15: Frequency distribution (%) of RTEs located within intergenic and intragenic regions……....91

Figure 16: Frequency distribution (%) of RTEs located within enhancer regions of the GeneHancer database…….....91

Figure 17: Frequency distribution (%) of RTEs located within different recombination regions using the standardised sex-averaged (female and male) recombination map described by Kong et al., (2010)……..…..93

Figure 18: Count of non-reference (polymorphic) RTEs enriched in the euchromatin domains of at least one epigenome reported in the Roadmap project……....99

Figure 19: Creating LD-blocks around each trait associated SNP (TAS) using tagging SNPs with an r2 ≥ 0.8……...…...…126

Figure 20: Density plot for the enrichment of polymorphic RTEs at GWAS risk loci……....131

List of Abbreviations:

1KGP: The 1,000 genomes project

AF: Allele frequency

ALS: Amyotrophic lateral sclerosis

ALS: Amyotrophic lateral sclerosis

Bp: Base pair

cDNA: Complementary DNA

CNV: Copy number variation

dbRIP: Database of Retrotransposon Insertion Polymorphisms

DPs: Discordant read pairs

DSBs: Double-strand breaks

DXP: X-linked Dystonia-Parkinsonism

EID: Epigenome identifier

EN: Endonuclease

eQTL: Expression quantitative trait loci

ERVs: Endogenous retroviruses

euL1db: The European database of L1HS retrotransposon insertions in humans

FL-L1: Full-length L1

GC-rich: DNA segments rich in guanine and cytosine nucleotides

GoNL: The Genome of the Netherlands project

GRC: Genome Reference Consortium

GWAS: Genome-wide association studies

GWS: Genome-wide significant

HERVs: Human Endogenous retroviruses

HLA: Human leukocyte antigen

KAP1: KRAB-associated protein 1

Kb: Kilo base, i.e. 1,000 base pairs

K-S test: Kolmogorov–Smirnov test

Kya: Thousand years ago

L1Hs: Human-specific L1

LD: Linkage disequilibrium

LINE: Long interspersed nuclear element

LncRNA: Long non-coding RNA

LTR: Long terminal repeats

MAF: Minor allele frequency

MaLR: Mammalian apparent long terminal repeats

MaLR: Mammalian apparent LTR

Mb: Mega base i.e. 1,000,000 base pairs

MiRNA: Micro RNA

MRNA: Messenger RNA

NAHR: Non-allelic homologous recombination

NCBI: The National Centre for Biotechnology Information

NGS: Next-generation sequencing

NHGRI-EBI: National Human Genome Research Institute (NHGRI) & European Bioinformatics Institute (EBI).

ORF: Open reading frame

ORFp: Open reading frame protein

PCR: Polymerase chain reaction

PiRNA: Piwi-interacting RNA

PIs: Principal investigators

PMID: PubMed id

Poly(A) tail: Adenine-rich tail

QC: Quality control

RefSeq: The reference Sequence collection

RNP: Ribonucleoprotein

RT: Reverse transcriptase

RTE: Retrotransposable element

SINE: short interspersed nuclear elements

SiRNA: Small interfering RNA

SNP: single nucleotide polymorphism

SNV: Single nucleotide variant

SR: Split sequencing reads

SRP: Signal recognition particle

SRR: Standardised recombination rate

SV: structural variant

SVA: SINE-VNTR-Alu

TAS: Trait associated SNP

TE: Transposable elements

TF: Transcription factor

TPRT: Target-primed reverse transcription

UCSC: University of California, Santa Cruz

UTR: Untranslated region

VNTR: Variable number of tandem repeat

WGS: Whole-genome sequencing

XDP: X-linked Dystonia-Parkinsonism

1. Introduction

1.1. Transposable elements (TEs):

Transposable elements (TE) are a common genomic feature in the genomes of many organisms, including prokaryotic and eukaryotic organisms (Kleckner, 1981; Bowen and Jordan, 2002; Touchon and Rocha, 2007). They are genomic segments capable of relocating their position in the genome through one of two mechanisms depending on their class (Bire and Rouleux-Bonnin, 2012). Although the evolutionary origin of TEs in eukaryotes remain murky, numerous studies has suggested horizontal transfer as a common occurrence involved in the evolutionary history of all major TE class (Smit, 1996; Bourque et al., 2018; Zhang et al., 2020).

There are two major classes of TEs that are known to exist within the human genome: DNA transposons and retrotransposable elements (Lander et al., 2001). DNA transposons mobilise through a cut-and-paste mechanism that results in the parent element inserting elsewhere in the genome so that it is no longer present in its original genomic location (Muñoz-López and García-Pérez, 2010). In contrast, retrotransposable elements (RTEs) mobilise through a copy-and-paste mechanism via an RNA intermediate resulting in two elements: the parent element at the original genomic location and a new copy elsewhere in the genome (Boeke et al., 1985; Viollet et al., 2014).

The initial sequencing of the human genome has revealed that at least 45% of the genome is composed of TEs (Figure 1), with potentially more of the human genome owing to sequences derived from TE activity (Lander et al., 2001).

The majority of TEs in the human genome are remnants of ancient TE retrotransposition events that amplified in the genome throughout the evolutionary lineage of modern humans (Boissinot et al., 2004; Ewing and Kazazian, 2010).

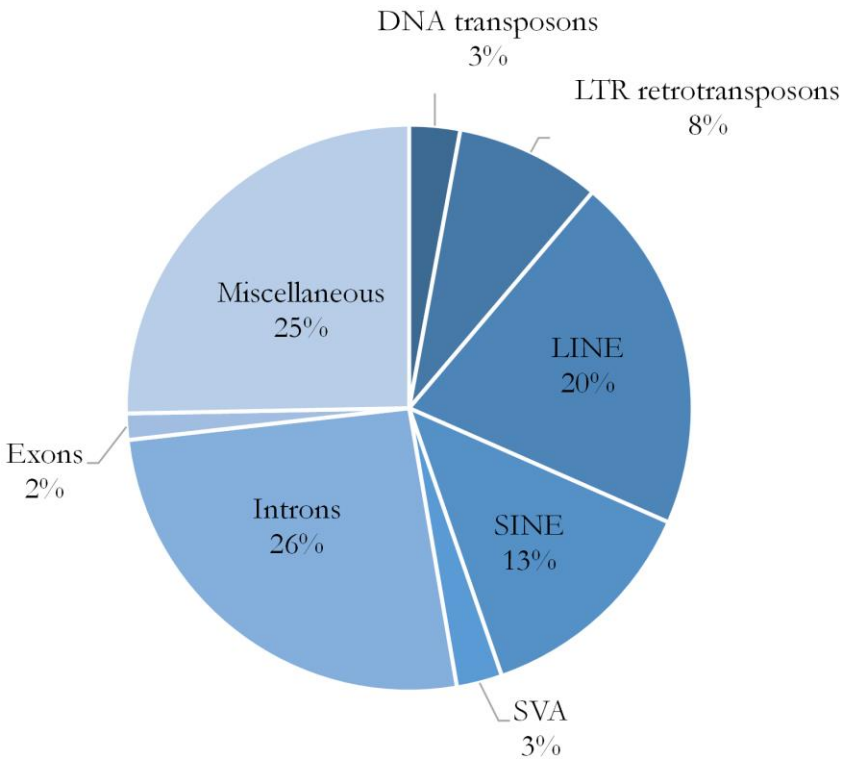


Figure 1: Components of the human genome. Over 45% of the human genome is derived from transposable elements, the majority of which are from the non-LTR (long terminal repeat) retrotransposons including LINEs, SINEs, and SVA elements. Figure adapted from “Synergy between sequence and size in Large-scale genomics” by Gregory, 2005, Nature Reviews Genetics, 6(9), p.702; “Transposable elements and psychiatric disorders” by Guffanti et al., 2014, Am J Med Genet B Neuropsychiatr Genet, 165B(3), p. 203.

Ancient TEs are no longer capable of mobilising in the human genome due to the build-up of random inactivating mutations or internal rearrangements (Lander et al., 2001; Wei et al., 2001; Hancks and Kazazian, 2016). DNA transposon activity has reportedly been non-existent in the human genome for the past 50 million years (Lander et al., 2001). Still, several RTE subfamilies retain their ability to

transpose, thus creating insertional polymorphisms in human populations (Wang et al., 2005; Mills et al., 2007; Huang et al., 2010).

RTE insertional polymorphisms can interfere with the function of the human genome (Kazazian, 2004; Hancks and Kazazian, 2016; Bourque et al., 2018), yet the extent to which RTE polymorphisms contribute to human health and disease has not yet been fully explored. This study investigates the effect of relatively recent RTE insertions from the active RTE subfamilies on genome function and how RTE variants from these subfamilies may influence predisposition to complex diseases. It is first necessary to review some of the relevant features of retrotransposable .elements found within the human genome before describing the analyses of this study.

1. 2. Retrotransposable elements (RTEs)

1.2.1. Classes of retrotransposable elements (RTEs)

RTEs are classified depending on the presence or absence of flanking long terminal repeats (LTR). LTR retrotransposons include mammalian apparent LTR (MaLR) and three classes of endogenous retroviruses (ERVs): ERV-class I, ERV (K)-class II, and ERV (L)-class III (Lander et al., 2001). The non-LTR retrotransposons include long interspersed nuclear elements (LINE), short interspersed nuclear elements (SINE) encompassing Alu elements, and SINE-VNTR-Alu (SVA) elements (Lander et al., 2001; Cordaux and Batzer, 2009).

Each RTE subclass is further organized into families and subfamilies, reflecting the evolution of RTE propagation in the human genome (Lander et al., 2001; Cordaux and Batzer, 2009). RTE subfamilies are distinguishable by sequence

variations from the consensus sequence shared by all members of the same RTE subclass (Ovchinnikov et al., 2002; Price et al., 2004; Wang et al., 2005). RTEs of the non-LTR class are the most abundant in humans, which together comprise over a third of the human genome (Figure 1) (Lander et al., 2001; Cordaux and Batzer, 2009). The non-LTR retrotransposons are the main focus of this study, which is why a more detailed description of these RTEs is discussed below.

1.2.2. Structural organisation of non-LTR RTEs

1.2.2.1. LINE

Three LINE families exist in the human genome: LINE1, LINE2, and LINE3 (Lander et al., 2001). The LINE1 (or L1 for short) family is the most recently evolved LINE family and the most abundant TE type in the human genome comprising 16.9% of the genome (Lander et al., 2001). A typical LINE is about 6 kilobases (kb) long and consists of: an internal promoter for RNA polymerase II located in its 5' untranslated region (UTR), two open reading frames (ORF1 and ORF2), separated by 63 base pairs (bp), a 3' UTR and an adenine-rich tail i.e. poly(A) tail (Dombroski et al., 1991) (Figure 2A).

Members of the human-specific L1 subfamily (L1Hs) are the only type of LINE elements that remain active in the human genome (Kazazian et al., 1988). L1s are described as autonomous elements due to their ability to code for the proteins required for their mobilisation in the genome (Moran et al., 1996). SINEs and SVA elements, on the other hand, do not encode any proteins. Instead, they hijack the L1 machinery for their own retrotransposition (i.e. the mechanism of mobilisation), thus they are non-autonomous elements (Dewannieux et al., 2003; Ostertag et al., 2003).

1.2.2.2. SINE: Alu

Alus are the most abundant RTEs in terms of copy number in the human genome, with over 1 million copies that accumulated in the genome from continuous retrotransposition over the past 80 million years of human evolution (Lander et al., 2001). Alu elements diverged from the 7SL RNA gene (Ullu and Tschudi, 1984). Three Alu families exist in the human genome, AluJ, AluS, and AluY, with AluJ showing the most sequence similarity to the 7SL RNA gene and AluY being the most diverged, representing the evolutionary age of the Alu families (Jurka and Smith, 1988; Price et al., 2004). Members of the AluY family, as well as a few subfamilies of AluS, remain active in the human genome (Bennett et al., 2008). A typical Alu element is about 300 bps and consists of two homologous monomers rich in guanine and cytosine nucleotides (GC-rich) and separated by an adenine-rich (A-rich) linker region. The 5' monomer contains A and B boxes, representing the internal promoter derived from the 7SL RNA gene for RNA polymerase III, and the 3' monomer is followed by a poly(A) tail (Figure 2B) (Fuhrman et al., 1981; Dewannieux et al., 2003).

1.2.2.3. SVA

SVA composite elements are the evolutionary youngest TE type in the human genome that potentially evolved during the divergence between humans and the other great apes (Orangutans, Gorillas, and Chimpanzees) about 15 million years ago (Wang et al., 2005). SVAs are also the most recently discovered TE type identified in 1994 by Shen et al. during their investigation of the RP gene structure (Shen et al., 1994). Sequence divergence analysis revealed the existence of 6 SVA subfamilies in the human genome named alphabetically from oldest

(SVA_A) to the most recently evolved (SVA_F) (Wang et al., 2005). SVAs from the E and F subfamilies are human-specific and remain active in the human genome (Wang et al., 2005). A typical SVA element is about 2kb long and is composed of five components: a 5' (CCCTCT)n hexamer tandem repeat region, a region homologous to Alu elements (i.e. an Alu-like region consisting of two antisense Alu fragments), a variable number of tandem repeat (VNTR) region made up of between 35 and 50 bp repeats, a SINE region about 490 bp long derived from the human endogenous retrovirus (HERV)-K10, and a poly(A) tail (Figure 2C) (Shen et al., 1994; Ostertag et al., 2003). Unlike L1 and Alu elements, an internal promoter region for SVA elements has not been identified and it has been suggested that these elements rely on the promoter activity of the flanking genomic regions (Wang et al., 2005). Wang et al. (2005) suggested that SVA elements are potentially transcribed by RNA polymerase II, similar to L1 elements, which was later confirmed in the literature via cell culture retrotransposition reporter assays (Hancks et al., 2011; Raiz et al., 2012).

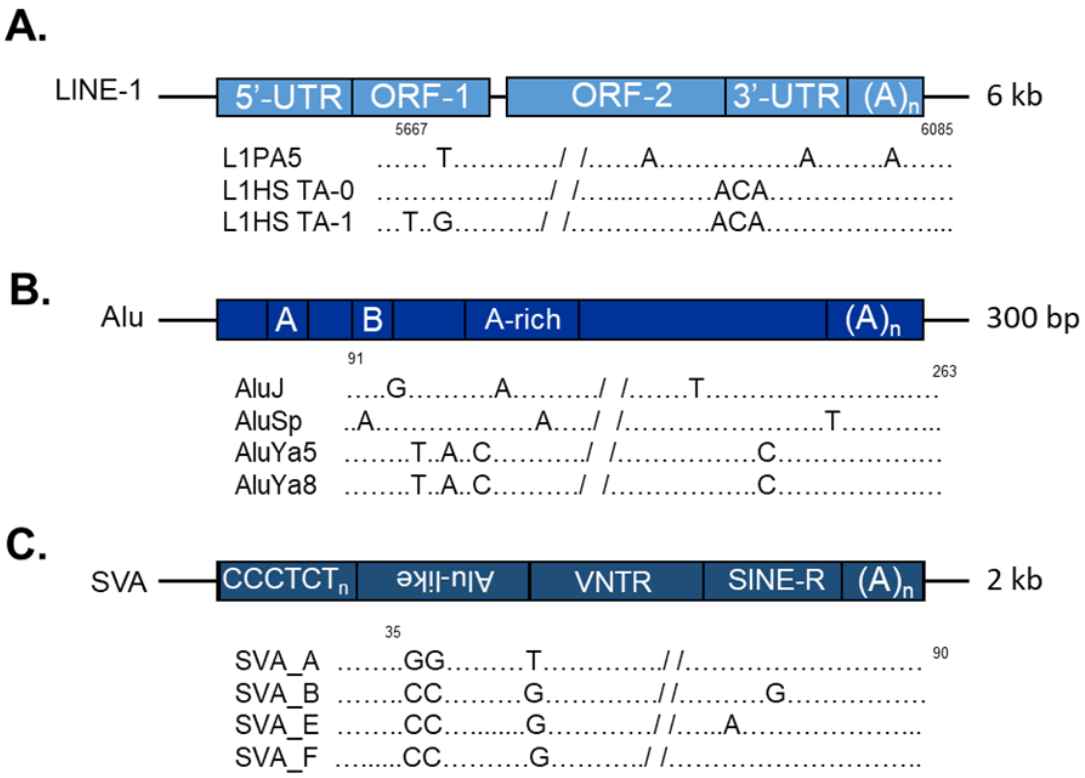


Figure 2: Schematic representation for the structural organization of (A) LINE1, (B) Alu, and (C) SVA elements and multiple sequence alignment from various subfamilies of each element. Sequence alignments are shown below each element, beginning with the oldest to the youngest subfamily that remain active in the human genome. The partial alignment shows the positions of some of the nucleotides that are diagnostic of each subfamily. A| A typical full-length long interspersed nuclear element-1 (LINE1 or L1) is about 6 kilobases (kb) long and consists of a 5' untranslated region (UTR), open reading frame 1 (ORF-1), and 2 (ORF-2) encoding the proteins required for L1 transposition, a 3' UTR, and ends with a polyadenylation tail (A)n. B| The structural organisation of a full-length Alu element. Alus are typically 300 base pairs (bp) long and consist of two homologous monomers separated by an adenine-rich (A-rich) linker region. The 5' monomer contains an internal promoter for RNA polymerase III (A and B boxes), while the 3' monomer ends with a polyadenylation tail (A)n. C| SVAs are composite elements consisting of 5 units: a 5' (CCCTCT)n hexamer tandem repeat region, an Alu-like region consisting of two antisense Alu fragments, a variable number of tandem repeat (VNTR), a SINE region derived from the endogenous HERV-K10 retrovirus, and a 3' polyadenylation tail (A)n. Figure adapted from "The insertional history of an active family of L1 retrotransposons in humans" by Boissinot et al., 2004, Genome Research, 14(7), p.1222; "Standardized nomenclature for Alu repeats" by Batzer et al., 1996, Journal of Molecular Evolution, 42(1), p.4; "Mobile DNA elements in the generation of diversity and complexity in the brain" by Erwin et al., 2014, Nature Reviews Neuroscience, 15(8), p.498; "Identification of polymorphic SVA retrotransposons using a mobile element scanning method for SVA (ME-Scan-SVA)" by Ha et al., 2016, Mobile DNA, 7(1), p.2.

1.2.3. Mechanism of L1 retrotransposition

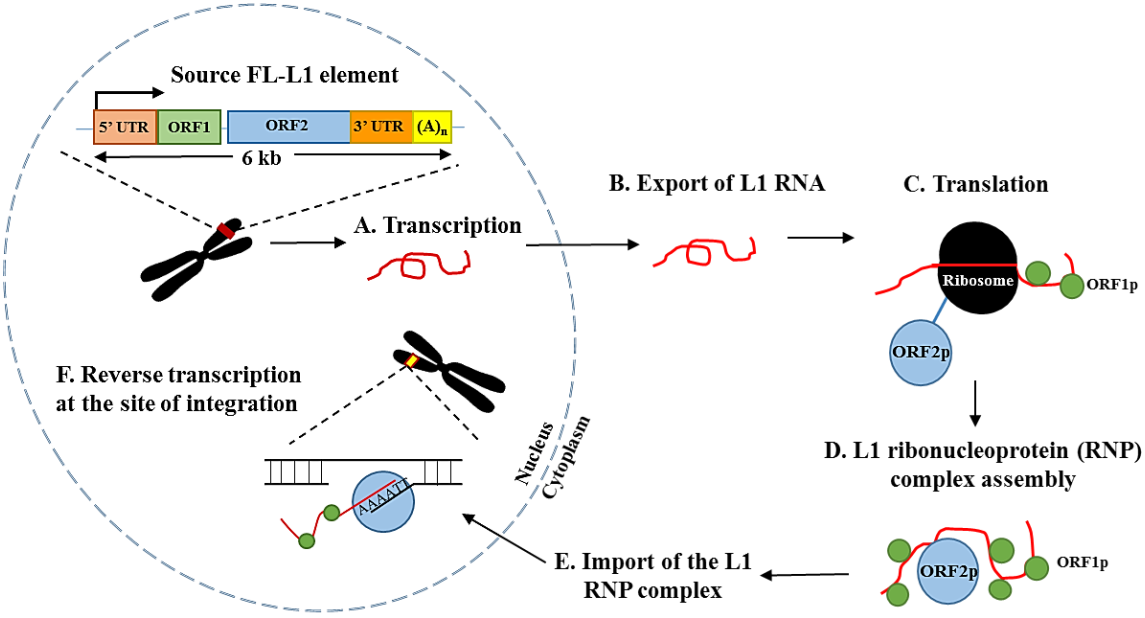


Figure 3: L1 retrotransposition cycle: A| Full- length L1 source element (FL-L1) is transcribed in the nucleus producing an L1 mRNA. B and C| The mRNA is exported into the cytoplasm, where its ORF1 and ORF2 proteins are translated into ORF1p and ORF2p. D| The translated proteins bind to the source L1 mRNA to form an RNP complex, which is imported back into the nucleus (E). F.| A new L1 that is a copy of the source L1 element is inserted into a new genomic location by target-primed reverse transcription. Figure adapted from "Mobile DNA elements in the generation of diversity and complexity in the brain" by Erwin et al., 2014, Nature Reviews Neuroscience, 15(8), p.498.

The L1 retrotransposition cycle begins with transcription of a full-length L1 source element (FL-L1) by RNA polymerase II from its internal promoter, generating a bicistronic (i.e. encoding two proteins) L1 messenger RNA (mRNA) (Figure 3.A). The generated mRNA is exported into the cytoplasm (B), where the two L1-encoded proteins are translated (C). ORF1p is a small protein with RNA-binding properties, and ORF2p is a larger protein, with both endonuclease (EN) and reverse transcriptase (RT) activities.

The source L1 mRNA and translated proteins bind to form a ribonucleoprotein complex (RNP) (D) that is imported into the nucleus (E). The EN domain of ORF2p creates a single-strand nick at the opposite strand of its target site (5'-TTTT/AA-3', where "/" indicates the site of EN cleavage). The cleavage exposes a 3'-OH which is used by the ORF2p-RT domain to prime reverse transcription of the L1 mRNA in the new genomic location (F), starting from its 3' -poly(A) tail end, through a process known as target-primed reverse transcription (TPRT). The outcome of this process is a novel L1 insertion, which is a copy of the original FL-L1 source element, at a second genomic location. The source FL-L1 element is capable of producing further L1 copies, flanked by target site duplications, which are characteristic of the retrotransposition process (Hancks, Kazazian and Jr., 2016; Scott and Devine, 2017).

1.2.4. Alu and SVA elements hijack the L1 machinery

Alu and SVA insertions are typically flanked by target site duplications characteristic of the L1 retrotransposition process and suggest that these elements hijack the L1 machinery since they lack any open reading frames (Dewannieux et al., 2003; Wang et al., 2005). A model describing the process by which Alu elements hijack the L1 machinery was described by Dewannieux et al. (2003). This model explains how a domain of the 7SL gene that is conserved in Alu elements associates with the binding site of the signal recognition particle (SRP), a ribonucleoprotein complex that can bind to specific signal peptides. The Alu-bound SRP then interacts with the ribosome, positioning the Alu transcript in close proximity to the L1 mRNA, thus allowing it to capture the L1 ORF2 protein as it is being translated (Figure 3.C). Alu elements that successfully capture the

L1 ORF2 protein during its translation can then replace the L1 transcript with their own during the TPRT process (Figure 3.F) (Dewannieux et al., 2003). Note that the L1 ORF1 protein is not required for Alu retrotransposition, therefore L1 elements with a non-functional ORF1 gene but a functional ORF2 gene can still facilitate Alu mobilisation.

Compared to Alu elements, the process by which SVA transcripts replace the L1 mRNA during the TPRT process is not very well defined. Reporter assays in cell culture confirmed the requirement for both L1 ORF1 and L1 ORF2 proteins for SVA retrotransposition (Hancks et al., 2011; Raiz et al., 2012). The Alu-like domain of the SVA element has been hypothesised to anneal to the SVA transcript with Alu transcripts bound to the SRP ribonucleoprotein complex, which will potentially allow the SVA transcript to capture the L1 ORF2 protein (Mills et al., 2007). This hypothesis is consistent with the results of Raiz et al. (2012) that reported a decrease in the retrotransposition activity of SVA reporter elements with a deleted Alu-like domain in cell cultures by an average of 32-46%, compared with a full-length reporter element.

1.2.5. Effect of retrotransposition on genome function and integrity

The L1 ORF2 protein activity is an obvious source for the potential negative impact of retrotransposition on genome integrity, as it is capable of inducing DNA breakage, as previously discussed. RTE activity can harm genome function and integrity through a variety of mechanisms (Cordaux and Batzer, 2009; Guffanti et al., 2014; Savage et al., 2019). The effect of a new RTE element on genome function depends on the location in which it is inserted. Insertions into gene regions can interfere with mRNA splicing or even introduce new exons within the

interrupted gene (Lev-Maor et al., 2008; Chénais, 2016). RTE insertions can also cause inactivating mutations through inserting into exonic regions (Kazazian et al., 1988; Hancks and Kazazian, 2016). The poly(A) tail of RTE elements can provide polyadenylation signals that can affect the elongation of gene transcription, either by resulting in premature termination or by reducing transcription, consequently reducing gene expression (Perepelitsa-Belancio and Deininger, 2003; Chen et al., 2009).

RTE insertions upstream of gene regions can also affect gene function in many ways. RTEs can induce local epigenetic modifications which could affect the expression of neighbouring genes (Goodier, 2016). Alu and L1 elements carry internal promoter regions that can modulate gene expression of nearby genes (Nigumann et al., 2002; Zhang et al., 2015). Although SVA elements do not have an internal promoter, they are still able to bind to transcription factors, thus are able to modulate the expression of nearby genes (Quinn and Bubb, 2014; Gianfrancesco et al., 2017).

RTE retrotransposition has the potential to destabilise local genomic stability and mediate post-insertional rearrangements. An *in vitro* study reported that the L1 ORF2 protein created more DNA double-strand breaks (DSBs) than the number associated with successful L1 insertions, suggesting the negative impact of L1 activity on genome stability (Gasior et al., 2006). DSBs can drastically effect genome function. Unrepaired DSBs can potentially result in cell death, while incorrectly repaired DSBs can create chromosomal abnormalities such as translocations (Jeggo and Löbrich, 2007).

RTE integration can also cause deletion of genomic DNA at the integration site (Callinan et al., 2005; Han et al., 2005; Lee et al., 2012). Significant deletions and

duplications, resulting from RTE-mediated homologous and non-allelic homologous recombination events, can occur due to the high copy number and sequence homology between RTE elements (Startek et al., 2015; Nazaryan-petersen et al., 2016). RTE elements are also able to duplicate 5' and 3' flanking genomic regions during their retrotransposition, a phenomenon known as sequence transduction. This occurs when RTE transcription starts or carries on outside the element itself as a result of transcription initiation, using a promoter located upstream of the element or transcription elongation past the polyadenylation signal of the element, resulting in 5'- and 3' transduction, respectively (Goodier et al., 2000; Xing et al., 2009). These and other various effects of RTE activity on genome function are extensively reviewed in the literature (Cordaux and Batzer, 2009; Beck et al., 2011; Hancks and Kazazian, 2012; Guffanti et al., 2014; Bourque et al., 2018; Savage et al., 2019).

1.2.6. Mechanisms of retrotransposition silencing

The human genome evolved many methods that act at every stage of the retrotransposition process to suppress RTE activity due to the many negative impacts associated with its activity on genome function and integrity. Pre-transcriptional silencing methods include epigenetic silencing methods such as: repressive histone modifications, heavy methylation of CpG dinucleotides in the promoter region, and KRAB-associated protein 1 (KAP1) mediated chromatin remodelling (Bestor and Bourc'his, 2004; Garcia-Perez et al., 2010; Jacobs et al., 2014).

Post-transcriptional silencing methods include degradation of RTE transcripts through endonucleases guided by short non-coding RNAs (typically 20-30 bp)

such as small interfering and Piwi-interacting RNAs (siRNA and piRNAs, respectively) (De Fazio et al., 2011; Chen et al., 2012; Goodier, 2016).

Additional host defence mechanisms exist that appear to limit the size of RTE insertions during the final stage of integration (Perepelitsa-Belancio and Deininger, 2003; Coufal et al., 2011). A detailed explanation of the named silencing mechanisms plus additional silencing mechanisms are reviewed in the literature (Goodier, 2016; Hancks and Kazazian, 2016; Yang and Wang, 2016).

1.3. Functional role of TE activity in genome evolution

The persistence of TE activity in the human genome despite their potential to negatively impact genome function and the many mechanisms of retrotransposition silencing poses the question about their role in genome function and evolution. Empirical studies owed the continuity of TE activity throughout the evolutionary history of eukaryotic genomes to the selfish and parasitic properties of these elements, evident by their ability to replicate faster than their host (Doolittle and Sapienza, 1980; Orgel and Crick, 1980; Kidwell and Lisch, 2001). There is now growing evidence supporting the original perspective of Barbara McClintock, the scientist responsible for TE discovery, who suggested that TEs may contribute to gene regulation (McClintock, 1956). TEs are now believed to have provided the source for the evolution of the majority of regulatory elements in the human genome through exaptation, i.e., the process by which TEs are harnessed to provide new functions, thereby facilitating the adaptation of their host to defined selective pressures (Jacques et al., 2013; Su et al., 2014; Chuong et al., 2017; Bourque et al., 2018).

Many mechanisms by which TE-derived sequences contribute towards the regulation of the human genome, both in cis and in trans, have been reported in the literature (Smalheiser and Torvik, 2005; Johnson and Guigó, 2014; Elbarbary et al., 2016; Trizzino et al., 2018). Genome-wide assays have revealed that most TE-derived regulatory elements originate from ancient insertions that are now fixed in the human genome (Lowe and Haussler, 2012; Lynch et al., 2015; Trizzino et al., 2018). TE-derived cis-regulatory sequences include promoters, enhancers, and transcription factor (TF) binding sites that can interact with regulatory elements such as activator and repressive elements (Jordan et al., 2003; Lowe and Haussler, 2012; Su et al., 2014). At least 475 experimentally validated promoters and ~20% of TF binding sites in the human genome contain sequences derived from TEs (Jordan et al., 2003; Sundaram et al., 2014). TEs also provided the source for many micro RNAs (miRNAs) and long non-coding RNAs (lncRNAs) that can act as cis- or trans-regulatory elements with the potential to modulating gene expression or contribute towards post-transcriptional regulation of many genes (Smalheiser and Torvik, 2005; Johnson and Guigó, 2014).

Qin et al. (2015) identified 409 miRNAs derived from TE sequences, while Kelley and Rinn (2012) reported that about 42% of >7,600 human lncRNA sequences are derived from TEs. Other TE-derived regulatory sequences include transcription terminators (Conley and Jordan, 2012) and chromatin looping binding sites (Diehl et al., 2020). In addition to donating regulatory elements, TEs can also source and modulate regulatory networks that control complex biological pathways (Wray et al., 2003; Feschotte, 2008; Sundaram et al., 2014).

Studies have demonstrated the contribution of TE elements in mediating novel regulatory networks in the uterus during the evolution of mammalian pregnancy (Lynch et al., 2011; Lynch et al., 2015). Genome-wide ChIP analysis revealed that about one-third of the p53 regulatory protein binding sites are mediated by primate-specific ERV TEs, suggesting the role of TE elements in mediating species-specific transcriptional networks (Wang et al., 2007). An integrated genome-wide analysis has also confirmed the greater impact of species-specific TEs in mediating gene regulations (Zeng et al., 2018). Taken together, the process by which TEs contribute towards their host adaptation to selective pressures by providing the source for novel gene regulations is still ambiguous. One of the main challenges in achieving this knowledge is working backward, to study events that have already taken place millions of years ago, to understand the specific steps that led to TEs exaptation.

1.4. RTEs as structural variants

1.3.1. RTE variants and their detection in the human genome

Despite the many genomic mechanisms to suppress RTE activity, some active elements escape such that individual genomes acquire additional RTE copies (Kazazian et al., 1988; Wang et al., 2005; Bennett et al., 2008; Guffanti et al., 2014; Hancks and Kazazian, 2016). When this occurs in tissues, the active RTEs create intra-individual somatic variations, which are not inherited by future generations (Reilly et al., 2013; Scott et al., 2017; Faulkner and Billon, 2018). In contrast, active RTEs that escape genome suppression in the germline create inter-individual variations that are inherited by future generations (Huang et al.,

2010; Akagi et al., 2013). Such insertions can create structural polymorphisms within and between populations depending on the time of their integration with respect to human evolution (Rishishwar et al., 2015). Initial studies investigating the rate of new RTE insertions per generation estimated that 1 new L1, Alu, and SVA elements occur in every 20-270, 20, and 900 births, respectively (Cordaux et al., 2006; Xing et al., 2009; Ewing and Kazazian, 2010; Beck et al., 2011). More recent pedigree-based estimates for the rate of new RTE elements are 1 new L1 and SVA elements for every 63 births, and 1 new Alu element per 40 births (Feusier et al., 2019). These new estimates suggest that SVA elements are more active while Alu elements are not as active in the human genome as previously thought. RTEs are often neglected in genomic studies despite their ongoing contribution to creating structural variations in humans, because of their repetitive nature and high sequence homology, which makes them difficult to detect and study (Rishishwar et al., 2017; Bourque et al., 2018). Detection of RTE variants from next-generation sequencing (NGS) data requires specialised computational tools (Ewing, 2015; Goerner-potvin and Bourque, 2018). Currently, the most accessible method of RTE detection rely on short-read NGS data and computational detection tools such as MELT (Gardner et al., 2017) and Mobster (Thung et al., 2014). These tools are designed with the consideration of the uniqueness of short-read NGS data alignment to the reference human genome and to RTE consensus sequences (Goerner-potvin and Bourque, 2018). However, different tools analysing the same sample produce varying results, despite the similarity in the fundamental algorithmic design (Rishishwar et al., 2016). Consequently, the full scope of structural variations mediated by RTE insertions is still under-represented in RTE detection studies.

Long-reads from sequencing technologies such as PacBio (Pacific Biosciences; Rhoads and Au, 2015) and MinION (Oxford Nanopore; Lu et al., 2016) are more likely to span the entire length of an RTE insertion plus their flanking genomic sequences. As such, the increasing length of NGS reads plus the continuous improvement of RTE detection tools is likely to resolve the issues of accuracy and precision of RTE detection within the human genome.

1.4.2. The effect of RTE variants on human health and disease

RTE elements that are fixed in the human genome tend to accumulate in non-functional genomic regions (Lander et al., 2001) yet *de novo* germline RTE insertions do insert in functional regions. Such insertions are responsible for ~1/1000 disease-causing mutations (Lutz et al., 2003), and about 124 germline RTE-mediated insertional mutations that are known to cause monogenic diseases are reported in the literature (Hancks and Kazazian, 2016). The lack of RTE fixation in coding regions, despite their ability to integrate into functional regions, supports the role of natural selection in shaping the genomic landscape of ancient insertions (Medstrand et al., 2002; Abrusán and Krambeck, 2006; Kvikstad and Makova, 2010; Zhang et al., 2011).

In contrast, the landscape of recent RTE insertions is unlikely shaped by selection, a slow process that takes place over many generations (Huang et al., 2010). Conflicting results about the genomic distribution of recent RTE insertions have been reported in the literature (Ovchinnikov et al., 2001; Medstrand et al., 2002; Boissinot et al., 2004; Beck et al., 2010). The integration of active RTEs into coding regions suggests the negative impact of retrotransposition on genome function and the potential implication of recent insertions with regards to human health and disease. Indeed, recent studies have shown that somatic RTE

insertions in epithelial cells frequently drive tumorigenesis, so much so that it is now believed to be a hallmark feature of epithelial cancers in humans (Shukla et al., 2013; Scott et al., 2016; Zampella et al., 2016). Moreover, accumulating evidence supporting the role of somatic retrotransposition in neuronal plasticity and possibly in neuropsychiatric disorders has been published (Baillie et al., 2011; Erwin et al., 2014; Upton et al., 2015; Doyle et al., 2017). Structural variants derived from germline RTE insertions are also likely contributors towards phenotypic variations and individual predisposition to complex disorders. RTE variants are often in linkage disequilibrium with nearby SNPs (Higashino et al., 2014; Kuhn et al., 2014), and they can modulate the expression of nearby genes (Wang et al., 2017; Spirito et al., 2019). This suggests that they may cause differential gene expression between individuals in the population. Nevertheless, the relationship between germline RTE variants and predisposition to complex disorders remains ambiguous.

1.5. Summary of current study

Active RTE elements are an ongoing source of threat to the human genome given their disruptive nature and ability to impact genome regulation, yet the full scope of their integration, genomic distribution, and contribution towards complex disorders remains ambiguous. A comprehensive database of polymorphic RTE insertions was curated using online RTE databases and peer-reviewed publications in the literature. The curated database was used to investigate the genomic distribution of polymorphic RTEs by comparing it with the distribution of fixed RTE elements. The distribution analysis revealed the extent of RTE retrotransposition impact on genome function. Polymorphic RTEs in linkage disequilibrium with SNPs significantly associated with risk of various complex

traits and disorders were identified, providing a list of putative causative variants in risk loci.

1.6. Aims and Objectives

1.6.1. Research Aims:

Several RTE elements retain their ability to transpose in the human genome, thereby creating new insertions that contribute to the genomic diversity of humans. Yet, the extent of the continuous effect of RTE activity on genome function and the potential contribution of recent insertional polymorphisms as causative variants of diseases remains an open question. This thesis aimed to investigate the impact of RTE activity on genome function and the potential association of RTE variants with disease susceptibility. To this end, the landscape of recent RTE insertions and their positional overlap with trait-associated SNPs has been investigated to establish whether RTE activity may pose a high risk to genome function and contribute to human health and disease.

1.6.2. Main Objectives:

1. Curate a comprehensive database of known RTE variants from the active L1, Alu, and SVA subfamilies using publicly available online databases plus peer-reviewed journal articles.

2. Use the curated database to investigate the potential effect of RTE activity on genome function by comparing the genomic distribution of polymorphic RTE insertions against the distribution of endogenous RTE elements fixed in the human genome.

3. Investigate the enrichment of polymorphic RTE elements in GWAS risk loci to establish the potential contribution of polymorphic RTEs as causative variants of complex diseases.

4. Curate a list of RTE variants in linkage disequilibrium with trait-associated SNPs that could potentially be causative variants of complex disease.

2. Database curation

2.1. Introduction:

The average human genome is composed of 17% L1s, 11% Alus, and 0.13% SVA elements, the majority of which are remnants of ancient retrotransposition events that took place throughout the evolutionary lineage of modern humans (Lander et al., 2001; Gregory, 2005; Quinn and Bubb, 2014). RTE insertions of ancient subfamilies are now fixed in the genome of individuals from all races i.e. they are present in the same location in the genome of all individuals in all populations (Lander et al., 2001; Ewing and Kazazian, 2010). These ancient RTE subfamilies are no longer capable of retrotransposition due to 5' and 3' truncations, build-up of random inactivating mutations, or internal rearrangements (Lander et al., 2001; Wei et al., 2001; Hancks and Kazazian, 2016). Only the youngest, most recently evolved RTE subfamilies remain capable of active retrotransposition in humans including elements from the L1Hs (for human-specific), AluY, and SVA_E/F subfamilies (Myers et al., 2002; Wang et al., 2005; Feusier et al., 2017). The evolutionary young and active RTE subfamilies are distinguishable from ancient subfamilies by sequence variations that deviate from the consensus sequence (Myers et al., 2002; Raiz et al.,2012;

Witherspoon et al., 2013; Konkel et al., 2015) as previously shown on Figure 2 (Chapter 1, page 7).

The average human genome is estimated to carry between 80-100 L1Hs elements (Brouha et al., 2003), 852 Alus (Bennett et al., 2008), and 56 SVAs (Bennett et al., 2004) that are actively transposing. Retrotransposition events of the active RTE subfamilies create insertional polymorphisms within and between populations. As such, an insertion within a defined genomic location can be either present or absent in the genome of two unrelated individuals within a population. Such insertions are not part of the reference genome assembly. These polymorphic RTEs contribute to the genetic diversity of the human genome, and have the potential to influence host susceptibility to disease depending on their genomic location (Bourque et al., 2018; Gardner et al., 2019; Hormozdiari et al., 2019).

2.1.1. Early methods of RTE discovery:

Characterising polymorphic RTE insertions in the human genome is troublesome due to their repetitive nature and high sequence homology with ancient sub-families. PCR display was one of the early methods of polymorphic RTE detection in the human genome (Sheen et al., 2000; Ovchinnikov et al., 2001). It involved two successive PCR experiments: the first experiment being amplified using a set of RTE-specific and non-specific arbitrary primers, and the second reaction amplified with a nested primer characteristic of the active RTE subfamily plus the same non-specific primer of the first experiment. The nested PCR experiment was then followed by southern blot hybridisation using an oligonucleotide probe that is also complementary to the diagnostic sequences of the active subfamily. The chromosomal location of polymorphic insertions were then characterised as

the insertions that are only visible in few of the tested individuals (Sheen et al., 2000; Ovchinnikov et al., 2001). A second example of the early efforts to characterise polymorphic RTE detection involved DNA shredding/fragmentation and PCR amplification with primers complementary to the diagnostic sequences of the active RTE subfamily (Boissinot et al., 2004; Mamedov et al., 2005). The PCR products were then cloned into vectors that were grown in bacterial cultures followed by a number of steps that were designed to identify clones holding polymorphic insertions. Clones that were long enough to contain a polymorphic insertion plus flanking genomic sequences were eventually sequenced via Sanger sequencing and the genomic location of the sequenced insertions were finally identified in the public databases using BLAST and BLAT (Boissinot et al., 2004; Mamedov et al., 2005). Consequently, previous studies characterizing polymorphic RTE insertions using these labour-intensive methods were limited by the number of samples and in the number of insertions they were able to characterise (Sheen et al., 2000; Ovchinnikov et al., 2001; Myers et al., 2002).

2.1.2. Current methods of RTE discovery:

Advances in high throughput next-generation sequencing (NGS) technologies and the development of computational detection tools facilitated their genome-wide detection. Genome-wide detection of polymorphic RTE insertions via short-read NGS typically involves the use of targeted enrichment of the active RTE elements in the sequencing libraries. PCR-based amplification (David et al., 2015; Streva et al., 2015; Ha et al., 2016) and hybridization-based capture (Baillie et al., 2011; Shukla et al., 2013; Upton et al., 2015) are the most commonly used enrichment techniques. These pre-sequencing enrichment techniques use

specific primers and oligonucleotide probes to target specific DNA sequences that are characteristic of the active RTE subfamilies. The enriched libraries are then sequenced typically using a short-read NGS platform (~150-200bp) such as illumina (Quail et al., 2012; Rishishwar et al., 2016). Reads passing the quality filtering criteria are then mapped to the reference human genome using a sequencing alignment tool such as BWA-MEM (Li and Durbin, 2009) or SOAP2 (Li et al., 2009). Putative insertion sites are then called using various computational tools that are specifically designed for the detection of polymorphic RTE insertions from short-read NGS data (Baillie et al., 2011; David et al., 2015; Streva et al., 2015; Gardner et al., 2017). The many polymorphic RTE detection tools available in the literature typically function using the same fundamental method. Essentially, two main types of sequencing reads that point to the presence of a polymorphic insertion: discordant read pairs and split-reads.

Paired-end sequencing reads are individual reads produced from both ends of the same DNA segment (Paterson et al., 2015). Such reads can be paired together as the distance between them is known, hence the name. The Paired-end sequencing method increases the accuracy of read mapping, especially in repetitive genomic regions, thus facilitates the detection of structural rearrangements. Discordant read pairs (DPs) where only one read aligns at the expected genomic location facilitate the detection of insertions, such as those produced by RTEs (Paterson et al., 2015; Goerner-potvin and Bourque, 2018). In contrast, split reads (SRs) describe sequencing reads where only a part of an individual read aligns with the genome while its other part does not (Goerner-potvin and Bourque, 2018). Polymorphic RTE detection tools use DPs and SRs, usually referred to as unmapped sequencing reads, to identify polymorphic RTE

insertions by aligning those unmapped reads with the consensus sequence of the active RTE subfamilies (Figure 4; Goerner-potvin and Bourque, 2018).

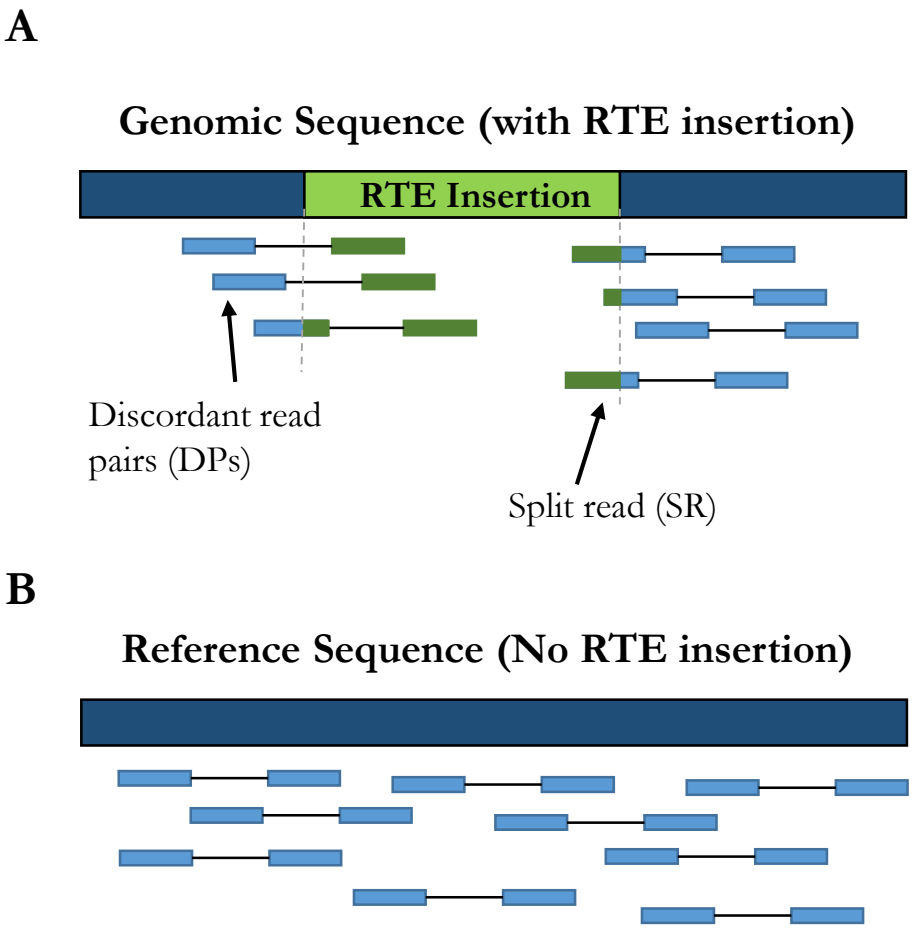


Figure 4: Schematic representation of polymorphic retrotransposable elements (RTE) detection using next-generation sequencing data. Paired-end reads are represented as rectangles connected by a solid line representing the un-sequenced middle region of the DNA segment. A| Genomic region with a polymorphic RTE insertion. Sequencing reads are initially mapped to the reference genome. Unmapped reads that do not partially or fully align with the reference genome as expected are suggestive of an RTE insertion. There are two main types of sequencing reads that are informative for polymorphic RTE identification: 1). Discordant read pairs (DPs) where one read pair aligns with the reference genome while the other align to an RTE sequence, and 2). Split reads (SRs) where only part of one read aligns with the reference genome while the other part aligns to an RTE sequence. B| Genomic region without a polymorphic RTE insertion. Both types of sequencing reads align to the reference genome as expected in the absence of an RTE insertion. Figure adapted from “Benchmarking computational tools for polymorphic transposable element detection” by Rishishwar et al, 2016, Briefings in Bioinformatics, (April).

2.1.3. Limitations of RTE discovery using NGS data:

Characterising and genotyping genomic variants mediated by RTE insertions using short-read NGS data and RTE detection tools, although convenient, remains a challenging task due to the limited length of sequencing reads that does not span the whole integration site. Consequently, the many computational tools developed for RTE detection that follow the same general method often produce unstandardized findings, with different calls being retrieved from the same sequencing sample due to discrepancies in the algorithmic design and choice of parameters (Ewing, 2015). Recent benchmarking studies have demonstrated the utility of combining multiple RTE detection tools in increasing the accuracy and precision of RTE calling (Rishishwar et al., 2016; Nelson et al., 2017).

2.1.4. Online databases for non-reference RTE insertions:

Active RTEs have the potential to implicate human health and disease as their mobilization activity provides a continuous source of structural variations and has the potential to interfere with gene function, epigenetic regulation, and local genomic stability. As such, a comprehensive database of RTE insertions from the active RTE subfamilies is an essential tool for studying RTE-derived SV and for investigating the effects of their continuous activity on genome function. Many studies have been performed for characterising RTE-mediated variants over the last couple of decades. Some of these RTE-derived variants have been collected in online databases, while the majority of them remain scattered in the literature.

Two online databases of polymorphic RTE insertions identified in the human genome exist: the database of retrotransposon insertion polymorphisms (dbRIP;

http://dbrip.brocku.ca/; Wang et al., 2006) and the European database of L1Hs retrotransposon insertions (euL1db; http://eul1db.unice.fr; Mir et al., 2014). At the time of its construction, dbRIP (Wang et al., 2006) queried all the available studies and curated a database of 2,095 unique RTE insertions, including 407 L1s, 1,625 Alus, and 63 SVAs from over 50 studies. At the time of this study, dbRIP (Wang et al., 2006) has not been updated comprehensively since 2009, as only five studies have been added in the second release of dbRIP, bringing the total number of unique insertions up to 2,761, including 598 L1s, 2,086 Alus, and 77 SVAs. Note that the third release of dbRIP, expected in May 2021, was pre-announced on the dbRIP website (https://dbrip.brocku.ca/announcements.html), yet there have been no new file uploads or announcement updates from the date of the pre-announcement up to this date.

The euL1db (Mir et al., 2014) provides a detailed source of recent L1Hs elements, including germline and somatic insertions. It consists of 8,991 non-redundant L1Hs insertions identified in the genomes of 741 individuals that have been curated from 32 studies published up to 2014. The non-redundant list of L1Hs insertions was generated by merging L1 elements located within 200 base pairs (bp) from each other to account for variations in the accuracy of different detection tools and to avoid splitting insertions corresponding to the same retrotransposition event.

2.1.5. Study aims:

Existing online RTE databases are not up-to-date, and thus do not include data from the most recent studies, such as RTE variants from phase 3 of the 1000 genome project (1kGP) (Sudmant et al., 2015). This study aims to update and build on data in existing online databases by curating a comprehensive list of

polymorphic RTE insertions from the L1Hs, AluY, and SVA_E/F subfamilies from peer reviewed journal articles. The updated database of polymorphic RTE insertions will be used for downstream analyses, including inferring the potential impact of RTE activity on genome function and identifying RTE variants that may influence complex traits and predisposition to multifactorial diseases.

2.2. Methods:

RTEs were curated from freely available online databases and peer-reviewed publications in the literature.

2.2.1. Criteria of RTE database curation:

The inclusion and exclusion criteria of the curated database, as summarized in Table 1, were established to ensure the specific inclusion of polymorphic RTE insertions that segregate within the population. Such RTE insertions have the potential to influence complex traits and predisposition to disease. The effect of recent RTE insertions on the fitness of its host depends on where it lands in the genome, thus only studies that report the exact genomic location of recent RTE insertions were selected for inclusion. Studies reporting engineered RTE insertions in transfected cell lines were excluded as such insertions may not be true representatives of insertions that segregate in the population. Similarly, Somatic RTE insertions such as those found in tumour cells were also excluded as such insertions does not get passed on to future generations. Germline insertions were included as these are the only type of insertions that segregate in the population, thus contributes towards inter-individual variation and disease risk.

Table 1: Summary of the inclusion and exclusion criteria for curating a comprehensive database of retrotransposable element insertions that contribute to the inter-individual genomic diversity.

Criteria	Included	Excluded
Genomic coordinates	Exact genomic coordinates reported in any human genome build	Does not report exact genomic coordinates
Tissue type	Fresh samples (e.g. Blood) or non-transfected cell lines.	Transfected cell lines
Tissue state	Healthy or matched non-tumour	Tumour cells
Insertion type	Germline	Somatic

2.2.2. Study selection:

Studies referenced in dbRIP (http://dbrip.brocku.ca/; Wang et al., 2006) and euL1db (http://eul1db.unice.fr; Mir et al., 2014) online databases that were published within the cut-off dates of this study (01/01/2009-11/04/2019) were selected. This cut-off date was decided based on the period all three RTE subtypes were simultaneously updated in dbRIP (http://dbrip.brocku.ca/; Wang et al., 2006). Additional studies were selected via PubMed using the search terms listed in table 2. MeSH terms refer to controlled PubMed vocabulary used for indexing journal articles for easier retrieval of relevant publications via automatic term mapping. The MeSH terms are arranged hierarchically by subject categories with more specific terms beneath the broader term (Ecker and Skelly, 2010). PubMed search results were refined by activating filters to exclude articles published outside the cut-off date. Additional filters were applied to ensure all search results concerned humans and are published in English.

Table 2: Search terms used for extract studies identifying retrotransposable element insertions in the human genome from PubMed.

Search terms	Description
Transposable elements	Two classes: Retrotransposons and DNA transposons. Narrow MeSH term(s) included in the results: 1. Insertions
Retrotransposons	Class I transposable elements that mobilise via an RNA intermediate. Narrow MeSH term(s) included in the results: 1. Endogenous Retroviruses 2. Genes, Intracisternal A-Particle 3. Long Interspersed Nucleotide Elements 4. Short Interspersed Nucleotide Elements 5. Alu Elements
Mobile element polymorphisms/insertions	Synonym of transposable elements and retrotransposons.
Structural Variation	Differences in genomic DNA segments between/within individuals of a population. Includes DNA Copy Number Variations (CNVs).

A second PubMed search was performed using the names of principal investigators (PIs) as keywords to avoid missing out on recent articles that may not have been indexed with the MeSH terms found in table 2 at the time of this study. PIs were identified from the transposable elements labs directory of the Mobile DNA journal (https://mobilednajournal.biomedcentral.com/labs).

PIs interested in RTE detection and retrotransposition in humans were selected (Table 3). Ewing A.D. was selected based on publication record and interest in RTE detection.

Table 3: Principle investigators queried in PubMed to extract potentially non-indexed recent studies identifying retrotransposable element (RTE) insertions in the human genome.

Principal Investigator	Research Interests
Burns K.H.	• The role of transposable elements in human disease. • Characterising human retrotransposon insertion polymorphisms. • LINE-1 regulation and RTE insertions in cancer.
Ewing A.D.	• Development of computational tools for RTE detection. • Inferring the functional consequences of mutations caused by RTE insertions.
Kazazian H.H.	• Detecting and understanding LINE-1 (L1) retrotransposons. • Better the understanding about RTE role in complex human disorders.
Moran J.V.	• Understanding the biology of human L1s. • Understanding host defences against the transposition process.
Jorde L.B.	• Evolution and effect of RTEs on human genome in health and disease.
Xing J.	• Understanding the mechanism and consequences of genomic variations caused by RTE insertions in humans health and disease. • RTEs in evolutionary genetics.

Citations retained using the terms listed in Tables 2 and 3 were collated in a text file. Identical citations retained from multiple search terms were removed based on their PubMed ids using an awk command in UNIX. Titles and abstracts of the

remaining studies were initially manually scanned to exclude articles unrelated to RTE detection/identification. Studies that did not relate to L1s, Alus, or SVAs were excluded, e.g. those relating to DNA transposons and LTR retrotransposons i.e. human endogenous retroviruses (HERV). Review articles, articles about the origin and evolution of RTEs, and articles assessing the effect of chemical (e.g. drugs) or physical (e.g. radiation) agents on RTEs mobilization/epigenetics were also excluded. Publications that passed this initial scanning step underwent thorough manual survey to investigate whether they meet the inclusion criteria of this study as discussed above (Table 1). Articles that meet the inclusion criteria of this study were retained for the database curation. An overview of the study selection processes is shown in figure 5.

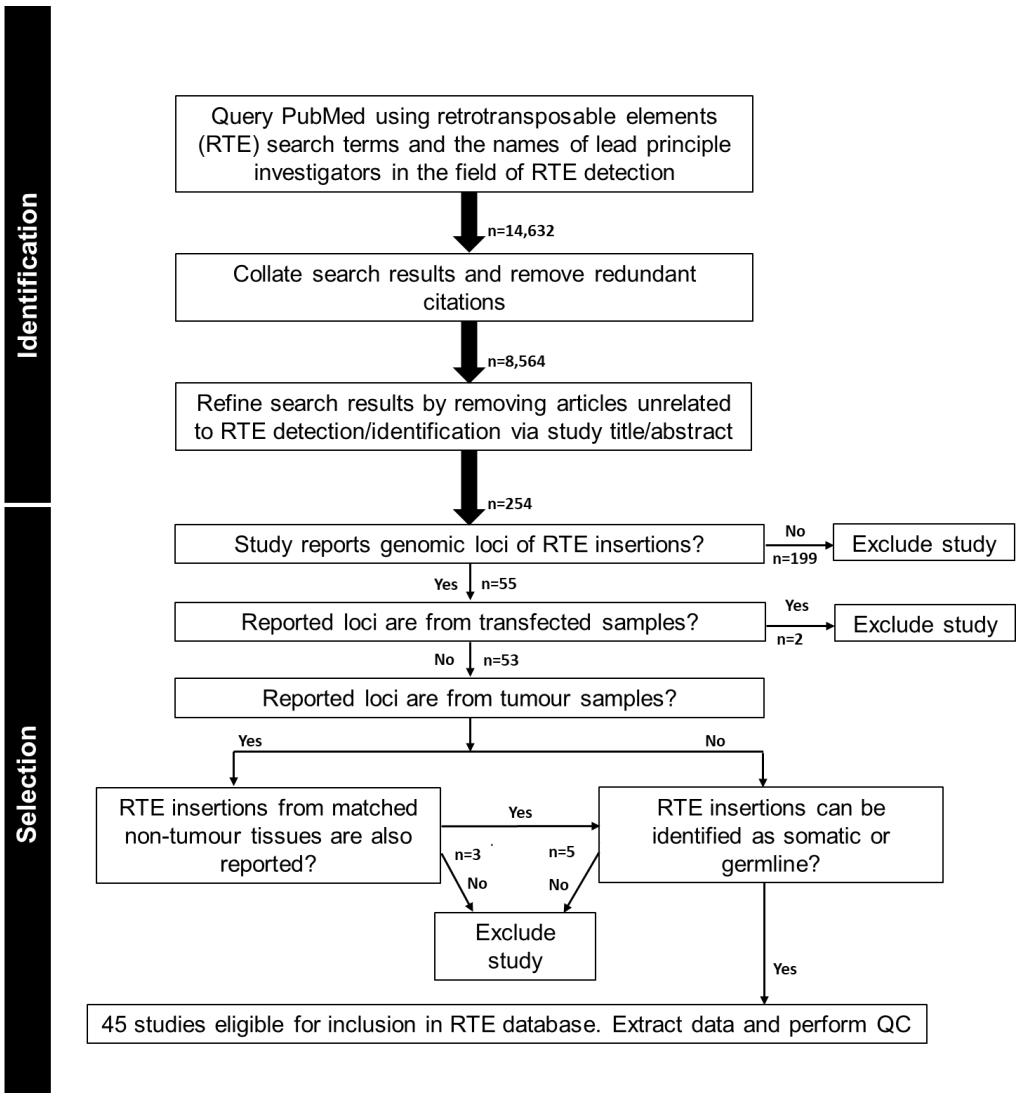


Figure 5: Overview of study selection process for curating a comprehensive database of retrotransposable elements (RTE) insertions.

2.2.3. Data curation:

Two types of datasets will be generated for each RTE subtype:

1. A general dataset of RTE insertions from all of the included studies.
2. A dataset by individual including insertions found with a matching sample id.

Six databases will be created: 2 for L1s, 2 for Alus, and 2 for SVAs. The study intends to use the general database to investigate the genomic landscape of polymorphic RTE insertions, which will provide insight into the effect of recent

RTE activity on genome function and integrity. The intended downstream analysis for the database by individual is to investigate the effect of structural polymorphisms mediated by RTE insertions on the fitness of its host using a more detailed approach as discussed in chapter 4.

The chromosomal locations of RTE insertions were extracted from the selected studies, either directly from the original publication, or from its supplementary files. Somatic RTE insertions, or insertions from subfamilies other than L1Hs, AluY or SVA_E/F were removed. RTE insertions not reported in GRCh37/hg19 coordinates were converted into hg19 coordinates using the UCSC Genome Browser liftOver tool (available at: https://genome.ucsc.edu/cgi-bin/hgLiftOver; Karolchik et al., 2004). The minimal information recorded from each insertion include:

1. Chromosomal location (chromosome name: insertion start-insertion end)
2. Study reference.
3. Study PubMed id (PMID).

Additional information recorded when available include:

1. Strand orientation.
2. RTE subtype.
3. Allele frequency.

The database by individual (db-individual) which includes non-reference RTE insertions (L1Hs, AluY, or SVA_E/F) identified in the genome of individuals recognised by a unique sample id will hereinafter be referred to as RTE profile(s). RTE profiles were extracted from studies reporting sample ID information

following the data extraction method described in the general database section. Sample ethnicity was also recorded when available.

2.2.4. Quality control:

As there is yet to be a gold-standard method for RTE detection, each of the studies included in the database follows its method of RTE detection consisting of a unique blend of computational parameters and quality control measures. As such, variations in the database are bound to exist, which jeopardize the consistency within the intended database. The following quality control (QC) measures have been applied to maximize the uniformity within the curated database:

1. Applying a minimum supporting reads threshold: Evrony et al. (2016) reported that applying a read count threshold of ≥ 3 supporting reads maintains the detection of at least half (53%) of true-positive insertions while excluding >99% of false-positive calls. As such, a minimum read count of ≥ 3 supporting reads was applied to reduce the likelihood of false positives, and insertions identified by less than 3 supporting reads were removed. In case of tumour studies, only the insertions supported by ≥ 3 reads in the matched non-tumour samples were retained.

2. PCR validation: Where PCR validation had been performed in a study, putative insertions that failed PCR validations or were located on unplaced/un-localised contigs were also excluded. Data from the offspring of trios were excluded to minimise allele frequency bias that may be introduced from related individuals. To minimise redundancy caused by

variation in breakpoint estimation of the different detection tools, RTEs within 200bp from each other were merged into a single insertion using the merge tool of BEDtools version 2.25.0 (Quinlan, 2014). The merge window was decided based on the analysis conducted by Mir et al. (2014) of the euL1db. Shared strand orientation was not required for the merging as the probability of two independent insertions being on different strands in the same location is extremely low (Sheen et al., 2000).

2.2.5. Addressing duplicate RTE profiles from the database by individual:

Where studies characterised the RTE profile of the same individual, only one RTE profile was selected for inclusion in db-individual to avoid inflation of allele frequencies in the curated database. The RTE profile retained was based on investigating the overlap between the identified RTE profiles in a single individual, as illustrated in figure 6. First, the duplicate RTE profiles were merged by 200bp to account for variability in breakpoint estimation of the different detection tools. A minimum of 50% overlap between both RTE profiles was required for profile selection. Duplicate RTE profiles that failed to meet this requirement suggested that at least one of the studies used a detection method with poor precision or recall therefore all the individual RTE profiles from such a study should be excluded to maintain consistency within the database. Of the 317 duplicate profiles, 293 resulted from a number of studies analysing samples that are part of phase 3 of the 1,000 genomes project (1KGP) dataset (Sudmant et al., 2015). The 1KGP is the biggest and most reliable human genomic variation study. Therefore, for duplicate RTE profiles where one profile is produced by the 1KGP and both profiles overlap by ≥ 50%, the 1KGP profile was selected for inclusion in db-individual. Of the remaining 24 duplicate profiles, three resulted from two

studies analysing samples that are part of the pilot phase of the 1KGP (Stewart et al., 2011). In these cases, when the overlap between both profiles was ≥ 50%, the profile with the number of insertions closest to the number observed per typical genome was retained (128 L1s, 915 Alus, and 51 SVAs; 1000 Genomes Project Consortium et al., 2015). For the 21 duplicate profiles not part of the pilot phase or phase three of the 1KGP dataset, the profile closest to the average per typical genome was retained in db-individual when both profiles overlapped by ≥ 50%. Finally, when the overlap between duplicate profiles was less than 50%, all individual profiles obtained from the study deemed unsatisfactory were excluded from db-individual to maintain consistency within the curated database.

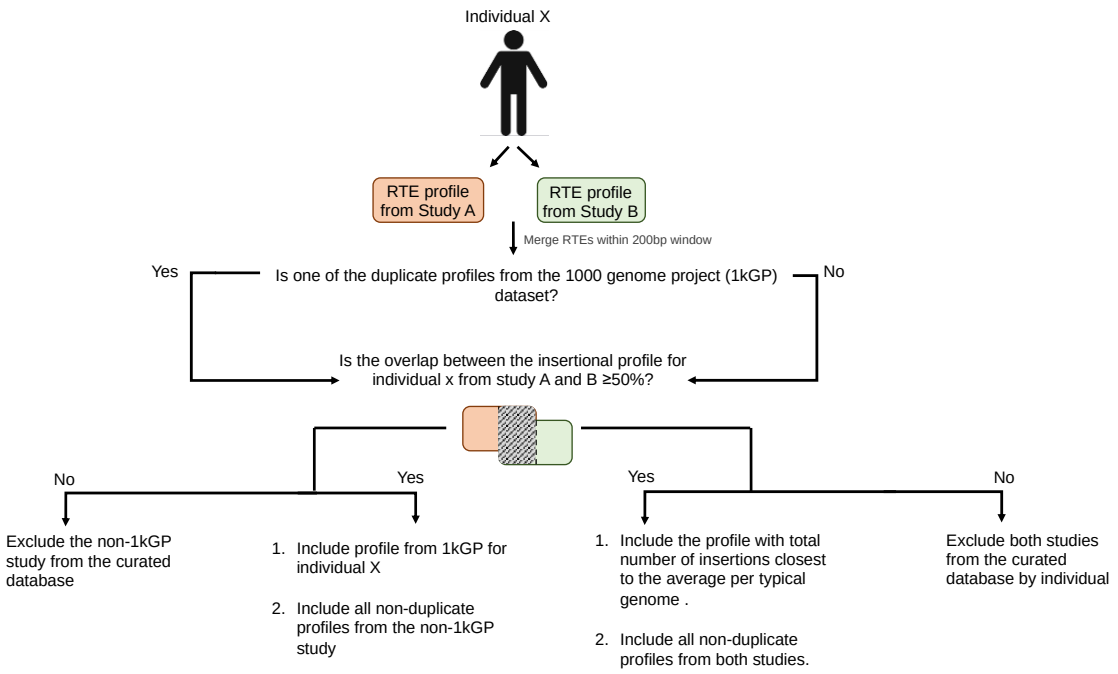


Figure 6: An overview of the procedure applied for selecting one retrotransposable element (RTE) profile for inclusion in the database by individual when duplicate RTE profiles for the same individual (individual x) were produced by two studies (study A and B). Note that an RTE profile refers to non-reference RTE insertions (L1Hs, AluY or SVA_E/F) identified in the genome of an individual.

2.3. Results:

2.3.1. Study selection:

Articles reporting non-reference retrotransposable element insertions (RTEs) were selected from PubMed (https://www.ncbi.nlm.nih.gov/pubmed/). Querying PubMed up to the cut-off date (11/04/2019) using the terms listed in tables 1 and 2 produced 14,632 articles in total. Removal of articles returned by multiple search terms (i.e. duplicates) and/or unrelated to RTE detection resulted in 254 publications, of which 45 studies meet the inclusion criteria of this study (table 4). An overview of study selection process is provided in figure 5.

Table 4: Studies included in the curated non-reference retrotransposable element (RTE) databases, including the database by individual (db-individual), along with the total number of samples and RTEs pre- and post- quality control (QC) steps. All 45 studies were included in the general database, of which 28 were included in the database by individual (db-individual).

#	Study ID	PMID	Total samples (S) | S in db-Indvl	Total RTEs pre|post QC		
				L1	Alu	SVA
1	Achanta et al., (2016)	27843500	6	10|10	--	--
2	Cardelli et al., (2012)	22495107	12	--	4|4	--
3	Doyle et al., (2017)	28585566	62	1161|903	--	--
4	Ewing et al., (2010)♦	20488934	25	367|365	--	--
5	Ewing et al., (2011)♦	20980553	310	67998|0	--	--
6	Ewing et al., (2015)●	26260970	18	1182|1052	--	--
7	Feusier et al., (2017)	28770012	213	--	5288|4890	--
8	Hehir-Kwa et al., (2016)	27708267	769	4011|1339	8670|5950	781|602
9	Helman et al., (2014)♦	24823667	200	1103|982	6248|6164	373|146
10	Kloosterman et al., (2015)	25883321	769	--	5|5	--
11	Konkel et al., (2015)	26319576	35	--	343|291	--
12	Kuhn et al., (2014)●	24847061	20	855|855	--	--
13	Kurnosov et al., (2015)●	25689626	1	19|8	9|0	--
14	Mir et al., (2014)‖	25352549	1023	142,495 S|21 M	--	--
15	Payer et al., (2017)	28465436	60 Pooled	--	809|579	--
16	Rouchka et al., (2010)	21044359	3	22|0	--	--
17	Tubio et al., (2014)●	25082706	244	1478|1478	--	--

(Table 4 continued)

Studies in db-individual

18	Arokium et al., (2014)●	25289675	1|1	100|73	--	--
19	Baillie et al., (2011)♦	22037309	3 & Pooled | 3	9279|96	19007|1673	2037|21
20	Beck et al., (2010)♦	20602998	6|1	68|21	--	--
21	Brandler et al., (2016)	27018473	235|138	1092|1092	6402|6402	417|417
22	Carreira et al., (2016)●	27843499	14|14	770|760	--	--
23	David et al., (2013)	23921633	7|5	--	6057|4521	--
24	Erwin et al., (2016)●	27618310	3|3	317|279	--	--
25	Evrony et al., (2012)♦	23101622	3|3	76|43	--	--
26	Evrony et al., (2015)●	25569347	1|1	24|24	48|45	8|8
27	Ha et al., (2016)	27478512	21|14	--	--	409|409
28	Hormozdiari et al., (2011)	21131385	8|7	--	4342|3554	--
29	Iskow et al., (2010)♦	20603005	30|30	2174|299	3799|52	--
30	Lee et al., (2012)♦	22745252	44|43	2639|2277	5531|5499	225|225
31	Nguyen et al., (2018)	29949758	19|19	554|186	--	--
32	Schauer et al., (2018)	29643204	61|61	559|558	2462|2459	34|27
33	Scott et al., (2016)●	27197217	1|1	104|104	--	--
34	Shin et al., (2019)	30699287	1|1	525|494	--	--
35	Shukla et al., (2013)♦	23540693	19|19	1018|863	6283|5301	329|316
36	Solyom et al., (2012)♦	22968929	21|21	7031|124	10429|1013	8476|18
37	Stewart et al., (2011)♦	21876680	156|153	998|172	4499|4405	79|76
38	Streva et al., (2015)●	25887476	7|7	228|228	--	--
39	Sudmant et al., (2015)●	26432246	2504|2504	3048|3048	12748|12748	835|835
40	Thung et al., (2014)●	25348035	3|2	233|227	1541|1488	62|54
41	Upton et al., (2015)●	25860606	5|5	395|385	--	--
42	Wildschutte et al., (2015)	26503250	53|53	--	1614|1599	--
43	Witherspoon et al., (2013)	23599355	169|160	--	5799|2674	--
44	Xing et al., (2009)	19439515	1|1	52|49	584|584	14|11
45	Yu Q et al., (2017)	28938719	90|90	2398|2398	6483|5975	400|367

‖ The euL1db publication. The number of insertions retained were from studies that did not require additional QC processing remove

♦ Studies included in euL1db. Post QC for L1 elements conducted by Roxane Dunbar.

● Studies in L1 database where QC was conducted by Roxane Dunbar.

2.3.2. Database structure/content:

Two databases were curated for each RTE type (L1Hs, AluY, and SVA_E/F): a general database that includes all non-reference RTE insertions, and a database by individual that includes RTE profiles identified within individual genomes. Both databases were curated using the same data extraction method and quality

control procedures as described in the method section. Details about the number of samples and RTE insertions obtained from each selected study, including pre- and post- the QC steps, are found in table 4. Additional information about each study, including data source, is available in appendix 1.

2.3.3. General database:

The general database (db-general) includes RTEs curated from 45 studies (listed in table 4) across 7,285 unrelated samples (including sample redundancy). RTEs within 200bp window were merged to minimise the chance of the same RTE being counted more than once due to varying breakpoint estimation of the detection tools estimations of the breakpoint location. The final database contains 39,798 non-reference RTEs from the L1Hs, AluY and SVA_E/F subfamilies. A breakdown of the number of RTEs per subtype is shown in table 5.

Table 5: Counts of retrotransposable elements (RTE) curated from 45 studies. RTEs of the same subfamily within a 200bp window were merged to minimise redundancy caused by variation in breakpoint estimation produced by differences in the RTE detection methods of the included studies. The general database (db-general) includes RTEs identified from all the included studies, while the individual database (db-individual) includes RTE profiles of individuals. The average number of RTEs per individual is calculated from db-individual.

RTE type	db-general	db-individual	
	RTE counts pre|post merge	RTE counts pre|post merge	Average per individual
L1Hs	20,813|10,211	13,800|6,377	134
AluY	77,875|27,699	59,915|18,698	1,064
SVA_E/F	3,532|1,888	2,784|1,085	51
Total	102,220|39,798	76,499|26,160	

Note that the number of pre-merge AluY elements in db-individual does not add up with the numbers of post-QC for Alu elements in table 4 as 77 Alu insertions from Baillie et al.*, (2011) were identified in the pooled sample only, thus were only included in db-general.*

2.3.4. Databases by individual for L1Hs, AluY, and SVA_E/F:

The pre-QC databases by individual collectively included 3,360 RTE profiles (including duplicate samples) curated from a total of 28 studies, as indicated in table 4. Any related individuals, including the offspring of trios, were excluded to control allele frequency bias that may be introduced by samples relatedness. Some studies have analysed samples collected from the same individual resulting in duplicate RTE profiles. Duplicate profiles pose the potential issue of overcalling for some of the individual profiles within the curated database, thus affecting the consistency of db-individual. Therefore, when duplicate RTE profiles were identified, a minimum of 50% overlap between both profiles was required before selecting one for inclusion in db-individual (Figure 6). All but three studies of the 13 studies that overlapped by a total of 317 individuals were retained, as summarised in table 6.

The post-QC databases collectively contains the insertional profile of 2,987 non-related individuals with an average of 134 L1Hs, 1,064 AluY and 51 SVAs per individual genome (Table 5). The average number of RTEs per individual genome is similar to the numbers reported by the 1000 genome project (128 L1s, 915 Alus, and 51 SVAs; 1000 Genomes Project Consortium et al., 2015). The majority of RTE insertions in db-individual were identified with an allele frequency (AF) below 1% (78% of L1Hs, 65.5% of AluY and 68% of SVA_E/F). RTEs in the curated database were identified in samples from diverse ethnic groups, including African, Asian, European, and American admixed. About one-third of the RTEs identified in db-individual were singletons, i.e., only present in one of the total samples in the database. (Table 7).

Table 6: Counts of duplicate RTE profiles in the database-by-individual (db-individual). A minimum of 50% overlap between duplicate profiles was required before selecting one of the profiles for inclusion in db-individual. All RTE profiles from the 3 studies that failed to meet this requirement (in bold) were excluded from db-individual.

Overlapping Studies	Number of overlaping samples	Average frequency of overlaping loci (%)	Included profile reference	Reason
Sudmant et al. (2015) and Stewart et al. (2011)	150	85% (Alu) 81% (SVA)	Sudmant et al. (2015)	1. MELT, the RTE detection tool of the 1kGP (Sudmant et al., 2015) is one of the best detection tools for RTE calling as shown by its superior performance in benchmarking studies (Rishishwar et al., 2016; Gardner et al., 2017; unpublished in-house benchmarking).
Sudmant et al. (2015) and Yu et al. (2017)	83	76% (Alu) 70% (SVA)	"	
Sudmant et al. (2015) and Lee et al. (2012)	2	72% (Alu) 65% (SVA)	"	
Sudmant et al. (2015) and Witherspoon et al. (2013)	43	73% (Alu)	"	
Sudmant et al. (2015) and Hormozdiari et al. (2011)	3	85% (Alu)	"	2. All RTEs in Sudmant et al. (2015) are genotyped and phased.
Sudmant et al. (2015) and David et al. (2013)	2	94% (Alu)	"	

(Table 6 continued)

Sudmant et al. (2015) and **Ha et al. (2016)**	10	33% (SVA)	"	
Thung et al. (2014) and Stewart et al. (2011)	2	98% (Alu) 71% (SVA)	Thun et al. (2014)	Thung et al. (2014) reported more RTE insertions, resembling the average number of insertions per individual as reported by the 1kGP and identified in db-individual.
Yu et al. (2017) and Stewart et al. (2011)	1	81% (Alu) 64% (SVA)	Yu et al. (2017)	Yu et al. (2017) reported more RTE insertions, resembling the average number of insertions per individual as reported by the 1kGP and identified in db-individual.
Shukla et al. (2013) and Schauer et al. (2018)	19	83% (Alu) 66% (SVA)	Shukla et al. (2013)	Shukla et al. (2013) reported more RTE insertions, resembling the average number of insertions per individual as reported by the 1kGP and identified in db-individual.

(Table 6 continued)

Wildschutte et al. (2015) and Hormozdiari et al. (2011)	1	32.7% (Alu)	Hormozdiari et al. (2011)	Hormozdiari et al. (2011) identified >80% of the insertions called by Sudmant et al. (2015) for the 3 overlapping samples between the two studies, therby increasing the confidence in the sensitivity of its method compared to the method of Wildschutte et al. (2015).
Xing et al. (2009) and Witherspoon et al. (2013)	1	46% (Alu)	Witherspoon et al. (2013)	Witherspoon et al. (2013) identified >70% of the insertions called by Sudmant et al. (2015) for the 43 overlapping samples between the two studies, therby increasing the confidence in the sensitivity of its method compared to the method of Xing et al. (2009).

Table 7: Count and frequency of singleton RTE insertions identified in each of the ethnic groups within the curated database.

Ethnicity	Singltons/Total insertions per ethnic group (%)		
	L1	Alu	SVA
African	242/1616 (15.0)	1093/9992 (10.9)	39/519 (7.5)
Ad mixed American	231/1527 (15.1)	1231/8890 (13.8)	36/496 (7.3)
American	4/188 (2.1)	11/1419 (0.8)	2/60 (3.3)
European	583/2418 (24.1)	1381/8557 (16.1)	91/540 (16.9)
East Asian	1052/2338 (45.0)	1055/7392 (14.3)	85/496 (17.1)
South Asian	170/982 (17.3)	461/5351 (8.6)	18/363 (5.0)
Other	796/2166 (36.7)	941/5880 (16.0)	74/381 (19.4)
Total Singltons/Total RTEs in db-individual (%)	345/1085 (31.8)	6173/18698 (33.0)	3078/6377 (48.3)

Note that the total number of insertions per ethnic group may not add up to the total number of insertions in db-individual as some RTE insertions are identified in more than one ethnic group.

2.4. Discussion:

A comprehensive collection of retrotransposable element insertions in humans have been curated and organised into two databases; a general database and a database by individual (Table 5). The information in this database has been curated from a total of 45 peer-reviewed articles (listed in table 4) published in the last decade up to the 12th of April 2019, including data from the final phase of the 1000 genome project (1kGP) (Sudmant et al., 2015) as well as the Genome of the Netherlands (GoNL) project (Hehir-Kwa et al., 2016). The selected articles have been identified via PubMed using the search terms listed in Tables 1 and 2. The general database (db-general) holds 10,211 L1Hs, 27,699 AluY, and 1,888 SVAs from the E and F subfamilies (SVA_E/F). The database of individual RTE profiles contains 6,377 L1Hs, 18,698 AluY, and 1,085 SVA_E/F identified in 3,360 non-related individuals from diverse ethnic backgrounds. All RTEs in the curated database are non-reference RTEs, referring to RTEs that are absent from the reference genome.

2.4.1. Study database vs. existing online databases

Compared to existing online RTE databases, the curated database holds 10-fold the amount of RTEs in dbRIP (http://dbrip.brocku.ca/; Wang et al., 2006; n=3,106), and an additional 36% germline L1Hs entries compared with the euL1db list (http://eul1db.unice.fr; Mir et al., 2014; n=8,012). In addition, the curated database only includes germline insertions that are polymorphic in the population as the main aim of the database is to provide a resource of RTE variants that potentially contribute towards host susceptibility to disease. This is in contrast to dbRIP and the euL1db databases that also hold insertions

associated with Mendelian disorders (dbRIP) that tend to segregate in specific families, or somatic insertions (euL1db) that do not get passed on to future generations (Wang et al., 2006; Mir et al., 2014).

To our knowledge, the curated database of this study is the first to apply a threshold of ≥3 supporting reads. This quality control (QC) measure was suggested by Evrony et al. (2016) to minimize false-positive calls. Applying this threshold may have potentially excluded a few low-frequency true-positive insertions, however, it may have also excluded the majority of false-positive calls as demonstrated in the analysis of Evrony et al. (2016). A recent benchmarking study has also demonstrated the importance of applying a minimum supporting reads threshold. Rishishwar et al. (2016) found that the best performing tools out of the seven tools analysed by their study had applied a minimum supporting reads threshold by default. The best performing tools applying such a threshold included MELT (Gardner et al., 2017) and Mobster (Thung et al., 2014) used by the 1kGP (Sudmant et al., 2015) and the GoNL project (Hehir-Kwa et al., 2016), respectively.

2.4.2. Issues with current methods of RTE detection

Accurate RTE detection is essential for understanding the effect of recent RTE activity on genome function. Detecting true-positive RTE insertions in the human genome is challenging, mainly due to the repetitiveness and high sequence homology of these elements in the human genome. Numerous computational tools based on the analysis of NGS data have been developed over the past two decades for genome-wide detection of RTE insertions (Ewing, 2015; Goerner-potvin and Bourque, 2018). However, most RTE detection tools rely on short sequencing reads that do not span the whole integration site. Consequently,

variations in recall sensitivity and positional accuracy do exist between the available RTE detection tools.

The discrepancies in the results of the duplicate RTE profiles curated by this study show the extent of variations in the recalls of different RTE detection tools when analysing the same sample (Figure 6). There has yet to be a standard approach against which the performance of current and future tools operate. Long-read sequencing technologies such as PacBio and MinION (Rhoads and Au, 2015; Lu et al., 2016) have the potential to improve the recall and precision of RTE detection, and eventually become the standard approach for RTE discovery.

2.4.3. Correlations between population growth rate and allele frequency spectrum

Most RTE insertions in the curated database have low allele frequencies (AF), including over 65% of the insertions with an AF below 1%. The greater proportion of rare variants (AF < 1%) is largely due to RTEs unique to individuals, i.e., singletons. This is consistent with the well-established effect of the recent explosive growth in population size (Keinan and Clark, 2012; Gao and Keinan, 2016). The increased load of singletons is explained by the growing population shifting the balance between the rate of RTE insertion and elimination from the population (Gazave et al., 2013; Bourgeois and Boissinot, 2019). Therefore populations with a higher growth rate are expected to have a higher load of singletons. Gravel et al. (2011) estimated the growth rate in the East Asian and European populations at 0.48% vs. 0.38% per generation, respectively, starting about 23 thousand years ago (kya) using exon and low-coverage sequencing

data from the 1kGP (n=40 per panel). These results are consistent with the higher fraction of singletons identified in East-Asian compared with European samples. Another recent study inferring population size changes from whole-exome sequencing data of 4,298 European vs. 2,217 African samples reported a higher growth rate in the European population compared to the African population. Chen et al. (2015) estimated the growth rate of the European population at 1.49% per generation starting ~7.26 kya vs. 0.74% per generation starting ~10.01 kya in the African population. These results are also consistent with the higher fraction of singletons identified in the European samples in comparison to the African samples (Table 7).

2.4.4. Correlations between population growth rate and efficiency of natural selection

The excess of singleton RTE variants as a consequence of the recent population explosion has the potential to increase the genetic risk of complex disorders within the average individual of the growing populations by impacting purifying selection against deleterious insertions. A simulation-based study by Gazave et al. (2013) suggested that individuals in a growing population show a moderate increase in the fraction of deleterious mutations in comparison to individuals in a non-growing population. This effect was suggested to be due to the increased efficiency of selection at eliminating the most extremely deleterious mutations, resulting in a slight increase in the number of weakly deleterious mutations (Gazave et al., 2013; Gao and Keinan, 2016). In addition, the total number of *de novo* RTE insertions is likely to be much higher in a growing population, which

consequently increases the number of RTE variants that have the potential to contribute towards the susceptibility of complex diseases.

Identifying rare RTE variants in the growing population that are caused by recent insertions requires sequencing of a very large sample size. This is because singletons identified in a small sample of 100 individuals have a frequency of 0.5% and are likely caused by older insertions compared to singletons with a frequency of 0.005%, identified in a sample of 10,000 individuals. However, increasing sample size will also increase type I error which may complicate the analysis and reduce the ability to distinguish between true insertions and false positives. These issues may potentially be tackled via the development of RTE detection tools analysing long sequencing reads. Such tools may increase the sensitivity of RTE detection and reduce the complexity of RTE discovery (Ewing, 2015; Jiang et al., 2019).

2.4.5. Study overview

In summary, a comprehensive database of RTE elements has been curated, updating the information reported in existing online databases. The curated database holds an excess of rare insertions, consistent with the well-established effect of the recent population explosion. Capturing such variants is of particular interest for studying the impact of population growth on the genetic architecture of complex disorders. In effect, the curated database is a useful resource for understanding the contribution of RTE variants towards human phenotype and disease.

3. Genomic Distribution of non-LTR RTEs

3.1. Introduction

Retrotransposable elements (RTEs) are the most abundant type of repetitive sequences, essentially constituting 45% of the human genome (Lander et al., 2001). The majority of RTEs in the human genome are remnants of ancient insertions, hence they are no longer capable of transposition due to inactivating mutations and internal rearrangements (Lander et al., 2001; Wei et al., 2001; Hancks and Kazazian, 2016). Only the most recently evolved RTE subfamilies retain the ability to transpose within the human genome, creating insertional polymorphisms within and between populations (Wang et al., 2005; Mills et al., 2007; Huang et al., 2010). The ongoing activity of the non-LTRs (Long Terminal Repeats) retrotransposons is driven by a single autonomous family, known as the long interspersed nuclear element-1 (L1 for short).

3.1.1. Retrotransposition of L1s, Alus, and SVAs

L1s are the most abundant type of transposable elements (TE) in humans, with over half a million copies of L1s constituting up to 17% of the human genome (Lander et al., 2001). Nevertheless, only about 100 L1s from the evolutionary young human-specific L1 subfamily (L1Hs) in each individual genome are full-length and capable of transposing, of which a handful have been described as 'hot' L1s, due to them being highly active (Brouha et al., 2003). A full-length L1 element is about 6 kilobases (kb) long and encode two proteins that are essential for its retrotransposition: a small RNA binding protein (ORF1p) and a large protein with endonuclease (EN) and reverse transcriptase (RT) activities (ORF2) (Dombroski et al., 1991; Lander et al., 2001). L1s retrotranspose via a process

termed target-primed reverse transcription (TPRT), whereby the ORF2p reverse transcribes an RNA copy of the parent L1 and integrates the complementary DNA (cDNA) elsewhere in the genome (Cost et al., 2002; Scott and Devine, 2017).

The non-autonomous Alu and SVA elements have evolved to capture the L1 ORF2p, thus L1s mediate the retrotransposition of Alu and SVA elements in *trans* during TPRT (Dewannieux et al., 2003; Raiz et al., 2012). Although the L1 machinery mediates the mobilisation of the non-LTR RTEs, endogenous L1s, Alus, and SVAs, referring to reference insertions that are fixed in the human genome, are known to accumulate in different genomic regions.

3.1.2. Distribution of endogenous RTEs

Endogenous L1 elements accumulate in AT-rich low-activity regions whereas Alus and SVA elements accumulate in GC-rich, high-activity regions (Smit, 1996; Smit, 1999; Lander et al., 2001; Pavlíček et al., 2002; Wang et al., 2005). The accumulation of L1 elements in AT-rich regions have been credited to the specificity of the L1 endonuclease target motif (5'-TTTT/AA-3'), which is significantly denser in AT-rich regions of the genome (Cost and Boeke, 1998; Lander et al., 2001; Graham & Boissinot, 2006). Investigating the surrounding GC content of endogenous RTEs by evolutionary age have shown that the evolutionary young subfamilies of L1s, Alus, and SVAs accumulate in regions of lower GC content in comparison to older subfamilies, a difference that is potentially masked when all the endogenous elements of each type are analysed as one (Lander et al., 2001; Medstrand et al., 2002; Wang et al., 2005; Kvikstad and Makova, 2010; Costantini et al., 2012). As such the observed distribution bias in the genomic distribution of Alus and SVAs, in comparison to L1 elements,

is thought to have been reshaped by differential selection forces acting on each element (Dewannieux et al., 2003; Raiz et al., 2012; Tang et al., 2018).

3.1.3. Effects of RTE activity on genome function

RTE insertions are known to affect genome function in a variety of mechanisms, including mediating post-insertional genomic rearrangements. Insertions in genic regions can interfere with gene function or induce loss-of-function mutations through interfering with gene transcription or splicing (Conley and Jordan, 2012; Chénais, 2016; Hancks and Kazazian, 2016; Bourque et al., 2018). In addition, the high copy number and sequence similarity between RTE elements can promote non-allelic homologous recombination (NAHR) events that can cause significant deletions and duplications (Lee et al., 2012; Startek et al., 2015; Nazaryan-petersen et al., 2016). The reader can refer to the following review articles for a more detailed overview of how TE activity can impact genome function both directly and indirectly: Hancks and Kazazian (2016), Bourque et al. (2018), and Saleh et al. (2019).

3.1.4. Effects of RTE activity on human health

In terms of causing disease or susceptibility to disease, RTE elements are, on whole, not as well studied as single nucleotide variants (SNVs). However, structural variants (SV) mediated by RTE insertions have been implicated with genetic diseases.

3.1.4.1. RTEs and monogenic diseases

The first reported case of a disease-causing RTE insertion is a *de novo* L1 insertion into exon 14 of the factor VIII (F8), identified in two unrelated

haemophilia A patients (Kazazian et al., 1988). Since then, 124 monogenic diseases mediated by RTE insertions have been reported (Hancks and Kazazian, 2016). Nevertheless, disease-causing RTE insertions are rare events, estimated to be responsible for ~1/1,000 disease-causing mutations (Lutz et al., 2003). Advances in genome sequencing technology, and the development of efficient computational detection tools capable of simultaneously analysing numerous genomic samples to identify RTE insertions on a genome-wide level, have led to a substantial increase in the number of recovered RTE insertions (Ewing, 2015; Rishishwar et al., 2017). Such advances have also improved the recovery of somatic and polymorphic insertions, including insertions implicated in complex disorders.

3.1.4.2. RTEs and complex diseases

Somatic L1 retrotranspositions are recognised as mutagenic agents in many epithelial cancers (Chénais, 2016; Burns, 2017; Scott and Devine, 2017). In addition, increased somatic L1 retrotransposition in the brain has been reported in several neurological and neuropsychiatric disorders, including schizophrenia, amyotrophic lateral sclerosis (ALS), and Alzheimer's disease (Guffanti et al., 2016; Savage et al., 2019; Terry et al., 2020). As for germline insertions, a recent study identified a polymorphic SVA element in an intron of the TAF1 gene, associated with an increased risk of X-linked Dystonia-Parkinsonism (XDP) through promoting intron 32 retention, resulting in reduced TAF1 gene expression (Aneichyk et al., 2018). The length of the hexanucleotide repeat domain of the SVA element was found to have a significant inverse correlation with the age of XDP onset (Bragg et al., 2017). Another study identified a polymorphic Alu element in an intron of the CD58 gene, associated with an increased risk of

multiple sclerosis through promoting skipping of exon 3, resulting in reduced CD58 gene expression (Payer et al., 2019).

Deleterious RTE insertions that affect genome function, consequently reducing the fitness of their host, are subject to purifying selection to such a degree that they cannot reach high allele frequencies in the population, and are eventually rendered extinct from the human genome (Loewe, 2008). In contrast, neutral insertions do not affect genome function, and as such, they can reach high frequencies and eventually become fixed in the human genome. Under these assumptions, insertions from the ancient subfamilies are more likely to be neutral than recent insertions from the actively transposing RTE subtypes.

3.1.5. Genomic distribution of recent RTE insertions

The genomic load of RTE activity on genome function can be inferred by comparing the distributions of RTEs from subfamilies of different evolutionary ages. Early studies, including ones from before the initial sequencing and analysis of the draft human genome, comprehensively characterised the genomic landscape of endogenous RTE insertions (Soriano et al., 1983; Smit, 1999; Gu et al., 2000; Lander et al., 2001; Wang et al., 2005). The extent to which the landscape of new insertions resembles that of the endogenous RTE elements which have become fixed in the human genome remains ambiguous. Studies comparing the distribution of fixed RTEs with insertions from the currently amplifying RTE subfamilies (Table 8), have suggested that the majority of new RTEs integrate in neutral or deleterious regions, in contrast to endogenous elements that accumulate in what have been called genomic safe-havens (Boissinot and Furano, 2005; Song and Boissinot, 2007). Nevertheless, the evolutionary fate of a new insertion, including its impact on genome function as

well as its ability to mobilise and generate new insertions, largely depends on where it inserts into the genome.

Some studies adopted an experimental approach for investigating the genomic landscape of RTE insertions in the human genome. In this approach, *de novo* insertions are induced using engineered elements in cell cultures. The induced insertions are then recovered and their pre-insertion loci are characterised and compared with the distribution of endogenous or polymorphic elements (Raiz et al., 2012; Sultana et al., 2019; Flasch et al., 2019; Chen et al., 2020). The main advantage of using this method is that the induced insertions experience minimal selection and thus recapitulate the initial integration site of RTEs. However, the distribution of such insertions may not be representative of natural elements segregating in the human genome, as the influence of cell-line specific factors on the retrotransposition assay cannot be ruled out. In addition, some *de novo* insertions may integrate into genomic locations that may potentially cause embryonic lethality (Boissinot et al., 2004).

The focus of the more recent RTE detection studies seems to have shifted away from characterising the landscape of RTE insertions. Few recent studies reported a partial characterisation of the insertional landscape of polymorphic RTE insertions, mainly reporting its distribution within gene regions (Stewart et al., 2011; David et al., 2013; Witherspoon et al., 2013; Thung et al., 2014; Ha et al., 2016). As such, a comprehensive study analysing the genomic distribution of polymorphic insertions against fixed RTE elements, in order to advance the current understanding regarding the role of germline RTE activity on genome function, has yet to be published.

3.1.6. Aims and objectives

The current study aims to investigate the potential effects of germline RTE activity on genome function and integrity by comparing the genomic distributions of polymorphic RTE insertions with the distribution of endogenous RTEs that are fixed in the human genome. The in-house curated database discussed in chapter 2 is utilised for this analysis. The curated database includes a high fraction of singleton and rare RTE insertions, as a result of its larger sample size, in comparison to previous studies conducting similar analysis (Ovchinnikov et al., 2001 [n=32 L1Hs vs. 30 ancient L1s]; Boissinot et al., 2004 [n=344 L1Hs vs. 300 ancient L1s]; Wang et al., 2005 [106 SVA_E/F vs. 2,656 SVA_A-D]; Cordaux et al., 2006 [43 polymorphic vs. 60 fixed AluY]; Ewing and Kazazian, 2010 [367 non-reference vs. 772 reference L1Hs]).

As such, the results of this study are expected to be intermediate between the reported distribution of *de novo* and polymorphic insertions. This study expands on previous studies in the literature by simultaneously comparing the distribution of L1, Alu, and SVA elements in order to understand the effect of each RTE type on genome function and integrity. Knowing the functional impact of recent germline insertions is fundamental for understanding the impact of RTEs, with respect to human health and disease, and the likely contribution of RTE-mediated SV to the missing heritability issue.

Table 8: Summary of the genomic distribution of old RTEs that are fixed in the human genome, in comparison to the distribution of polymorphic RTE elements.

Genomic feature	Old (fixed)	Young (polymorphic)	Supporting references
Chromosomal distribution			
L1	Enriched on the X-chromosome as a result of positive selection due to their role in propogating the X silencing signal. Chromosomal distribution correlated positively with chromosome size.	Not enriched or overrepresented on any chromosome. Distribution correlated positively with chromosome size.	Smit (1999); Bailey et al., (2000); Lander et al., (2001); Medstrand et al., (2002); Boissinot et al., (2004); Ewing and Kazazian, (2010); Tang et al., (2018); Sultana et al., (2019); Flasch et al., (2019); Chen et al., (2020)
Alu	Overrepresented on chromosome 19 more than expected for its size. Alu density correlated with chromosome gene density.	Not enriched or overrepresented on any chromosome. Distribution correlated positively with chromosome size.	Lander et al., (2001); Grover et al., (2004); Carter et al., (2004); Otieno et al., (2004); Wagstaff et al., (2012); Tang et al., (2018)
SVA	Overrepresented on chromosome 19 more than expected for its size. SVA density correlated with chromosome gene density.	Not enriched or overrepresented on any chromosome. Distribution correlated positively with chromosome size and gene density.	Wang et al., (2005); Savage et al., (2013); Tang et al., (2018); Gianfrancesco et al., (2019)

(Table 8 continues)

Local GC content			
L1	Accumulate in the most AT-rich regions of the genome	Accumulate in AT-rich regions, but are more evenly distributed in this region in comparision to fixed elements	Gu et al., (2000); Lander et al., (2001); Medstrand et al., (2002); Boissinot et al., (2004); Sultana et al., (2019); Flasch et al., (2019); Chen et al., (2020)
Alu	Accumulate in GC-rich regions	Accumulate in AT-rich regions	Gu et al., (2000); Lander et al., (2001); Medstrand et al., (2002): Jurka et al., (2004); Belle et al., (2005); Costantini et al., (2012); Wagstaff et al., (2012)
SVA	Accumulate in GC-rich regions	Accumulate in regions of higher GC-content in comparision to fixed elements	Wang et al., (2005); Raiz et al., (2012); Savage et al., (2013); Gianfrancesco et al., (2019)
Genic distribution			
L1	Accumulate in intergenic regions. Most intragenic insertions are intronic and significantly more often in the antisense strand.	Accumulate in intergenic regions. Significantly depleated in genic and intronic regions compared with reference insertions.	Lander et al., (2001); Boissinot et al., (2004); Ewing and Kazazian, (2010); Tang et al., (2018); Sultana et al., (2019); Flasch et al., (2019); Chen et al., (2020); Watkins et al., (2020)

(Table 8 continues)

Alu	Accumulate in intragenic regions, mostly in introns.	Significantly depleated in gene regions in comparision to fixed Alu elements.	Lander et al., (2001); Cordaux et al., (2006); Hormozdiari et al. (2011); Wagstaff et al., (2012); Witherspoon et al., (2013); David et al., (2013); Watkins et al., (2020)
SVA	Accumulate in intragenic regions, mostly in introns. The number of fixed SVA elements in gene deserts is lower than expected if the elements inserted randomly.	Accumulate in intergenic and intronic regions. The yonger subfamilies were underrepresented in gene deserts in comparision to the expected distribution, however, they were found in lower frequencies in gene deserts in comparision to the fixed SVA elements.	Wang et al., (2005); Hancks et al., (2011); Raiz et al., (2012); Savage et al., (2013); Ha et al., (2016); Tang et al., (2018)
Local Recombination rate			
L1	Studies suggested that fixed L1s accumulate in low and non-recombining regions as a result of selection	Accumulate in low recombination regions.	Lander et al., (2001); Medstrand et al., (2002); Boissinot et al., (2004); Abrusán et al., (2006); Song and Boissinot, (2007);

(Table 8 continues)

Alu	Accumulate in regions of higher recombination. Studies suggested that fixed Alu elements contain a motif associated with recombination and genome instability	No relationship between the distribution of polymorphic elements and local recombination rate	Lander et al., (2001); Medstrand et al., (2002); Hackenberg et al., (2005); Myers et al., (2008); Witherspoon et al., (2009)
SVA	No direct results, however, SVA elements have been reported to cause recombination-mediated deletions in the human genome suggesting their occurrence in recombining regions. In addition, SVA elements accumulate in GC-rich regions, which have a positive correlation with recombination rates.		Fullerton et al., (2001); Wang et al., (2005); Supplementary table 3 of Myers et al., (2008); Lee et al., (2012)

3.2. Methods

An overview of the methods workflow is presented in figure 7.

3.2.1. RTE Datasets

All datasets are based on the GRCh37/hg19 genome build co-ordinates.

3.2.1.1. Non-reference database:

The chromosomal locations of non-reference RTEs were taken from the in-house databases described in chapter 2. The non-reference database consists of 10,211 L1Hs, 27,699 AluYs, and 1,888 SVAs from the E and F human-specific subfamilies.

3.2.1.2. Reference database:

In this context, the reference database consists of those RTEs that are fixed in the human genome, such that any two individuals will carry the same insertion. The RepeatMasker Table was downloaded from the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables; Karolchik, 2004) and 38,366 L1PA2-5, 307,612 AluJ, and 1,005 SVAs from the A and B subfamilies were extracted. These elements are thought to be incapable of retrotransposition due to the build-up of random inactivating mutations (Lander et al., 2001; Wei et al., 2001; Hancks and Kazazian, 2016).

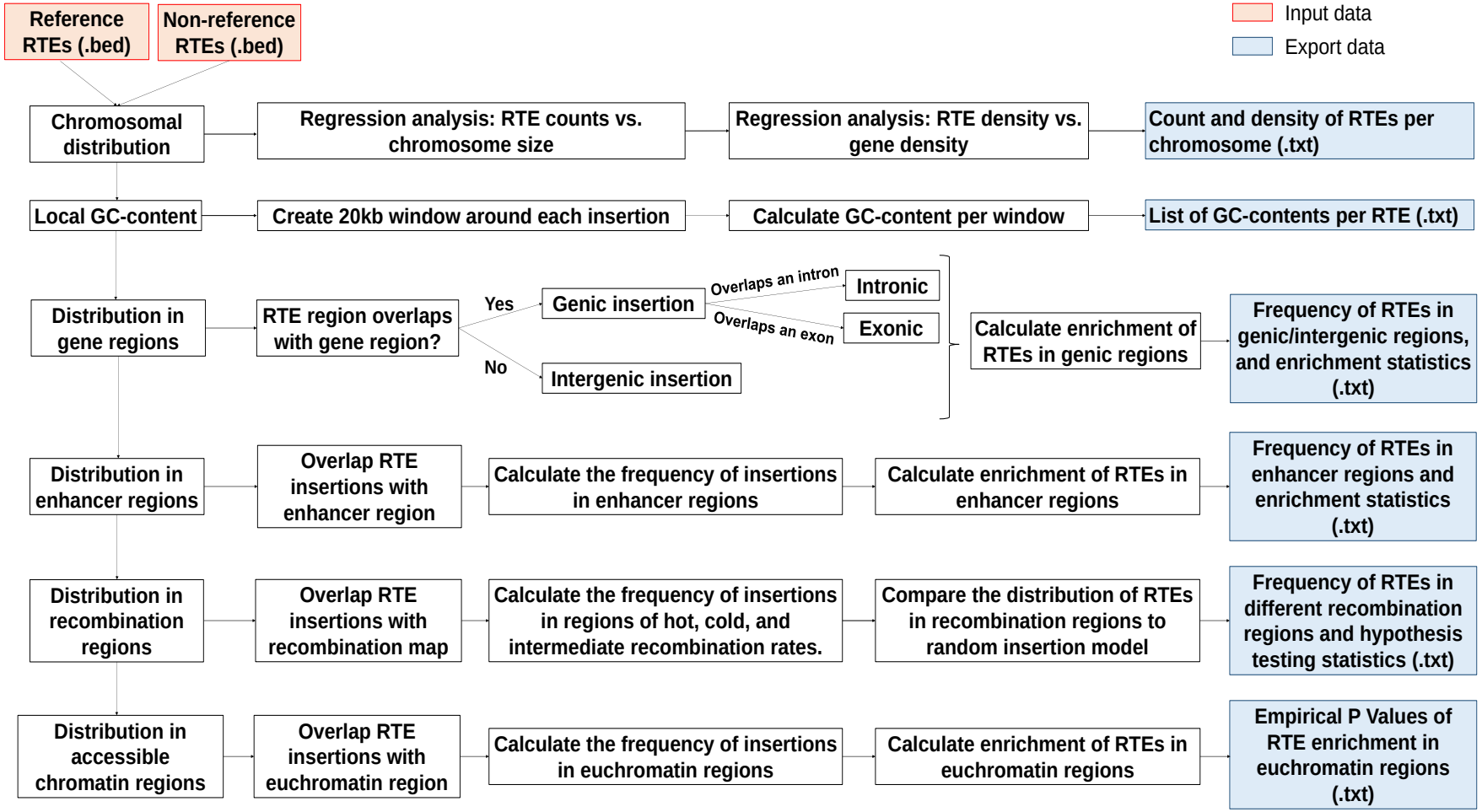


Figure 7: Genomic Distribution workflow. Each of the distribution analyses are conducted on the reference and non-reference RTEs input files. The number of RTE insertions per chromosome are counted. Subsequently the local genomic environment is investigated including local GC-content, overlap with gene and enhancer regions, and local chromatin accessibility is investigated by overlapping the location of RTEs with H3K4me3 epigenetic profiles associated with euchromatin domains.

3.2.2. Genomic Distribution analyses

3.2.2.1. Chromosomal distribution

The total number of RTEs per chromosome within the reference and non-reference L1Hs, Alu, and SVA databases were counted in UNIX. To investigate if each of the elements were randomly distributed throughout the genome, the chromosomal distribution was investigated based on the size of each chromosome, using linear regression performed in core R (R Core Team, 2012) version 3.4.0. The length of each chromosome was obtained from the human assembly data at the Genome Reference Consortium (GRC) (https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh37).

Given that endogenous L1s are known to accumulate in AT-rich regions, while endogenous Alu and SVA elements accumulate in GC-rich regions, and that genes are known to reside in GC-rich regions of the genome, the chromosomal distribution of RTEs based on gene density was investigated. Understanding the chromosomal distribution of the reference and non-reference RTEs based on gene density, may disentangle the combined influence of both negative selection against the potentially deleterious RTEs within genes, and the insertional bias of the particular RTE. To this end, a second linear regression analysis was performed to investigate the relationship between the chromosomal distribution of reference and non-reference RTEs and gene density across the genomes, defining the density of RTEs and genes as the numbers of each per one million base pairs. Chromosomal locations of RefSeq genes were obtained from the UCSC Genome Browser (University of California Santa Cruz) (https://genome.ucsc.edu/cgi-bin/hgTables) (Karolchik, 2004). R codes for both

regression analyses are provided in GitHub at: https://github.com/RandaAli1/MyPhDproject/tree/master/MyAnalysis.

3.2.2.2. Local GC content

The hg19/GRCh37 human genome assembly (chromFa.tar.gz) was downloaded from the UCSC genome browser ftp website (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/). Individual fasta files were concatenated in UNIX. The start and end positions of each RTE insertion were extended by 10 kilobases (Kb) to create a 20 Kb window around each insertion using the slop function of BEDTools (Quinlan, 2014) version 2.25.0. This window size was chosen in concordance with previous studies (Lander et al., 2001; Boissinot et al., 2004; Gasior et al., 2007). The GC-content for each window was calculated using the nuc function of BEDTools (Quinlan, 2014) version 2.25.0, and the hg19 human genome assembly. To account for variations in GC content within the human genome, the hg19 genome assembly was divided into 20 Kb adjacent windows from start to end using the makewindows function of BEDTools (Quinlan, 2014) version 2.25.0, and the GC-content of each genomic window was calculated. Windows of different GC-contents were grouped into bins of 2% width using R (R Core Team, 2012) version 3.4.0. The distribution of reference RTEs in genomic regions of different GC-contents was compared with the distribution of non-reference RTEs and with the GC distribution of the genome, using the two-sample Kolmogorov–Smirnov (K-S) test. The K-S test is a nonparametric goodness-of-fit test used to compare the cumulative distribution function of two samples to determine whether the tested samples share the same underlying distribution (Lall, 2015). The K-S test returns a D-statistic, representing the maximum difference between the cumulative distribution

functions of the tested samples and a P-value of significance (Frank and Massey, 1951; Lall, 2015). The GC-content analysis codes are provided in GitHub at: https://github.com/RandaAli1/MyPhDproject/tree/master/MyAnalysis.

3.2.2.3. Distribution in functional regions

3.2.2.3.1. RefSeq genes

The chromosomal locations of genes across the genome were based on the Reference Sequence collection (RefSeq), providing a non-redundant and well-annotated record of sequences submitted to the National Centre for Biotechnology Information (NCBI). The annotated NCBI RefSeq genes were downloaded from the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables; Karolchik, 2004) (last update 11.09.2017); under the group 'genes and gene predictions' and table 'NCBI RefSeq (refGene)'. Annotations with both the accession prefix NM and NR (mRNA or non-protein-coding RNAs, respectively) were included. The 'output format' was set to 'BED – browser extensible data' allowing additional user choices.

Two files were generated containing the chromosomal locations of:

I. All genic regions: 'Whole gene' option selected under the heading 'Create one BED record per' to generate a file of the start and stop locations of all genes. The resultant genic regions file contains the transcript start and end of 19,407 coding and 11,173 non-coding genes.

II. Exonic and intronic regions: To generate a file including the chromosomal start and stop location of all genes exons and introns,

the ‘Exons plus’ and ‘Intron plus’ options were selected under the heading ‘Create one BED record per’. The two separate files indicating the exons and introns were combined using the ‘cat’ command in UNIX.

Non-coding genes (prefixed with NR_*) such as microRNAs (MiRNAs) were included, as they have been shown to have key regulatory roles within the human genome, and there are numerous cases where mutations in non-coding genes have been shown to cause diseases in humans (Kornienko et al., 2013; Patrushev et al., 2014; Patil et al., 2014; De Almeida et al., 2016; Quinn et al., 2016).

3.2.2.3.2. Enhancer file

GeneHancer (Fishilevich et al., 2017) is a database of human enhancers generated by computationally integrating data from the following genome-wide databases while eliminating redundancy:

[1] The Encyclopedia of DNA Elements (ENCODE; Zerbino et al., 2015).

[2] Ensembl regulatory build (The ENCODE Project Consortium, 2012).

[3] The functional annotation of the mammalian genome (FANTOM; Andersson et al., 2014) project.

[4] The VISTA Enhancer Browser (Visel et al., 2017).

The chromosomal locations of the Enhancer elements were obtained from the GeneHancer database (Fishilevich et al., 2017) via the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables; Karolchik, 2004), selecting the ‘GH Reg Elems (geneHancerRegElements)’ table as part of the ‘GeneHancer’ track under the ‘Regulation’ group, including a total of 250,718 entries across the human genome (last updated 2-09-2018). Of these enhancers, approximately

44% (n=110,165) were defined as ‘elite’ status based on their defined strength of the identification evidence (Fishilevich et al., 2017).

3.2.2.3.3. Analysis:

The distribution of RTEs within functional genomic regions was identified by overlapping the genomic position of RTEs, with that of genes and enhancer regions, using the intersect function of bedtools (Quinlan, 2014) version 2.25.0. RTE insertions that did not interrupt a gene region were considered intergenic. Fisher's exact tests were carried out to compare the distribution of reference and non-reference RTEs, in intergenic vs. intragenic regions, and in enhancer vs. enhancer-free regions. The analysis codes for the distribution of reference and non-reference RTEs in genic and enhancer regions are provided in GitHub at: https://github.com/RandaAli1/MyPhDproject/tree/master/MyAnalysis.

3.2.2.4. Local recombination rate

The Decode recombination map (https:www.decode.com/; Kong et al., 2010) avoids the limitations of linkage disequilibrium (LD)-based maps, such as biases in recombination rate estimates due to the effect of natural selection on LD, and lack of information about sex differences (Kong et al., 2010). DeCODE produced sex-specific recombination maps using phased haplotypes of 15,257 parent-offspring pairs and a total of 298,069 genome-wide SNPs. Recombination was placed in the region between the two closest flanking markers in the parent within a resolution of 10 Kb. The X-chromosome and 5 Mb (mega base) regions at both ends of autosomal chromosomes were excluded due to reduced reliability in

placing recombination in these regions. The map used in this study is the standardised sex-averaged map, which is essentially the average of the male and female recombination maps.

DeCODE's sex-averaged map was downloaded from the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables; Karolchik, 2004); accessing the 'decodeSexAveraged' table under the group 'All Tables' under the 'hg19' database. The downloaded list consisted of 244,308 non-overlapping bins. Each bin is given a standardised recombination rate (SRR) that indicates the tendency of recombination in a specific 10 Kb genomic region. Bins with an SRR of 0 represent recombination cold-spots while bins with an SRR of 10 or above represent recombination hotspots.

The nearest recombination rate for each RT insertion was identified using the closest function of BEDTools (Quinlan, 2014) version 2.25.0. The expected number of RTEs in each recombination region was determined by calculating the number of elements that would be present in each region in relation to the size of the region. Specifically, if half of the human genome is non-recombining, the expected number of RTEs in non-recombining regions would be 50%. Fisher's exact tests were carried out to compare the distribution of reference and non-reference RTEs in non-recombining and recombination regions of the genome. The distribution of each RTE in regions of different recombination rates was compared with the expected distribution using the chi-squared goodness of fit test. The codes for the local recombination rate analyses are provided in GitHub at: https://github.com/RandaAli1/MyPhDproject/tree/master/MyAnalysis.

3.2.2.5. RTEs Enrichment in functional regions and accessible chromatin domains

Roadmap segmented the human genome into various regulatory classes, reflecting different degrees and types of regulatory activity (Roadmap Epigenomics Consortium, 2015). This study took advantage of this classification to define accessible chromatin regions using the H3K4me3 profiles associated with euchromatin domains. Roadmap-annotated BED files for each of the 127 cell types, across 30 types of human tissues, were downloaded (https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak/; last modified: October 2013). One file per cell type was downloaded using the epigenome identifier (EID) (EID-H3K4me3.broadPeak.gz). Each file was decompressed using the gunzip command in UNIX. An additional column was added to each file containing the associated EID identifier. The resulting 127 files were then grouped into one file using the cat command in UNIX. Broad peak calls were used since genomic regions bound with chromatin domains can be very wide.

A control database, known as the Random Database, was generated to test whether the associations observed between the genomic locations of RTEs with functional regions were random or based on the composition of the genomic features. The control database was also used to investigate whether RTEs are enriched in euchromatin domains in any cell type or tissue. The enrichment of non-reference RTEs in the euchromatin domain of any cell type or tissue may potentially affect its function, as active retrotransposition is more likely to occur in accessible chromatin regions. To make the random dataset, the random function of BedTools (Quinlan, 2014) version 2.25.0 was used to generate a random set of intervals in a Bed file format. For each RTE the random dataset

was subsampled based on the non-reference curated database, including 1000 x the size of each database; 10,211 L1Hs; 27,699 AluYs and 1,888 SVAs.

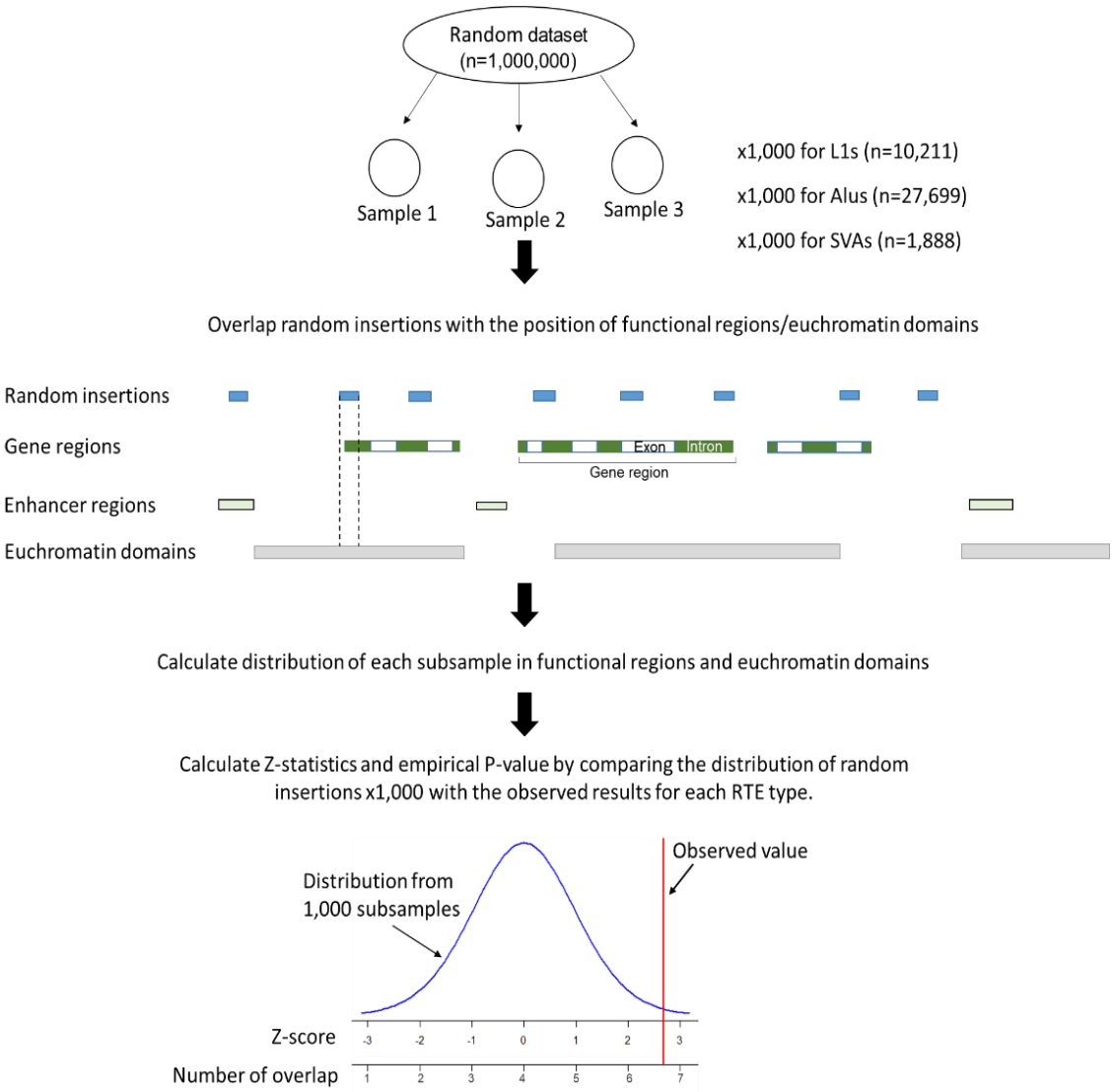


Figure 8: Schematic representation of RTEs enrichment in functional and accessible chromatin domains. A random dataset is generated using BEDtools. First, a samples is extracted based on the size of each of the non-reference curated database. Next, the random sample is overlapped with functional genomic regions including genic and enhancer regions plus accessible chromatin (euchromatin) domains. The dotted line shows a random insertion overlapping with an exon region (green) and is located in euchromatin domain. The process of subsampling and calculating the distribution of random insertions in functional regions and euchromatin domains is reiterated 1000x. Finally, the observed value is compared with the distribution of the 1000 samples using Z-statistics for the enrichment of RTEs in functional regions, or using the empirical P-value for the enrichment of RTEs in euchromatin domains.

The distribution of the random datasets in functional and accessible chromatin regions was determined by overlapping the locations of RTEs with the RefSeq genes, GeneHancer, and Roadmap files using the intersect function of bedtools (Quinlan, 2014) version 2.25.0. The mean overlap with functional regions and standard deviation of the random datasets were identified using the summary () function in R (R Core Team, 2012; V.3.4.0). These statistics were then used to calculate the Z-statistics (McLeod, 2019) for the overrepresentation of RTEs in functional regions including intronic, exonic, and enhancer regions. The enrichment of RTEs in euchromatin domains was calculated using the Empirical P Value (North et al., 2002) calculated as:

$$(r + 1) \div (n + 1)$$

Where r is the number of replicates that produce a number of overlap greater than or equal to the number of overlaps observed for RTEs in euchromatin regions, and n is 1,000 representing the number of replicate samples that have been created. The enrichment analysis is illustrated in figure 8. The codes for the enrichment analyses are provided in GitHub at: https://github.com/RandaAli1/MyPhDproject/tree/master/MyAnalysis.

3.3. Results

3.3.1. Chromosomal distribution

The distribution of the three classes of RTEs (L1s, Alus, SVAs) across the chromosomes and genome was studied, comparing number of RTEs, and both chromosome size and the gene density. Both reference and non-reference RTEs were studied separately in order to understand how natural selection forces play a part in the distribution over time.

Table 9: Count of reference and non-Reference LINE 1s, Alus, and SVAs per chromosome, displaying the percentage of the whole genome. Non-reference RTEs were taken from the curated RTE databases. Abbreviation: Chr = Chromosome.

	LINEs 1		Alus		SVAs	
Chr	Reference L1s/Chr (%)	Non-Reference L1s/ Chr (%)	Non-Reference Alus/ Chr (%)	Reference Alus/Chr (%)	Non-Reference SVAs/ Chr (%)	Reference SVAs/Chr (%)
1	2641 (6.88)	751 (7.35)	27842 (9.05)	2099 (7.58)	94 (9.35)	204 (10.81)
2	2983 (7.78)	849 (8.31)	21589 (7.02)	2320 (8.38)	66 (6.57)	151 (8.00)
3	2780 (7.25)	726 (7.11)	17233 (5.60)	1947 (7.03)	61 (6.07)	109 (5.77)
4	3035 (7.91)	808 (7.91)	13871 (4.51)	1957 (7.07)	58 (5.77)	81 (4.29)
5	2807 (7.32)	675 (6.61)	14194 (4.61)	1773 (6.40)	58 (5.77)	101 (5.35)
6	2334 (6.08)	654 (6.40)	14775 (4.80)	1814 (6.55)	62 (6.17)	108 (5.72)
7	1915 (4.99)	590 (5.78)	19350 (2.29)	1672 (6.04)	45 (4.48)	88 (4.66)
8	2097 (5.47)	512 (5.01)	13267 (4.31)	1406 (5.08)	41 (4.08)	70 (3.71)
9	1578 (4.11)	427 (4.18)	11827 (3.84)	1138 (4.11)	44 (4.38)	71 (3.76)
10	1500 (3.91)	456 (4.47)	15407 (5.01)	1278 (4.61)	54 (5.37)	98 (5.19)
11	1990 (5.19)	541 (5.30)	12709 (4.13)	1261 (4.55)	49 (4.88)	103 (5.46)
12	1759 (4.58)	472 (4.62)	16451 (5.35)	1393 (5.03)	51 (5.07)	92 (4.87)
13	1214 (3.16)	374 (3.66)	7134 (2.32)	1058 (3.82)	22 (2.19)	53 (2.81)
14	1143 (2.98)	347 (3.40)	9780 (3.18)	871 (3.14)	36 (3.58)	69 (3.65)
15	849 (2.21)	264 (2.59)	10034 (3.26)	825 (2.98)	24 (2.39)	75 (3.97)
16	650 (1.69)	264 (2.59)	14686 (4.77)	675 (2.44)	23 (2.29)	63 (3.34)
17	494 (1.29)	197 (1.93)	14131 (4.59)	769 (2.78)	36 (3.58)	92 (4.87)
18	908 (2.37)	265 (2.60)	6543 (2.13)	802 (2.89)	26 (2.59)	25 (1.32)
20	535 (1.39)	182 (1.78)	7859 (2.55)	522 (1.88)	20 (1.99)	68 (3.60)
19	348 (0.91)	145 (1.42)	15069 (4.90)	542 (1.96)	40 (3.98)	86 (4.56)
22	184 (0.48)	51 (0.50)	5998 (1.95)	310 (1.12)	14 (1.39)	41 (2.17)
21	395 (1.03)	168 (1.65)	2960 (1.00)	446 (1.61)	10 (0.99)	13 (0.69)
X	3525 (9.19)	438 (4.29)	12562 (4.08)	727 (2.62)	62 (6.17)	25 (1.32)
Y	702 (1.83)	55 (0.54)	2341 (0.76)	94 (0.34)	9 (0.90)	2 (0.11)
Total	38,366	10,211	307,612	27,699	1,005	1,888

The counts and frequency for each of the three RTEs studied are indicated in table 9 for both the reference and non-reference elements across the human chromosomes. The autosomes are arranged in descending order of size. Based on the frequency of the RTEs for individual chromosomes, table 9 suggests that the frequency is not consistently proportional to the size of the chromosomes. For example, a larger proportion of reference and non-reference LINE elements are located on chromosome 4 than in the larger chromosomes 1, 2, and 3 (reference LINEs only). A larger proportion of reference Alu and SVA elements are located on chromosomes 16, 17 and 18 than found on larger chromosomes; whereas chromosome 17 carries more non-reference Alus and SVAs than the larger chromosome16.

3.3.1.1. Linear regression analysis of RTE and chromosome size or gene density

To understand the chromosomal distribution of RTEs, based on the size of the chromosome, and the effect of selection, a linear regression analysis was performed for each RTE, separated into age categories (reference and non-reference). The confident and prediction intervals were also computed. Data points within the confident interval are likely within the range of the true population mean with a 95% certainty, while the prediction interval provides a 95% likelihood estimate that the observed data points are the outcome of the population, as predicted by the regression model.

Overall, the regression analysis shows that L1s, Alus and SVAs are not distributed randomly between chromosomes. Although at the chromosomal level, the number of elements, both the reference (fixed) and non-reference

(polymorphic), is generally related to the chromosome size, with a positive correlation to the physical length of the chromosome. There is a strong linear correlation between the number of insertions and chromosomal size, particularly for the non-reference elements (except for non-reference SVAs), as shown in Figures 1, 2, and 3. When comparing the RTE and gene density, there is a marked difference between the RTE classes, demonstrating either negative correlation, positive correlation, or no correlation at all. In the subsequent sections this will be discussed in more detail.

3.3.1.1.1. L1 Chromosomal distribution:

Initial integration of L1s, as shown by the distribution of non-reference L1s, do tend to show a greater correlation with chromosome size in comparison to reference L1s, evidenced by a greater correlation coefficient (r^2 =0.81 & 0.93 respectively for reference and non-reference L1s). There are some exceptions, for example, chromosome 4 for which there is an overrepresentation of non-reference L1 elements made evident by its position just below the upper limit of the prediction interval. In addition, there is an enrichment of reference L1s on chromosome X (Figure 9.A). The weaker correlation of reference L1s with chromosome size compared to non-reference L1s is mainly due of the enrichment of reference L1s on the X-chromosome, which has been previously linked to their role in propagating the X-silencing signal. Conducting the regression analysis for reference L1s without the X-chromosome data point increases the correlation coefficient (r^2) and statistical significance (r^2 from 0.81 to 0.90 and the P-value from 1.76×10^{-9} to 4.97×10^{-12}, respectively for reference and reference minus the X chromosome). Extending the analysis to include gene and L1 density across each chromosome revealed a weak inverse correlation for reference L1s

(r^2=0.27, P-value = 9.38E-3). (Figure 9.B). However, no such correlation was observed between the non-reference L1s and the gene density.

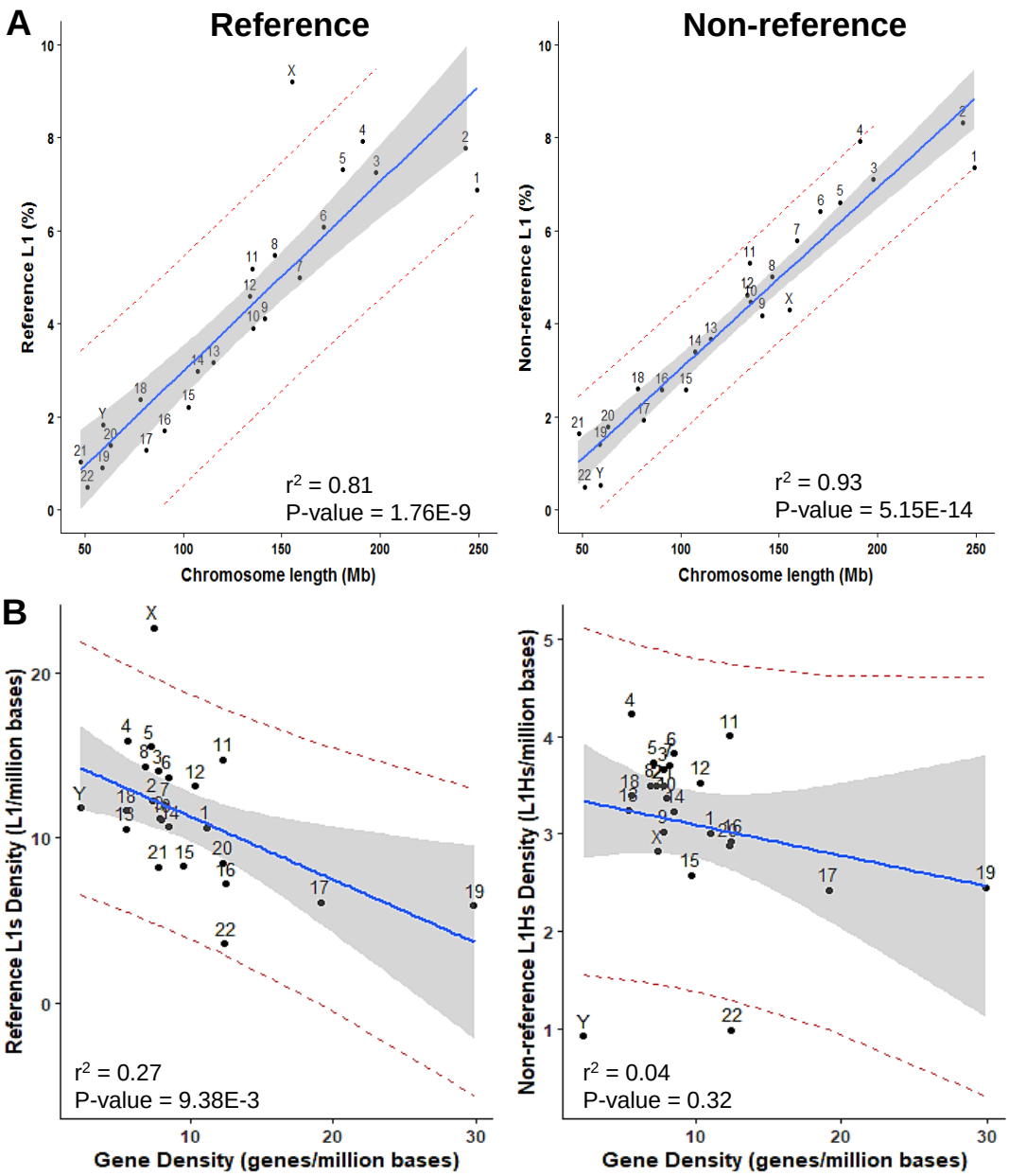


Figure 9: Scatter plots of the distribution of L1 elements across chromosomes and gene density. Elements counts as percentages of overall counts per chromosome are plotted against chromosomes size for reference and non-reference plots (left and right plots respectively) are shown in plots A. Plots B represent percentage of elements counts against gene density per chromosome. The fitted regression line is shown (blue line), along with 95% confidence intervals (grey cloud) and 95% prediction interval (red dotted line) for each scatter plot. Correlation between the count of elements and chromosome size, and element count and gene density are indicated within the plot, R^2 and P-value are given.

3.3.1.1.2. Alu Chromosomal distribution

Non-reference Alu elements appear to be distributed somewhat differently to the reference Alu elements. The initial integration of Alus, as shown by the distribution of non-reference Alus, resembles the initial integration of L1 elements. Non-reference Alus also tend to show a greater correlation with chromosome size (Figure 10.A) in comparison to reference Alus (r^2 = 0.64 & 0.89 respectively for reference and non-reference Alus). There are some exceptions, for example the X chromosome is depleted of non-reference Alus, evidenced by its position below the lower limit of the prediction interval (Figure 10.A). In addition, the correlation between the density of non-reference Alus and gene density is not significant. By contrast, there is a positive correlation between reference Alus density and gene density, suggesting that reference Alus are enriched in gene regions (Figure 10. B). These results reflect the enrichment of reference Alus on chromosome 19 more than expected for its physical length, as the density of reference Alus on chromosome 19 is proportional to its gene density. Reference Alus tend to show a greater correlation with gene density than chromosome size (r^2= 0.64 & 0.88 respectively for chromosome size and gene density). Chromosome 16 is an exception, where reference Alus are better correlated with its physical length than its gene density.

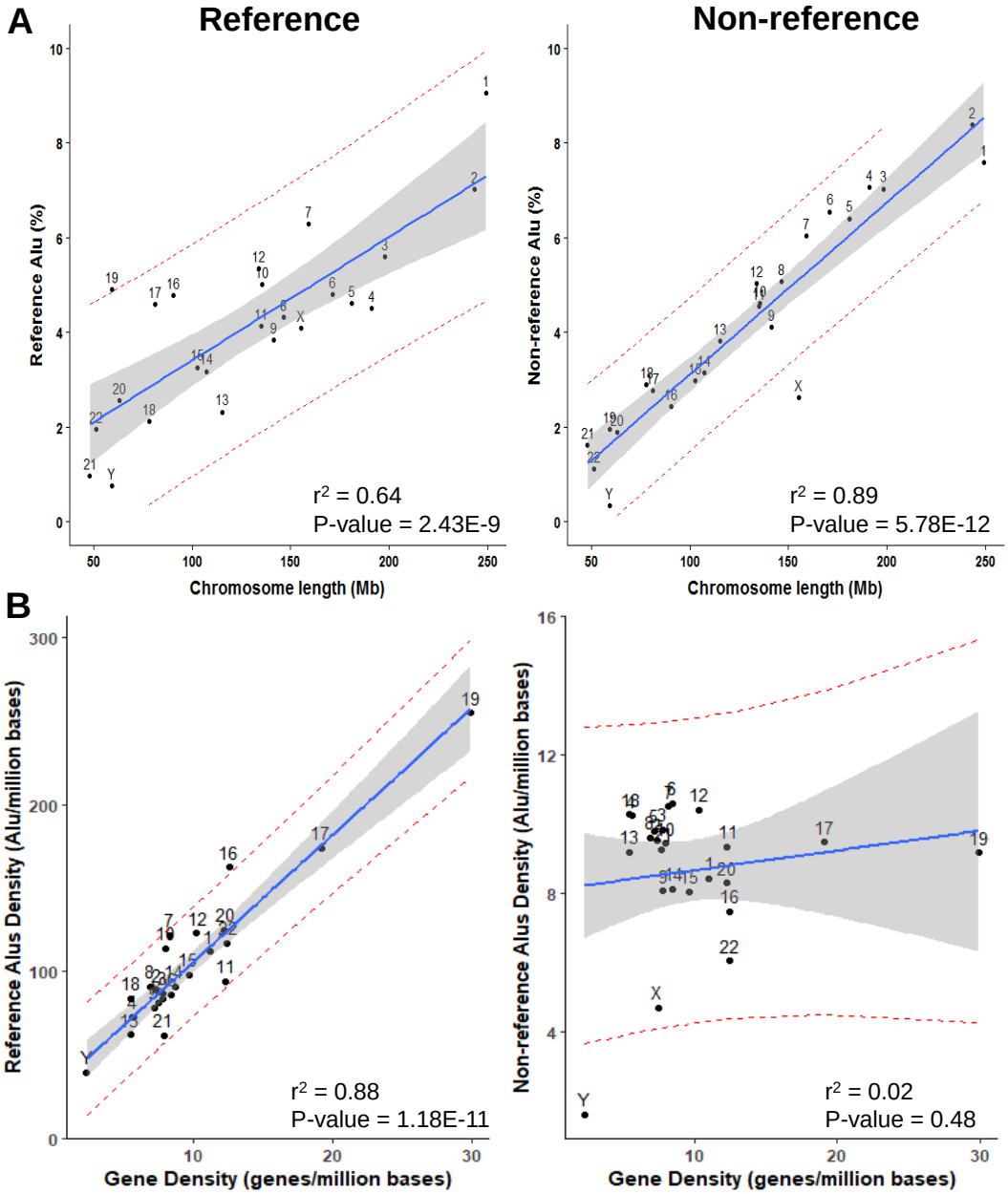


Figure 10: Scatter plots of the distribution of Alu elements across chromosomes and gene density. Elements counts as percentages of overall counts per chromosome are plotted against chromosomes size for reference and non-reference plots (left and right plots respectively) are shown in plots A. Plots B represent percentage of elements counts against gene density per chromosome. The fitted regression line is shown (blue line), along with 95% confidence intervals (grey cloud) and 95% prediction interval (red dotted line) for each scatter plot. Correlation between the count of elements and chromosome size, and element count and gene density are indicated within the plot, R^2 and P-value are given.

3.3.1.1.3. SVA Chromosomal Distribution

Unlike L1 and Alu elements, reference SVAs tend to show a stronger correlation with chromosome size in comparison to non-reference SVAs (r^2 = 0.82 & 0.58 respectively for reference and non-reference SVAs). There are some exceptions, for example chromosome 19, for which there is an enrichment of reference SVA elements evidenced by its position above the upper limit of the prediction interval (Figure 11.A). In addition, non-reference SVAs are depleted on chromosome X. There also tends to be a positive correlation with SVA density and gene density, suggesting that SVAs are enriched in gene regions (Figure 11.B). The initial integration of SVAs, as shown by the distribution of non-reference SVAs, does tend to show a greater correlation with gene density in comparison to reference SVAs (r^2 = 0.63 & 0.77 respectively for reference and non-reference SVAs). Chromosome X is an exception, as it is depleted of non-reference SVAs, evidenced by its position below the lower limit of the prediction interval (Figure 11.B). Nevertheless, non-reference SVAs tend to show a greater correlation with gene density than chromosome size (r^2 = 0.58 & 0.77 respectively for chromosome size and gene density), while the opposite trend is observed for reference SVAs. These results suggest that SVAs do integrate preferentially in gene-rich regions and reflect the combined role of chromosome size and gene density in shaping the chromosomal distribution of reference SVAs. As such, recent SVA insertions are likely to have the most negative impact on gene function and regulation, in contrast to L1s and Alus, which tend to avoid integrating into gene-rich regions.

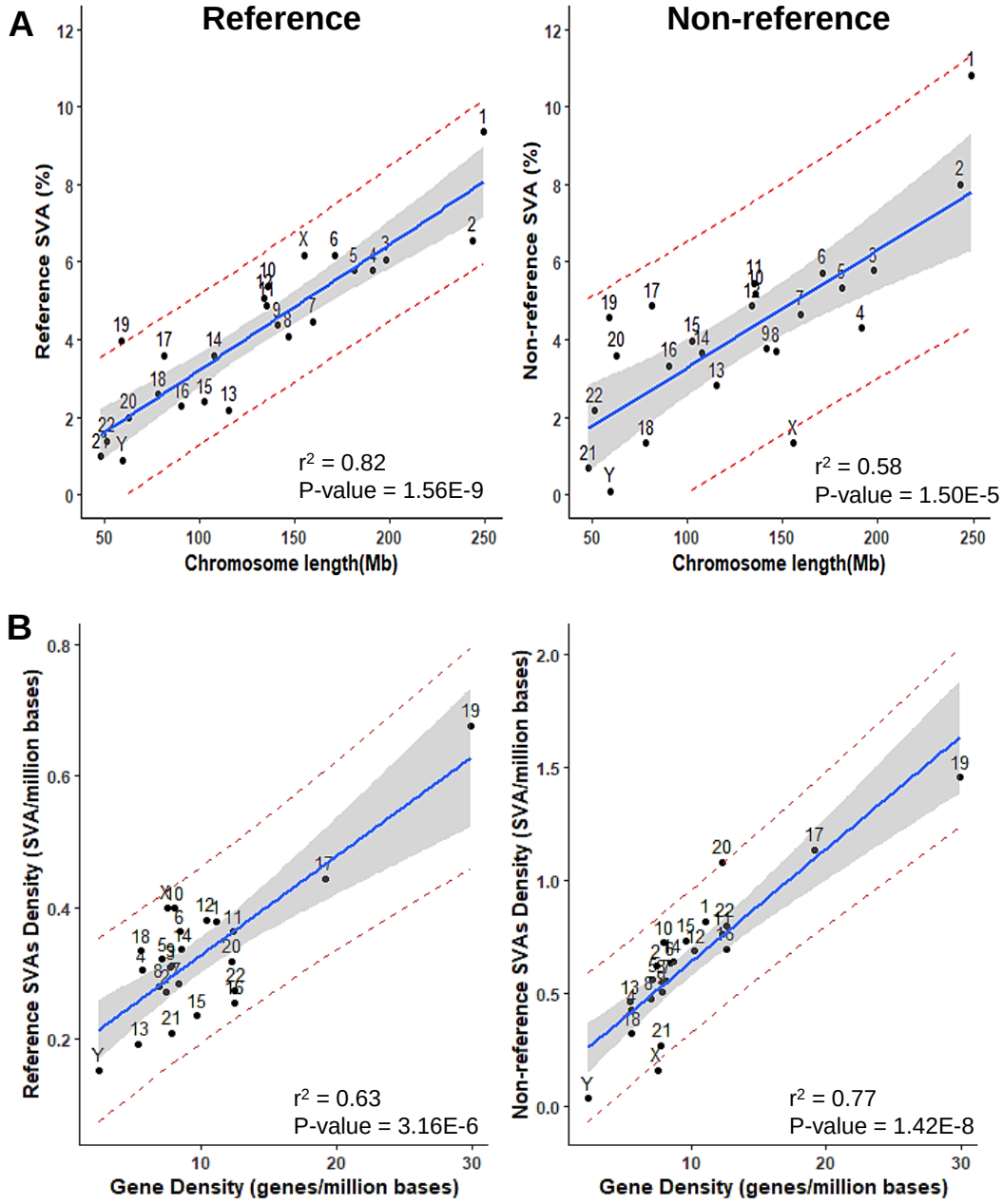


Figure 11: Scatter plots of the distribution of SVA elements across chromosomes and gene density. Elements counts as percentages of overall counts per chromosome are plotted against chromosomes size for reference and non-reference plots (left and right plots respectively) are shown in plots A. Plots B represent percentage of elements counts against gene density per chromosome. The fitted regression line is shown (blue line), along with 95% confidence intervals (grey cloud) and 95% prediction interval (red dotted line) for each scatter plot. Correlation between the count of elements and chromosome size, and element count and gene density are indicated within the plot, R^2 and P-value are given.

3.3.2. Local GC content

To understand the local base-composition of RTEs, the GC content for each RTE separated into age categories (reference and non-reference) was studied, alongside the genome GC content in 20 Kb windows. Genomic regions of higher GC content tend to have a higher gene density and recombination rate (Fullerton et al., 2001; Lander et al., 2001). The GC content was calculated to determine if the integration of RTEs is random within the chromosomes, or have been shaped by factors such as selection and/or preferential integration, that may have been driven by genomic regions of specific base-composition.

L1s, Alus and SVAs are not distributed randomly within different GC regions. Non-reference RTEs do show a tendency to accumulate in higher GC regions than reference elements, except for Alus. The distribution of L1s and SVA elements generally shifts towards lower GC bins with time, whilst the distribution of Alu elements shows the opposite trend and shift towards higher GC bins.

3.3.2.1. L1 GC content

Reference and non-reference L1 elements, at the genome level, appear to cluster in low GC regions, with maximum densities in the bin corresponding to 36-38% GC content. The overall accumulation of L1s in low GC regions is supported by the position of their frequency distribution curves being to the left of the genome curve, and their average GC content being below the genome-wide average of 41% (average GC content of 38% & 39% respectively for reference and non-reference L1s). Although L1s generally accumulate in low GC regions, non-reference L1s do tend to accumulate in higher GC regions than reference L1s, made evident by the position of its frequency distribution curve to the right of the

distribution curve of reference L1s. In addition, the non-reference distribution peak is below the peak of reference L1 elements (Figure 12). The statistical significance of the reported observations was confirmed using the two-sample K-S test, as shown in figure 12.

The K-S test returns a D-statistic that varies between 0 and 1 (Lall, 2015). The closer the D-statistic of the K-S test is to 0, the more similar the underlying distributions of the two samples are to each other. That is because a D value of 0 occurs when there is no difference between the cumulative distribution functions of the two samples. The D-statistic for the distribution of reference L1 elements vs. the genome is greater in comparison to the distribution of non-reference L1s. These results suggest that the initial integration of L1s, as shown by the distribution of non-reference L1s, does tend to be more uniform in the genome in comparison to reference L1s. In addition, the lowest D-statistic is observed for the K-S test between the distributions of reference vs. non-reference L1 elements, supporting the comparable accumulation of L1 elements in GC-poor regions.

3.3.2.2. Alu GC content

Non-reference Alu elements appear to accumulate in different GC regions to the reference Alus. The initial integration of Alu elements with respect to local base-composition, as shown by the distribution of non-reference Alus (Figure 13), is similar to the distribution of L1 elements. Although the frequency curve of non-reference Alus resembles the distribution curve of the genome, non-reference Alus appear to cluster in low GC regions with maximum density in the 36-38% GC bin. This observation is supported by the position of the frequency distribution curve of non-reference Alus being slightly to the left of the genome curve, and

their average GC content being below the genome-wide average of 41% (average GC content of non-reference Alus is 40%). In addition, the D-statistic of the K-S test is lowest for the K-S test between the distributions of non-reference Alus and the genome.

In contrast, reference Alus appear to cluster in genomic regions of higher GC content, with maximum density in the 40-42% GC bin. The frequency distribution curve of reference Alus is skewed to the right and is positioned to the right of the genome curve. In addition, the average GC content of reference Alus is above the genome-wide average (average GC content of reference Alus is 43%). These results indicate a tendency of Alu elements to shift towards regions of higher GC content over time. The D-statistic is highest for the K-S test between the distributions of reference vs. non-reference Alus, supporting the marked difference in the local base-composition of Alu elements of different evolutionary age.

3.3.2.3. SVA GC content

Reference and non-reference SVA elements, at the genome level, appear to cluster in higher GC regions with maximum densities in the 40-42% GC bin (Figure 14). The overall accumulation of SVA elements in higher GC regions is supported by the position of their frequency distribution curves being to the right of the genome curve, and their average GC content being above the genome-wide average (average GC content of 42% & 43% respectively for reference and non-reference SVAs). In addition, the lowest D-statistic is observed for the K-S test between the distributions of reference vs. non-reference SVA elements, supporting their comparable accumulation in GC-rich regions. Although SVAs are generally found in higher GC regions, non-reference SVAs do tend to accumulate

in higher GC regions, more so than reference SVAs, made evident by the frequency distribution curve of non-reference SVAs being positioned to the right of the curve of reference SVAs. In addition, the D-statistic is greatest for the K-S test between the distributions of non-reference SVAs vs. the genome. These results suggest that the initial integration of SVA elements, as shown by the distribution of non-reference SVAs, do tend to prefer higher GC regions of the genome.

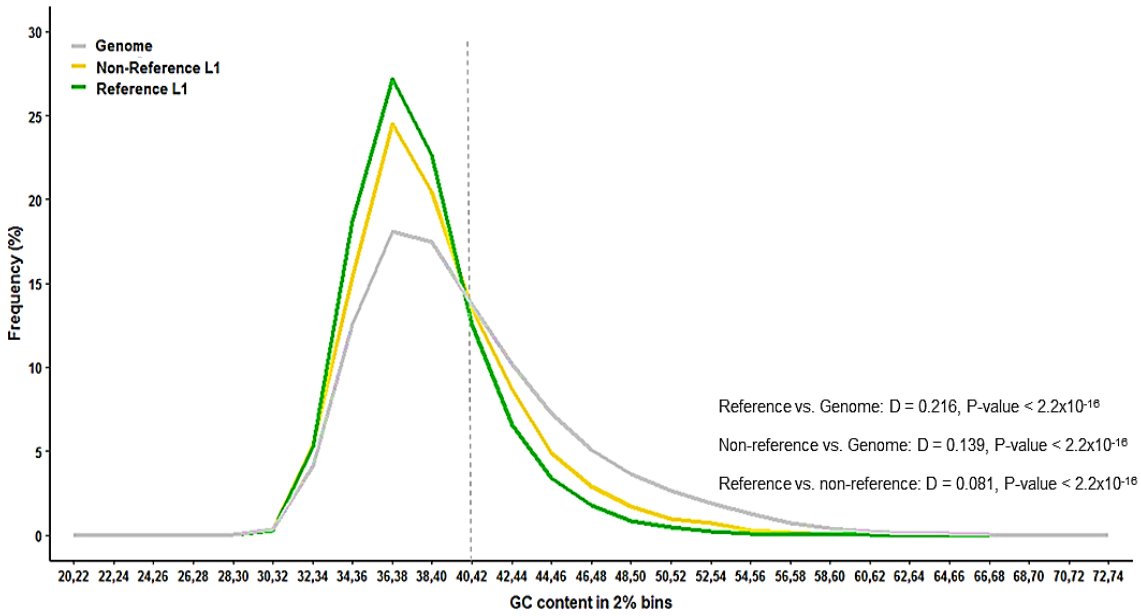


Figure 12: Frequency distribution of L1 elements in different GC fractions of the human genome. The GC content around each element is calculated over 20 Kb window (10 Kb on each side of the insertion). The x-axis from left to right correspond to increasing 2% GC content, and the frequency of L1s is show on the y-axis. The genome GC content was calculated by dividing the genome into 20 Kb windows from start to end. The dotted vertical line represents the position of the average genome-wide GC content of 41%. K-S test statistics are indicated within the plot, D statistics and P-value are given.

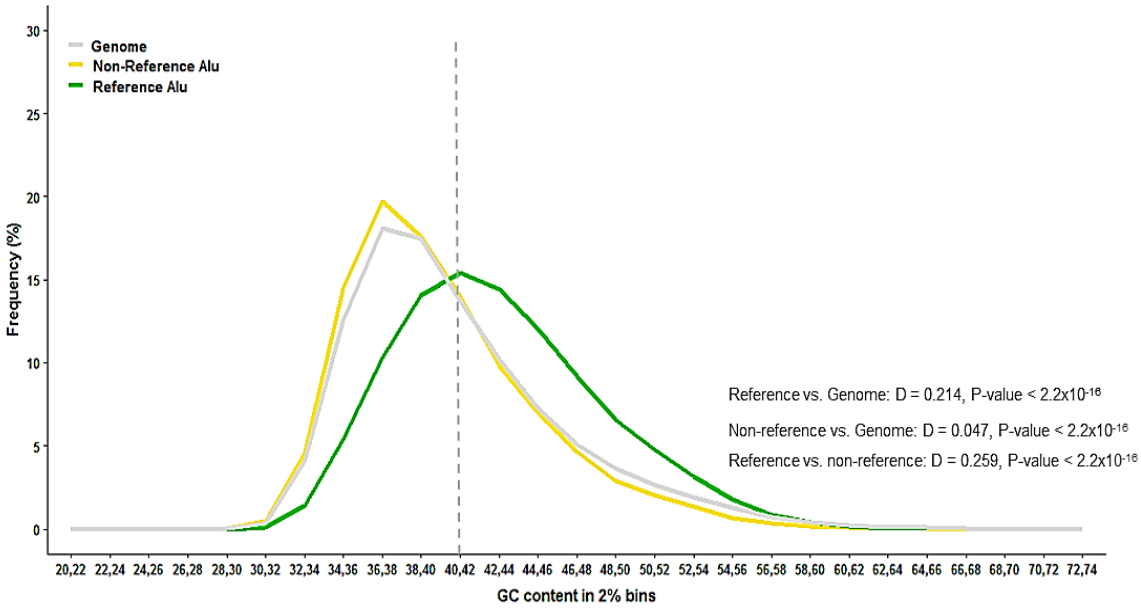


Figure 13: Frequency distribution of Alu elements in different GC fractions of the human genome. The GC content around each element is calculated over 20 Kb window (10 Kb on each side of the insertion). The x-axis from left to right correspond to increasing 2% GC content, and the frequency of L1s is show on the y-axis. The genome GC content was calculated by dividing the genome into 20 Kb windows from start to end. The dotted vertical line represents the position of the average genome-wide GC content of 41%. K-S test statistics are indicated within the plot, D statistics and P-value are given.

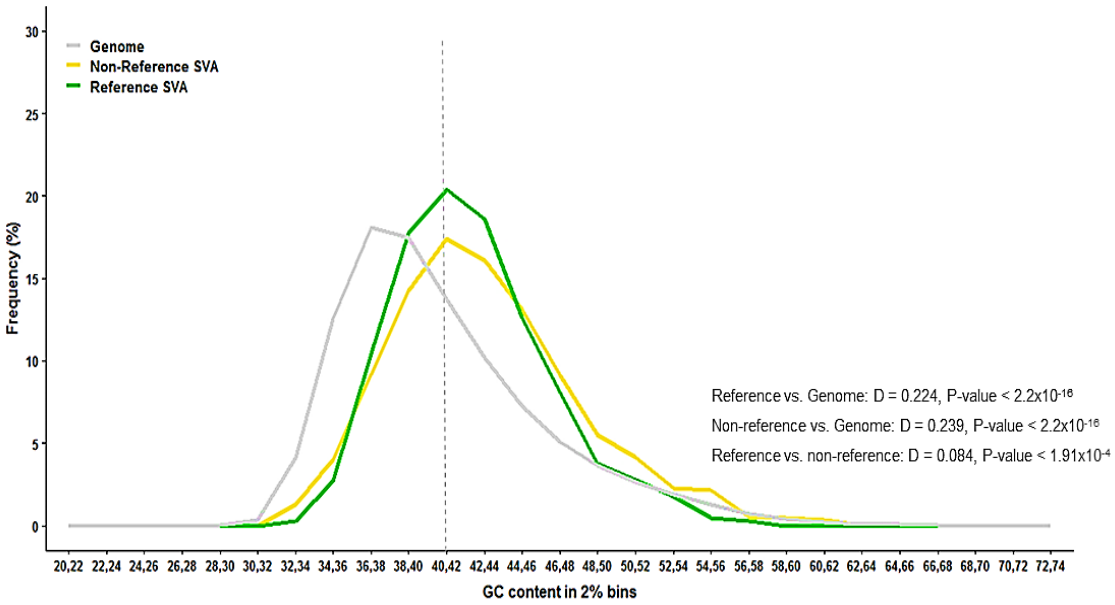


Figure 14: Frequency distribution of SVA elements in different GC fractions of the human genome. The GC content around each element is calculated over 20 Kb window (10 Kb on each side of the insertion). The x-axis from left to right correspond to increasing 2% GC content, and the frequency of L1s is show on the y-axis. The genome GC content was calculated by dividing the genome into 20 Kb windows from start to end. The dotted vertical line represents the position of the average genome-wide GC content of 41%. K-S test statistics are indicated within the plot, D statistics and P-value are given.

3.3.3. Distribution in functional regions

RTE integration in functional regions has the potential to alter gene expression, splicing, or dosage. The distribution of each RTE in gene and enhancer regions was studied to determine whether the integration of RTEs is random, or if it is biased by factors that favour or restrict its integration, such as selection and/or preferential integration.

L1s, Alus, and SVAs are not randomly distributed in functional genomic regions. All RTE classes show a significant difference in the distribution of their reference vs. non-reference elements across the functional genomic regions analysed, as confirmed by Fisher's exact test (Tables 10 & 11). Although RTEs are generally

less frequent in functional regions, non-reference RTEs do show a greater tendency to accumulate in intragenic (within genes) and enhancer regions than reference elements, with the exception of Alu elements (Figures 15 & 16). Overall, the distribution of L1 and SVA elements generally shifts away from functional regions over time, while the opposite trend is observed for the distribution of Alu elements.

3.3.3.1. L1 distribution in functional regions

L1 elements, from both age categories, are more frequent in intergenic regions than expected by chance alone, made evident by their positive Z-statistics in intergenic regions (Table 12). The initial integration of L1s, as shown by the distribution of non-reference L1s, do tend to accumulate more frequently in functional genomic regions in comparison to reference L1s. Non-reference L1s are 1.19 and 2.83 times more frequent in intragenic and enhancer regions, respectively, than reference L1s (Tables 10 & 11; Figures 15 & 16). The distribution of L1 elements within gene regions (from transcription, start to end, including the 5' and 3' untranslated regions) show a greater accumulation of L1s in intronic regions, although it should be noted, more non-reference L1s, in terms of frequencies, are found within intronic and exonic regions in comparison to reference L1 elements (Figure 15). Still, L1 elements (in both age categories) are depleted in functional regions in comparison to the random insertion model, evidenced by their negative Z-score values in intronic, exonic, and enhancer regions (Table 12). These results suggest that, although L1 elements in general tend to be depleted in functional genomic regions, the initial integration of L1 elements in functional regions is more uniform in comparison to reference (fixed)

L1s. The distribution of L1 elements in functional regions is in line with their accumulation in low GC regions that tend to be less functional than GC-rich regions.

Table 10: Counts and percentages of RTEs located within intergenic and intragenic regions. Intragenic regions include the gene region from transcription start to end including the 5' and 3' untranslated regions of the NCBI RefSeq genes. Reference RTEs are fixed elements in the reference genome obtained from the RepeatMasker table of repeats. Non-reference elements are polymorphic insertions curated in-house from published studies. P-value and OR for Fisher's exact test statistics are included. Abbreviations: OR, Odds Ratio; CI, Confident interval.

	L1		**Alu**		**SVA**	
	Reference (%)	Non-Reference (%)	Reference (%)	Non-Reference (%)	Reference (%)	Non-Reference (%)
Intergenic	24,476 (63.80)	6,104 (59.78)	142,735 (46.40)	15,045 (54.32)	575 (57.21)	962 (50.95)
Intragenic	13,890 (36.20)	4,107 (40.22)	164,877 (53.60)	12,654 (45.68)	430 (42.79)	926 (49.05)
P-value	1.02E-13		1.06E-140		0.0013	
OR (95% CI)	1.19 (1.13-1.24)		0.73 (0.71-0.75)		1.29 (1.10-1.51)	

Table 11: Counts and percentages of RTEs located within enhancer regions of the GeneHancer database. Reference RTEs are fixed elements in the reference genome obtained from the RepeatMasker table of repeats. Non-reference elements are polymorphic insertions curated in-house from published studies. P-value and OR for Fisher's exact test statistics are included. Abbreviations: OR, Odds Ratio; CI, Confident interval.

	L1		**Alu**		**SVA**	
	Reference (%)	Non-Reference (%)	Reference (%)	Non-Reference (%)	Reference (%)	Non-Reference (%)
Enhancer	1,266 (3.30)	900 (8.81)	47,305 (15.38)	3,356 (12.12)	65 (6.47)	300 (15.89)
Non-enhancer	37,100 (96.70)	9,311 (91.19)	260,307 (84.62)	24,343 (87.88)	940 (93.53)	1,588 (84.11)
P-value	1.42E-108		3.05E-50		4.25E-14	
OR (95% CI)	2.83 (2.59-3.10)		0.76 (0.73-0.79)		2.73 (2.10-3.67)	

Table 12: Z-test statistics for the distribution of RTEs located within intergenic and intragenic regions of the NCBI RefSeq genes in comparison with a random database including 1,000x iterations. Z-test statistics for the distribution of RTEs in comparison with the random database in enhancer regions of the GeneHancer database are also given. Reference RTEs are fixed elements in the reference genome obtained from the RepeatMasker table of repeats. Non-reference elements are polymorphic insertions curated in-house from published studies.

	L1		**Alu**		**SVA**	
	Reference Z-score (Pval)	Non-Reference Z-score (Pval)	Reference Z-score (Pval)	Non-Reference Z-score (Pval)	Reference Z-score (Pval)	Non-Reference Z-score (Pval)
Intergenic	16.35 (4.10E-60)	7.60 (2.91E-14)	-40.74 (< 2.2E-16)	-8.01 (1.17E-15)	0.79 (0.43)	-4.54 (5.73E-6)
Intronic	-9.71 (2.63E-22)	-3.91 (9.31E-5)	50.77 (< 2.2E-16)	12.64 (1.22E-36)	1.81 (0.07)	5.95 (2.74E-9)
Exonic	-18.44 (6.05E-76)	-10.24 (1.27E-24)	-26.34 (6.15E-153)	-12.12 (8.17E-34)	-7.39 (1.47E-13)	-3.53 (4.17E-4)
Enhancer	-29.51 (1.09E-191)	-12.71 (5.53E-37)	14.38 (6.55E-47)	-4.44 (9.04E-6)	-8.47 (2.56E-17)	3.48 (5.03E-4)

3.3.3.2. Alu distribution in functional regions

Non-reference Alu elements are distributed differently in functional regions in comparison to the distribution reference Alus. This result was expected, since Alu elements from both age categories are also distributed differently in different GC regions. Alu elements, from both age categories, are less frequent in intergenic regions than expected by chance alone, made evident by their negative Z-statistics in intergenic regions (Table 12). In addition, the initial integration of Alus do tend to be less frequent in functional regions in comparison to reference Alus. Non-reference Alus are 0.73 and 0.76 times less common in intragenic and enhancer regions, respectively, than reference Alus (Tables 10 & 11; Figures 15 & 16). Although Alu elements, both reference and non-reference, overall do show a tendency to be overrepresented in intronic regions in comparison to the random distribution model, the significance level of accumulation is greater for reference Alus. In addition, reference Alus are significantly more frequent in enhancer

regions in comparison to random insertions, while non-reference Alus are significantly less frequent, evidenced by the positive and negative Z-score values for reference and non-reference Alus, respectively (Table 12). Alu elements are generally depleted in exon regions in comparison to random insertions, although it should be noted, more non-reference Alus, in terms of frequency, are found in exon regions. These results suggest that the initial integration of Alu elements, as shown by the distribution of non-reference Alus, tend to avoid functional genomic regions. However, post-integration factors such as selection can alter the relative distribution of Alus among genic and enhancer regions, increasing the frequency of Alus within functional regions with age.

3.3.3.3. SVA distribution in functional regions

Although SVA elements, from both age categories, are more frequent in intergenic regions than genic regions (Table 10), non-reference SVAs appear to be depleted in intergenic regions when compared with the distribution of the random dataset. In addition, the accumulation of reference SVAs in intergenic regions is not significantly different from the distribution of random insertions (Table 12). The initial integration of SVAs, as shown by the distribution of non-reference SVAs, do tend to accumulate more frequently in functional genomic regions in comparison to reference SVAs. Non-reference SVAs are 1.29 and 2.73 times more frequent in intragenic and enhancer regions in comparison to reference SVAs (Tables 10 & 11; Figures 15 & 16). Although SVA elements, both reference and non-reference, overall do show a tendency to accumulate more frequently in intronic regions in comparison with the random insertion model, the significance level of accumulation is greater for non-reference SVAs. In addition,

non-reference SVAs are significantly more frequent in enhancer regions than reference SVAs, when compared with the distribution of the random dataset (Table 12). SVA elements are generally depleted in exon regions in comparison to the random database, although it should be noted, more non-reference SVAs, in terms of frequency, are found in exon regions. These results suggest that SVA elements tend to integrate preferentially in functional genomic regions, as shown by significant accumulation of non-reference SVAs in intronic and enhancer regions, in comparison to the distribution of random insertions. However, post-integration processes such as selection can alter the relative distribution of SVAs within functional genomic regions, reducing the frequency of SVAs frequency within functional regions with age. The distribution of SVA elements in functional regions is in line with their accumulation in GC-rich regions that tend to be more functional than GC-poor regions. The distribution of all RTEs in functional regions, overall, suggests that non-reference SVA elements may have the strongest potential to affect functional genomic regions, due to its significant overrepresentation in both genes and enhancer regions in comparison to the other RTE types.

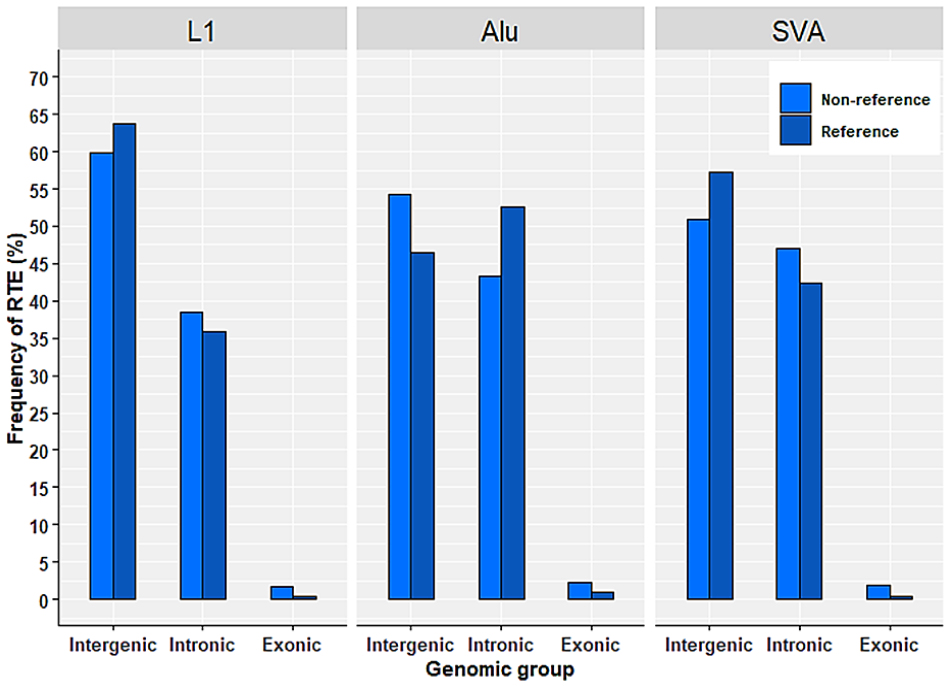


Figure 15: Frequency distribution (%) of RTEs located within intergenic and intragenic regions. Intragenic regions include intronic and exonic gene region from transcription start to end including the 5' and 3' untranslated regions of the NCBI RefSeq genes. Reference RTEs are fixed elements in the reference genome obtained from the RepeatMasker table of repeats. Non-reference elements are polymorphic insertions curated in-house from published studies.

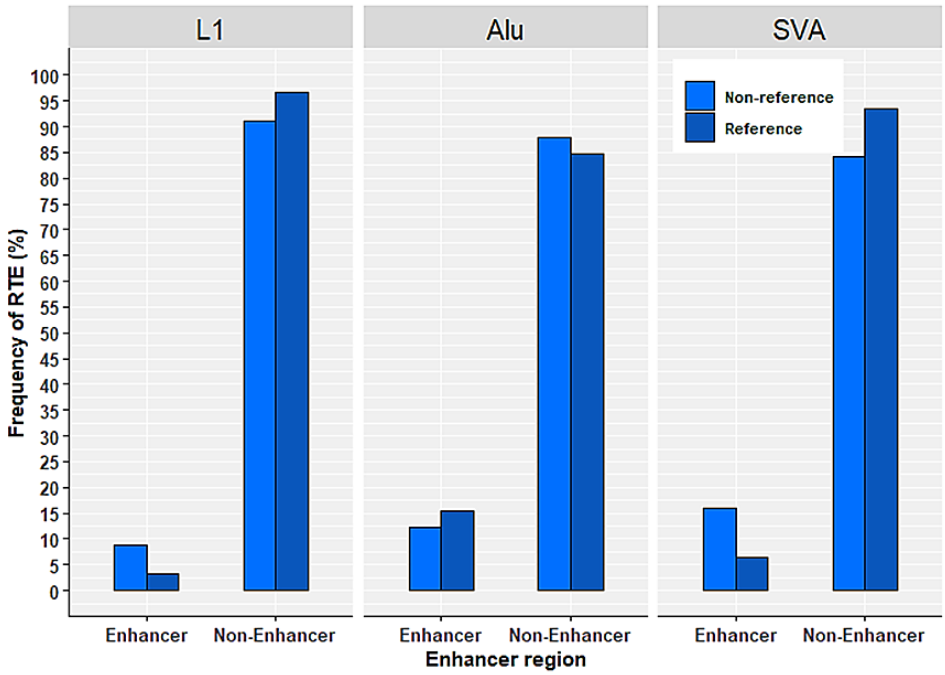


Figure 16: Frequency distribution (%) of RTEs located within enhancer regions of the GeneHancer database. Reference RTEs are fixed elements in the reference genome obtained from the RepeatMasker table of repeats. Non-reference elements are polymorphic insertions curated in-house from published studies.

3.3.4. Local recombination rate

The standardised sex-averaged (female and male) recombination map, described by Kong et al., (2010), was used to study the distribution of RTEs in regions of different recombination capacity. This recombination map provides a total of 244,308 bins, which were divided into three groups based on their standardised recombination rate (SRR): 104,488 bins (42.77%) are cold, 135,814 (55.59%) are intermediate, and 4006 (1.64%) are hot. Previous studies have reported that recombination rates tend to be highest in genomic regions surrounding genes. Owing to the high copy number and sequence homology between RTE elements, RTE integration in higher recombination regions has the potential to mediate significant deletions and duplications resulting from non-allelic homologous recombination events. Such deletions/duplications can result in the loss of function of many genes or an upset of the gene dosage balance. The distribution of each RTE, separated into age categories, in different recombination regions was studied to determine whether the integration of RTEs, with respect to local recombination rate, is random or whether it is biased by factors such as selection and/or preferential integration.

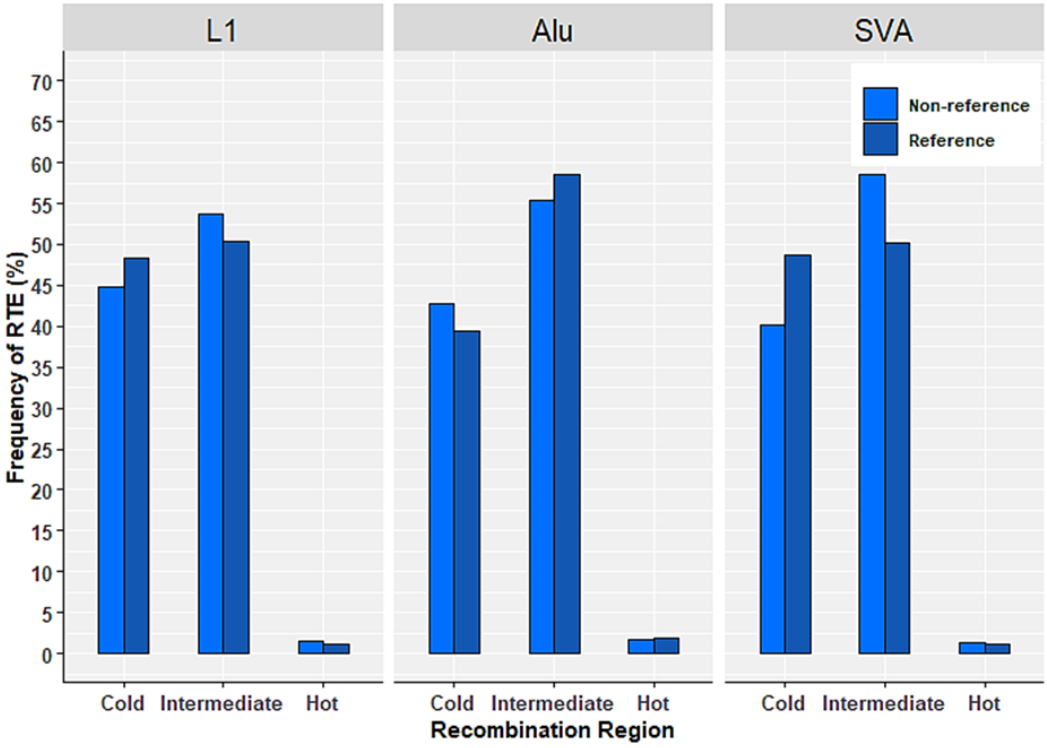


Figure 17: Frequency distribution (%) of RTEs located within different recombination regions using the standardised sex-averaged (female and male) recombination map described by Kong et al., (2010). Reference RTEs are fixed elements in the reference genome obtained from the RepeatMasker table of repeats. Non-reference elements are polymorphic insertions curated in-house from published studies.

Table 13: Counts and percentages of RTEs located within recombination regions using the standardised sex-averaged (female and male) recombination map described by Kong et al., (2010). Reference RTEs are fixed elements in the reference genome obtained from the RepeatMasker table of repeats. Non-reference elements are polymorphic insertions curated in-house from published studies. P-value and OR for Fisher's exact test statistics are included. Abbreviations: OR, Odds Ratio; CI, Confident interval.

	L1		**Alu**		**SVA**	
	Reference (%)	Non-Reference (%)	Reference (%)	Non-Reference (%)	Reference (%)	Non-Reference (%)
Non-recombining	16,517 (48.38)	4,354 (44.80)	115,431 (39.44)	11,496 (42.77)	454 (48.61)	747 (40.14)
Recombining	17,622 (51.62)	5,364 (55.20)	177,278 (60.56)	15,382 (57.23)	480 (51.39)	1,114 (59.86)
Total	34,139	9,718	292,709	26,878	934	1,861
P-value	4.71E-10		1.71E-26		2.11E-05	
OR (95% CI)	1.15 (1.10-1.21)		0.87 (0.85-0.89)		1.41 (1.20-1.66)	

Table 14: Counts and percentages of RTEs located within recombination regions using the standardised sex-averaged (female and male) recombination map described by Kong et al., (2010). Recombining regions are divided into intermediate and hot regions, defined by a standardized recombination rate (SRR) > 0 & <10 for intermediate regions, and SSR of 10 and above for hot recombination regions. Reference RTEs are fixed elements in the reference genome obtained from the RepeatMasker table of repeats. Non-reference elements are polymorphic insertions curated in-house from published studies. The random insertion model assumes that the fraction of insertions is proportional to the fractional size of each of the recombination regions. P-value for Chi-squared goodness of fit test statistics are given.

	L1		**Alu**		**SVA**	
	Reference (%)	Non-Reference (%)	Reference (%)	Non-Reference (%)	Reference (%)	Non-Reference (%)
Cold	16,517 (48.38)	4,354 (44.80)	115,431 (39.44)	11,496 (42.77)	454 (48.61)	747 (40.14)
Intermediate	17,219 (50.44)	5,224 (53.76)	171,626 (58.63)	14,907 (55.46)	469 (50.21)	1,088 (58.46)
Hot	403 (1.18)	140 (1.44)	5,652 (1.93)	475 (1.77)	11 (1.18)	26 (1.40)
P-value compared with random insertion model	2.99E-100	1.48E-04	1.06E-304	0.26	1.16E-03	0.040

The initial integration of RTE elements, with the exception of Alu elements, is not random with respect to local recombination rates. RTEs are found in all recombination regions, with the highest fraction observed in intermediate regions for all RTEs (Figure 17). In addition, all RTE classes show a significant difference in the distribution of their reference vs. non-reference elements, across the various recombination regions analysed, as confirmed by Fisher's exact test (Table 13). Although RTEs are generally more frequent in intermediate regions, non-reference RTEs do show a greater tendency to accumulate in recombining regions than reference elements, with the exception of Alus (Figure 17; Table 14). Overall, the distribution of L1 and SVA elements generally shifts away from recombining regions with time, while the opposite trend is observed for the distribution of Alu elements.

3.3.4.1. L1 local recombination rate

The initial integration of L1s, as shown by the distribution of non-reference L1s, tends to be more frequent in recombining regions. Non-reference L1s are 1.15 times more frequent in recombining regions in comparison to reference L1s (Table 13). The distribution of L1s within recombining regions shows a greater accumulation of L1s, both reference and non-reference, in intermediate regions in comparison to hot recombination regions (Figure 17; Table 14). Overall, L1 elements from both age categories are not randomly distributed within the different recombination regions analysed, as confirmed by the Chi-square goodness-of-fit test (Table 14). Although it should be noted, more L1 elements, in terms of frequencies, are found within non-recombining regions, in comparison to the fractional size of non-recombining bins in the genome (48.38% & 44.80% respectively for reference and non-reference L1s vs. 42.77% of non-recombining bins in the genome), suggesting the preferential integration of L1 elements in non-recombining regions. Overall, the initial integration of L1s in recombining regions is more uniform in comparison to reference (fixed) L1s. The distribution of L1 elements in recombination regions is in line with their preferential accumulation in GC-poor, non-functional genomic regions that tend to have a low recombination rate.

3.3.4.2. Alu local recombination rate

Non-reference Alu elements appear to be distributed differently to the reference Alu elements. This result was expected, since Alu elements from both age categories are also distributed differently in different GC and functional regions. The initial integration of Alus, as shown by the distribution of non-reference Alus, appears to be random. In comparison, the distribution of reference Alus is significantly different from the distribution expected by chance, as confirmed by the Chi-square goodness-of-fit test (Table 14). Although the distribution of Alus, both reference and non-reference, show a greater accumulation in recombining regions, non-reference Alus are 0.87 times less frequent in recombining regions than reference Alus (Table 13). The distribution of Alu elements within recombining regions shows a greater accumulation of Alus in regions of intermediate recombination rates, although it should be noted, more reference Alus, in terms of frequencies, are found within intermediate and hot regions in comparison to non-reference Alu elements (Table 14). In addition, the frequency of non-reference Alu elements in non-recombining regions (referred to as cold regions in the table) is higher in comparison to the frequency of reference Alus (Table 14). The increased accumulation of reference Alus in recombining regions is in line with their significant accumulation in GC-rich, gene-rich regions of the genome. These results suggest that the initial integration of Alus, as shown by the distribution of non-reference Alus, in regions of different recombination rates, tends to resemble a random integration pattern. It appears that post-integration factors, such as selection, can alter the relative distribution of Alus in different recombination regions, increasing their frequency within recombining regions with age.

3.3.4.3. SVA local recombination rate

Although SVA elements, from both age categories, are more abundant in recombining regions, non-reference SVAs are 1.41 times more frequent in recombining regions in comparison to reference SVAs (Table 13). The distribution of SVAs, both reference and non-reference, within recombining regions shows a greater accumulation in intermediate regions in comparison to hot recombination regions (Figure 17; Table 14). The distribution of SVA elements from both age categories, overall do appear to be significantly different from the distribution expected by chance, as confirmed by the Chi-square goodness-of-fit test (Table 14). Although it should be noted, more reference SVA elements, in terms of frequencies, are found within non-recombining regions in comparison to non-reference SVAs (48.61% & 40.14% respectively for reference and non-reference SVAs). In addition, reference SVA elements are more frequent in non-recombining regions in comparison to the fractional size of non-recombining bins in the genome. The increased frequency of reference SVA elements in non-recombining regions, in comparison to non-reference SVAs, suggests the role of post-integration processes in re-shaping the fractional distribution of SVA elements in different recombination regions. Overall, the initial integration of SVA elements in recombining regions, as shown by the distribution of non-reference SVAs, is more uniform in comparison to reference (fixed) SVAs. These results are in line with the reduced frequency of reference SVA elements in functional genomic regions, and accumulation in lower GC regions, in comparison to non-reference SVAs.

3.3.5. Local chromatin accessibility

This section analyses the enrichment of non-reference RTEs in accessible chromatin regions marked with the epigenetic modification H3K4me3, using 127 epigenomes divided into 30 anatomical regions, characterised by the Roadmap epigenomics project (Roadmap Epigenomics Consortium, 2015). The non-reference RTEs are the younger, potentially active RTEs that will have more of an effect in disrupting genome function in comparison to reference (fixed) RTEs that are no longer active. The observed density of non-reference RTEs in each of the 127 epigenomes was compared with the densities produced by 1,000 datasets that have been randomly generated. RTE elements overall appear to be enriched in the accessible chromatin domains of a wide variety of cell types and anatomical regions. Non-reference RTEs were significantly enriched (empirical P-values > 0.05) in a total of 103/127 cell types belonging to 27/30 unique anatomical groups (Table 15: Figure 18). L1s, Alus, and SVAs tend to show a greater enrichment in cell types belonging to the blood anatomical group. In addition, RTEs are enriched in the accessible domains of many types of epithelium cell groups, such as skin, heart, breast, and gastrointestinal. L1 elements are enriched in a narrower variety of cell types and anatomical groups (36 &16, respectively for cell types and anatomical groups) in comparison with Alu and SVA elements (Alu=65 & 22, SVA=68 & 25, respectively for cell types and anatomical groups). These results suggest that non-reference RTEs that remain active in the genome have the potential to negatively affect the function of a wide variety of cell types and organs. Alu and SVA elements have a greater potential to influence genome function in comparison to L1 elements.

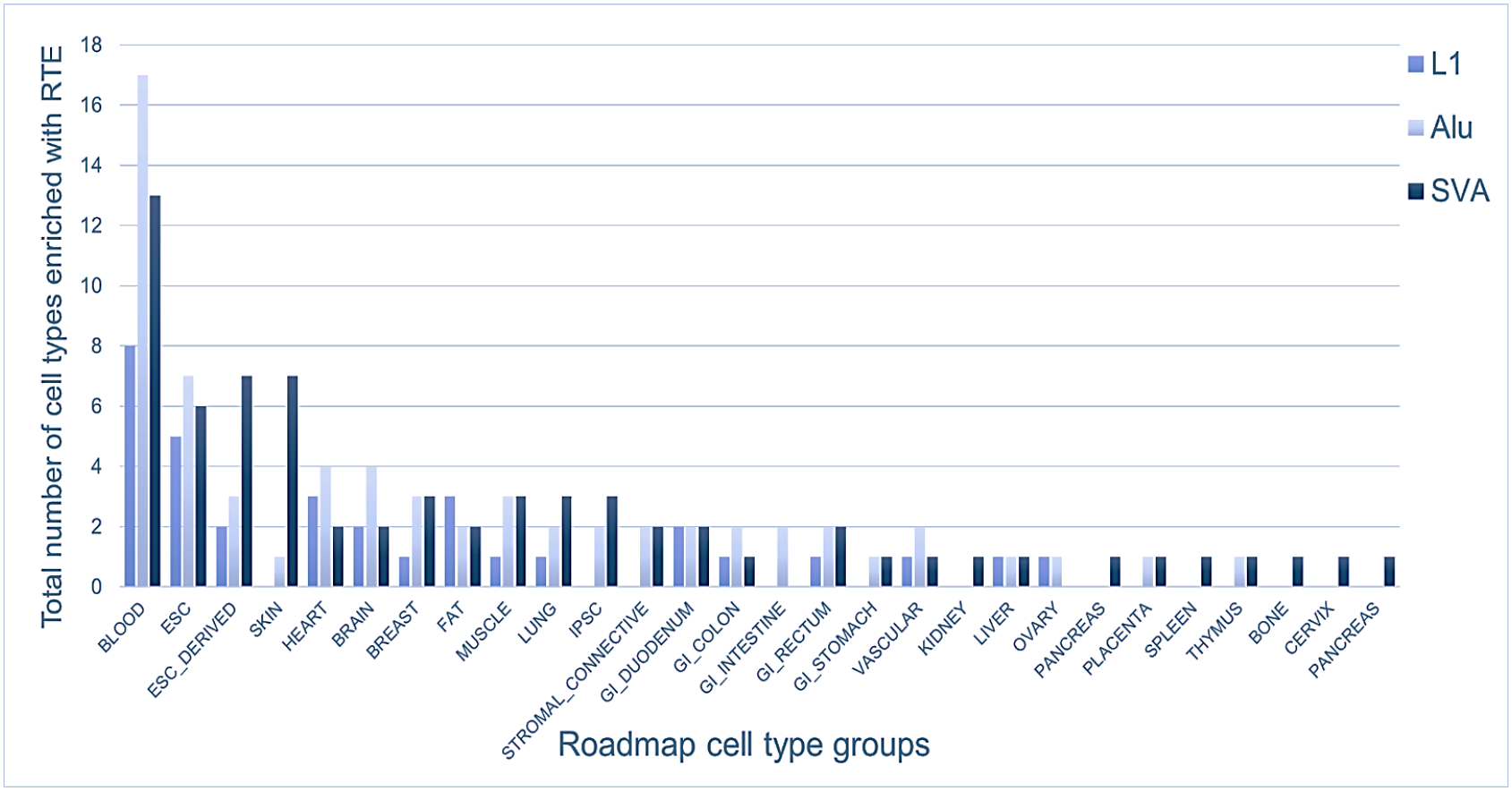


Figure 18: Count of non-reference (polymorphic) RTEs enriched in the euchromatin domains of at least one epigenome reported in the Roadmap project. Axes show anatomical groups (x-axis) against the count of cell types with an enrichment of RTEs in its euchromatin domains (y-axis). Abbreviations: ESC, Embryonic stem cells; IPSC, Induced pluripotent stem cells; GI, Gastrointestinal.

Table 15: Empirical P values for the enrichment of RTEs located within euchromatin domains using the H3K4me3 profiles of the Roadmap project. RTEs are polymorphic insertions curated in-house from published studies.

EID	P-value			Anatomy	Standardized Epigenome Name
	L1	Alu	SVA		
E001	1.0E-03	1.0E-03	1.0E-03	ESC	ESnaI3 Cells
E002	1.0E-03	1.0E-03	-	ESC	ESnaWA7 Cells
E003	-	2.7E-02	1.0E-03	ESC	H1 Cells
E004	-	1.0E-03	-	ESC_DERIVED	H1 BMP4 Derived Mesendoderm Cultured Cells
E005	-	-	1.0E-03	ESC_DERIVED	H1 BMP4 Derived Trophoblast Cultured Cells
E006	-	-	3.4E-02	ESC_DERIVED	H1 Derived Mesenchymal Stem Cells
E007	-	-	1.3E-02	ESC_DERIVED	H1 Derived Neuronal Progenitor Cultured Cells
E008	-	-	1.0E-03	ESC	H9 Cells
E009	-	-	2.9E-02	ESC_DERIVED	H9 Derived Neuronal Progenitor Cultured Cells
E010	-	-	2.8E-02	ESC_DERIVED	H9 Derived Neuron Cultured Cells
E011	1.0E-03	1.0E-03	1.0E-03	ESC_DERIVED	hESC Derived CD184+ Endoderm Cultured Cells
E012	1.0E-03	2.0E-03	-	ESC_DERIVED	hESC Derived CD56+ Ectoderm Cultured Cells
E013	-	-	2.0E-03	ESC_DERIVED	hESC Derived CD56+ Mesoderm Cultured Cells
E014	1.0E-03	1.0E-03	1.0E-03	ESC	HUES48 Cells
E015	1.0E-03	1.0E-03	1.0E-03	ESC	HUES6 Cells
E016	1.0E-03	1.0E-03	1.0E-03	ESC	HUES64 Cells
E017	-	-	1.0E-03	LUNG	IMR90 fetal lung fibroblasts Cell Line
E018	5.0E-03	-	1.0E-03	IPSC	iPSna15b Cells
E019	1.0E-03	1.6E-02	4.0E-03	IPSC	iPSna18 Cells
E020	1.0E-03	1.0E-03	2.0E-03	IPSC	iPSna20b Cells
E023	1.0E-03	1.0E-03	1.0E-03	FAT	Mesenchymal Stem Cell Derived Adipocyte Cultured Cells

(Table 15 continues)

E024	-	1.0E-03	-	ESC	ESnaUCSF4 Cells
E025	1.0E-03	1.0E-03	2.2E-02	FAT	Adipose Derived Mesenchymal Stem Cell Cultured Cells
E026	-	1.3E-02	3.0E-03	STROMAL_CONNECTIVE	Bone Marrow Derived Cultured Mesenchymal Stem Cells
E027	-	2.2E-02	1.0E-03	BREAST	Breast Myoepithelial Primary Cells
E028	5.0E-03	1.0E-03	1.0E-03	BREAST	Breast variant Human Mammary Epithelial Cells (vHMEC)
E030	-	-	3.0E-03	BLOOD	Primary neutrophils from peripheral blood
E031	-	-	1.4E-02	BLOOD	Primary B cells from cord blood
E032	-	1.2E-02	-	BLOOD	Primary B cells from peripheral blood
E033	-	1.0E-03	-	BLOOD	Primary T cells from cord blood
E035	-	3.0E-02	1.0E-03	BLOOD	Primary hematopoietic stem cells
E036	-	-	6.0E-03	BLOOD	Primary hematopoietic stem cells short term culture
E037	1.0E-03	1.0E-03	-	BLOOD	Primary T helper memory cells from peripheral blood 2
E038	1.0E-03	1.0E-03	-	BLOOD	Primary T helper naive cells from peripheral blood
E039	-	4.0E-03	1.0E-03	BLOOD	Primary T helper naive cells from peripheral blood
E040	1.0E-03	1.0E-03	-	BLOOD	Primary T helper memory cells from peripheral blood 1
E041	1.0E-03	1.0E-03	9.0E-03	BLOOD	Primary T helper cells PMAnaI stimulated
E042	5.0E-02	1.0E-03	-	BLOOD	Primary T helper 17 cells PMAnaI stimulated
E043	-	-	1.0E-03	BLOOD	Primary T helper cells from peripheral blood
E044	-	-	1.0E-03	BLOOD	Primary T regulatory cells from peripheral blood
E045	1.0E-03	1.0E-03	-	BLOOD	Primary T cells effector/memory enriched from peripheral blood
E046	-	1.0E-03	-	BLOOD	Primary natural Killer cells from peripheral blood
E047	-	1.0E-03	-	BLOOD	Primary T CD8+ naive cells from peripheral blood
E048	2.0E-03	1.0E-03	-	BLOOD	Primary T CD8+ memory cells from peripheral blood
E049	-	1.7E-02	6.0E-03	STROMAL_CONNECTIVE	Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells
E051	-	-	2.0E-03	BLOOD	Primary hematopoietic stem cells GnaCSFnamobilized Male
E052	-	1.1E-02	3.7E-02	MUSCLE	Muscle Satellite Cultured Cells
E055	-	-	1.0E-03	SKIN	Foreskin Fibroblast Primary Cells skin01

(Table 15 continues)

E056	-	-	1.0E-03	SKIN	Foreskin Fibroblast Primary Cells skin02
E057	-	-	8.0E-03	SKIN	Foreskin Keratinocyte Primary Cells skin02
E058	-	-	1.0E-03	SKIN	Foreskin Keratinocyte Primary Cells skin03
E059	-	-	1.0E-03	SKIN	Foreskin Melanocyte Primary Cells skin01
E061	-	-	1.0E-03	SKIN	Foreskin Melanocyte Primary Cells skin03
E062	-	1.0E-03	-	BLOOD	Primary mononuclear cells from peripheral blood
E063	1.0E-03	-	-	FAT	Adipose Nuclei
E065	1.0E-03	1.0E-03	-	VASCULAR	Aorta
E066	1.0E-03	1.0E-03	-	LIVER	Liver
E067	-	1.0E-03	-	BRAIN	Brain Angular Gyrus
E070	-	-	4.6E-02	BRAIN	Brain Germinal Matrix
E074	-	2.0E-03	-	BRAIN	Brain Substantia Nigra
E075	-	-	6.0E-03	GI_COLON	Colonic Mucosa
E076	-	1.0E-03	-	GI_COLON	Colon Smooth Muscle
E077	1.7E-02	1.0E-02	3.0E-03	GI_DUODENUM	Duodenum Mucosa
E078	1.0E-03	1.0E-03	3.0E-03	GI_DUODENUM	Duodenum Smooth Muscle
E081	8.0E-03	1.0E-03	-	BRAIN	Fetal Brain Male
E082	-	-	1.1E-02	BRAIN	Fetal Brain Female
E083	1.0E-03	1.0E-03	-	HEART	Fetal Heart
E084	-	4.0E-03	-	GI_INTESTINE	Fetal Intestine Large
E086	-	-	7.0E-03	KIDNEY	Fetal Kidney
E087	-	-	4.0E-03	PANCREAS	Pancreatic Islets
E088	1.0E-03	1.0E-03	-	LUNG	Fetal Lung
E093	-	6.0E-03	-	THYMUS	Fetal Thymus
E095	1.0E-03	1.0E-03	-	HEART	Left Ventricle
E097	1.0E-03	1.0E-03	-	OVARY	Ovary
E099	-	1.0E-03	1.0E-03	PLACENTA	Placenta Amnion
E100	1.0E-03	1.0E-03	-	MUSCLE	Psoas Muscle

(Table 15 continues)

E101	-	1.0E-03	1.0E-03	GI_RECTUM	Rectal Mucosa Donor 29
E102	1.0E-03	1.0E-03	-	GI_RECTUM	Rectal Mucosa Donor 31
E103	-	-	1.9E-02	GI_RECTUM	Rectal Smooth Muscle
E104	1.0E-03	1.0E-03	2.2E-02	HEART	Right Atrium
E105	-	2.0E-03	4.7E-02	HEART	Right Ventricle
E106	1.1E-02	1.0E-03	-	GI_COLON	Sigmoid Colon
E109	-	1.0E-03	-	GI_INTESTINE	Small Intestine
E110	-	1.0E-03	-	GI_STOMACH	Stomach Mucosa
E111	-	-	4.9E-02	GI_STOMACH	Stomach Smooth Muscle
E112	-	-	6.0E-03	THYMUS	Thymus
E113	-	-	4.0E-03	SPLEEN	Spleen
E114	-	1.0E-03	2.0E-03	LUNG	A549 EtOH 0.02pct Lung Carcinoma Cell Line
E115	-	-	1.0E-03	BLOOD	Dnd41 TCell Leukemia Cell Line
E116	-	1.1E-02	2.0E-03	BLOOD	GM12878 Lymphoblastoid Cells
E117	-	-	1.0E-03	CERVIX	HeLanaS3 Cervical Carcinoma Cell Line
E118	-	-	2.0E-03	LIVER	HepG2 Hepatocellular Carcinoma Cell Line
E119	-	1.0E-03	9.0E-03	BREAST	HMEC Mammary Epithelial Primary Cells
E120	-	-	1.0E-03	MUSCLE	HSMM Skeletal Muscle Myoblasts Cells
E121	-	1.0E-03	1.0E-03	MUSCLE	HSMM cell derived Skeletal Muscle Myotubes Cells
E122	-	4.0E-02	1.2E-02	VASCULAR	HUVEC Umbilical Vein Endothelial Primary Cells
E123	-	1.0E-03	1.0E-03	BLOOD	K562 Leukemia Cells
E124	1.0E-03	1.0E-03	1.0E-03	BLOOD	MonocytesnaCD14+ RO01746 Primary Cells
E125	1.0E-03	1.0E-03	-	BRAIN	NHnaA Astrocytes Primary Cells
E126	-	1.5E-02	-	SKIN	NHDFnaAd Adult Dermal Fibroblast Primary Cells
E127	-	-	1.0E-03	SKIN	NHEKnaEpidermal Keratinocyte Primary Cells
E128	-	-	1.4E-02	LUNG	NHLF Lung Fibroblast Primary Cells
E129	-	-	3.0E-03	BONE	Osteoblast Primary Cells

3.4. Discussion

A comparison between the genomic distributions of reference RTEs (fixed) and non-reference RTEs (polymorphic) was conducted. Polymorphic RTEs, specifically rare insertions (MAF ≥ 1%), experience selection to a lesser extent than ancient RTE insertions that are now fixed in the human genome. As such, the genomic landscape of non-reference RTEs was expected to be in-between that observed for fixed endogenous elements and *de novo* insertions that potentially reflect integration site preference. To this end, a database of polymorphic non-reference RTEs was curated. The in-house curated database is, to the best of our knowledge, the most comprehensive list of non-reference RTEs in the human genome, encompassing 39,798 RTE insertions identified in over 7,000 individuals. The landscape of L1, Alu, and SVA elements in the human genome was investigated by studying the distribution of RTEs within the following genomic features:

I. Chromosomal distribution; investigated to determine whether the properties of any chromosome allow it to harbour more RTE insertions.

II. Local base composition and functionality; investigated to understand the potential effect of RTE insertions on genome function and stability.

III. Local recombination rate; investigated as there is a strong positive correlation between selection and recombination, and also because of the risk NAHR between homologous RTEs can pose on genome integrity and rearrangement.

IV. Chromatin accessibility; investigated to understand whether RTEs integrate preferentially into regions that allow them to propagate and create new insertions.

Overall, the distribution of non-reference RTEs do display aspects of the interplay between integration preference and differential selection. Reference and non-reference RTEs are characterised by unique genomic distributions, with polymorphic non-reference RTEs displaying a closer genomic landscape to *de novo* insertions than reference RTEs. On average, and in agreement with previous studies, non-reference L1 and Alu elements were similarly distributed in GC-poor, low activity regions, while non-reference SVAs accumulated in GC-rich, high activity regions of the human genome. Below is a discussion of the distribution results for each of the studied RTE elements in each of the studied genomic features.

3.4.1. Chromosomal distribution

The distribution of L1 elements is in line with previous studies. The strong linear correlation between the number of non-reference L1s and chromosome size is consistent with the chromosome-wide distribution of *de novo* L1s (Flasch et al., 2019; Sultana et al., 2019; Chen et al., 2020). The chromosomal distribution analysis shows an over-representation of reference L1 elements on the X-chromosome, which have been associated with its role in propagating the X-inactivation signal (Lyon, 1988; Bailey, et al., 2000). In contrast, non-reference L1s are not significantly enriched on the X-chromosome (Figure 9). The lack of obvious enrichment of non-reference L1s on the X-chromosome is consistent with the reported distribution of *de novo* L1 insertions (Sultana et al., 2019; Chen

et al., 2020), confirming that the accumulation of reference L1s in the X-chromosome results from post-integration selection processes. Non-reference L1s are slightly over-represented in chromosome 4, consistent with a previous study analysing the distribution of 344 non-ref L1s (Boissinot et al., 2004). Boissinot et al. (2004) suggested that the observed bias of L1Hs elements on chromosome 4 is not likely due to preferential integration on this chromosome, but rather a reflection of the amplification of active L1 elements. The chromosomal distribution of *de novo* insertions obtained using engineered L1s supports this suggestion. Two recent studies reported an over-representation of *de novo* L1s in chromosome 1 and chromosome 5 (Sultana et al., 2019 and Flasch et al., 2019, respectively), suggesting that the integration of L1 elements is not perfectly random. The accumulation of non-reference L1s in low GC regions of the genome is consistent with the non-random distribution reported for *de novo* L1 elements. The study by Sultana *et al*. (2019) attributed the biased integration of L1s in GC-poor regions to pre-existing biases in the distribution of the L1 ORF1 target motif. Nevertheless, it was found that new L1s integrate into higher GC regions and broadly target all genomic regions in comparison to endogenous L1s, a difference that was predominantly attributed to evolutionary selection (Flasch et al., 2019; Sultana et al., 2019; Chen et al., 2020).

Sultana *et al.* (2019) reported that *de novo* L1s are not significantly enriched or depleted in gene regions. In contrast, non-reference L1Hs elements, endogenous in HeLa-S3 cells and reference L1s, are significantly depleted in genic regions and enriched in low activity regions. In addition, Chen et al. (2020) reported that the distribution of polymorphic L1s in both gene and active regions is closer to the distribution of reference L1Hs elements than *de novo* L1s. The reported significant depletion of polymorphic L1s in active regions is consistent with the

results of this study, yet their comparison with the control dataset suggests less significant depletion in contrast to the depletion of reference L1s. As such, the results of this study do show that selection pressures acting on polymorphic L1s are in-between those of *de novo* and fixed L1 elements.

Previous studies investigating the genomic distribution of Alu elements have suggested the occurrence of differences between the density of Alus from different evolutionary families across different chromosomes (Grover et al., 2004; Kim et al., 2004). Grover et al. (2004) when examining Alu elements from repeat masker (AluS, AluJ, and Alu Y) reported that, in terms of Alu elements density, chromosome 19 is the densest and chromosome Y is the least dense chromosome. When examining the reference Alu density across chromosomes in this study and others, the same trend is observed, and chromosome 19 appears to hold an increased Alu density in comparison to its size (Figure 10). Reference SVAs are also overrepresented on chromosome 19 (Figure 11). In line with previous studies, the higher density of fixed Alu and SVA elements on chromosome 19 is proportional to the high gene density of this chromosome, as shown by the strong correlation between Alu and SVA density with gene density (Lander et al., 2001; Grover et al., 2004; Wagstaff et al., 2012; Tang et al., 2018; Gianfrancesco et al., 2019).

In contrast to previous studies, non-reference Alu and SVA elements appear to be underrepresented on the X-chromosome (Figures 10 & 11). Wang et al. (2006) and Cotton et al. (2014) both reported a positive correlation between the Alu content of genes on the X-chromosome and their ability to escape the inactivation signal. These observations may explain the under-representation of non-reference Alu and SVA elements on the X-chromosome, given that SVA elements

contain an Alu-like segment in their sequence. Another possibility for the underrepresentation of non-reference Alu and SVA elements on the X-chromosome may be due to sampling bias caused by the heterozygosity of the X-chromosome in males, however, this is unlikely since this phenomenon was not observed for the distribution of non-reference L1 elements.

3.4.2. Local GC content

Alu and SVA elements are both non-autonomous RTEs that use the L1 machinery for transposition, yet SVAs and reference Alu elements (Figures 13 and 14) accumulated in GC-rich regions while L1s (Figure 12) and non-reference Alus accumulated into AT-rich regions. L1 and Alu elements initially integrate into AT-rich regions as the target site of the L1 ORF2 (5'-TTTT/AA-3') are more frequent in these genomic regions (Costantini et al., 2012). In contrast, SVA elements preferentially integrate into GC-rich regions, even though their target site resembles the L1 consensus sequence (Raiz et al., 2012). The reasons behind the preferential integration of SVA elements in GC-rich regions are yet to be understood. Nevertheless, the shift in the GC distribution pattern observed between reference and non-reference RTEs has been associated with natural selection. Alu elements in GC-rich regions were subject to positive selection, due to their role in regulating gene expression and increasing gene stability (Costantini et al., 2012). This explains the shift in GC distribution patterns observed between reference and non-reference Alu elements (Figure 13). Another explanation is that negative selection mostly acts on Alu elements in AT-rich regions, as their removal from these regions is less likely to be harmful to the genome function than the removal of Alu elements in GC-rich/gene-rich regions (Abrusán & Krambeck, 2006). Either explanation would have resulted in shifting

the distribution pattern of Alu elements over time to higher GC regions and would explain the over-representation of reference Alu elements in intron regions (Figures 13 & 15). The removal of deleterious L1 and SVA elements by unequal crossover is thought to contribute to the shifting of the GC distribution pattern of these elements (Wang et al., 2005; Song & Boissinot, 2007). Nevertheless, the preferential integration of SVA elements in GC-rich regions, together with their low density in the human genome (Figure 11.B), gives them a better chance of becoming fixed in GC-rich gene-rich regions. This may explain the over-representation of SVA elements in intron regions, more so than expected by chance (Figure 15, Table 12).

3.4.3. Distribution in genomic regions of functional relevance

The functional impact of RTEs on genome function was examined by analysing their location in relation to known genes and enhancer regions. Non-reference L1 and SVA elements are more frequent in functional regions than their reference counterparts (Figures 15 and 16). The shift in the distribution of L1 and SVA elements in functional genomic regions with time has been attributed to the role of purifying selection in removing deleterious insertions from the genome (Boissinot et al., 2001; Belle et al., 2005; Wang et al., 2005). The over-representation of L1 elements in gene-free regions has been suggested to be due to its retrotransposition mechanism, that may have evolved to target what was called “genomic safe havens” (Levin and Moran, 2014), to minimise the damage of its transposition on the genome function of its host (Cost & Boeke, 1998; Levin & Moran, 2014). A more recent study analysing the target sites of 1,565 L1 insertions that had been experimentally induced suggested that pre-

existing biases in the distribution of the L1 ORF2 motif in the human genome may have impacted the observed distribution of recent L1 insertions (Sultana et al., 2019).

Reference Alu elements are also over-represented in intronic and enhancer regions. The over-representation of reference Alus in intron and enhancer regions suggests their preferential retention in functional regions, possibly due to their positive effect on genome function. This observation is in line with previous studies that suggested the contribution of ancient Alu elements in gene regulation, by providing genes with promoters and enhancers (Cordaux & Batzer, 2009; Su et al., 2014; Trizzino, Kapusta, & Brown, 2018). New Alu insertions do not necessarily offer ‘‘ready-to-use’’ regulatory elements (Warnefors et al., 2010), but their potential ability to gain functional advantages may explain the over-representation of non-reference Alu elements in intron regions compared to the control distribution. Non-reference SVA elements are over-represented in intron and enhancer regions, compared with the control distribution and the distribution of reference SVAs (Figures 15 & 16, Table 12). These results suggest that SVA elements in functional regions are subject to purifying selection due to their negative impact on genome function. The preferential integration of SVA elements in gene regions has been previously reported in the literature (Raiz et al., 2012; Savage et al., 2013; Gianfrancesco et al., 2019).

SVA elements have the potential to influence genome function and regulation through a variety of mechanisms (Kwon et al., 2013; Savage et al., 2013; Bragg et al., 2017; Gianfrancesco et al., 2017; Pontis et al., 2019). The GC-rich sequence of SVAs gives them the potential to create G-quadruplexes (G4) structures (Savage et al., 2013; Bragg et al., 2017). These secondary DNA

structures are strongly associated with their negative impact on genomic and epigenomic stability (Bragg et al., 2017; Spiegel, et al., 2020). Evidence of the ongoing co-evolution of SVA retrotransposons and zinc finger genes that are known to suppress SVA mobilisation, supporting the harmful effect of SVA insertions on genome function and integrity (Jacobs et al., 2014; Gianfrancesco et al., 2019).

3.4.4. Local recombination rate

The distribution of RTEs in regions of different recombination rates may be associated with their local GC content, due to the significant correlation between recombination rate and local base composition. Local GC content is positively correlated with recombination rate, while local AT content correlates negatively with recombination rate (Kong et al., 2002; Mugal et al., 2015). Consistent with previous studies, L1 and SVA elements shift to regions associated with lower recombination rate, while Alu elements shift to regions of higher recombination rate with age (Lander et al., 2001; Medstrand et al., 2002; Wang et al., 2005; Myers et al., 2008). In contrast, Alu elements shift to genomic regions of higher recombination rate with age (Figure 17). This observed shift may be associated with the reduced strength of selection in genomic regions of low recombination rates (Boissinot et al., 2001; Abrusán & Krambeck, 2006; Dolgin & Charlesworth, 2008) and complements the observed shift in the GC distribution pattern of L1 and SVA elements with time to regions of lower GC-content (Figures 12 and 14). In comparison, the observed shift in local recombination rate of Alus from cold to intermediate with age reflects their preferential retention in gene regions, as mentioned above. Interestingly, Abrusán & Krambeck (2006) reported the lack of

difference in the distribution of L1 and Alu elements over time in gene-poor regions on chromosomes 4 and X, suggesting the importance of genes in shaping the genomic distribution of RTEs in the genome. In addition, a recent review article has reported the growing number of studies supporting the co-evolution between RTEs integration and local recombination rate, implicating the importance of transposons in the evolution of the genome structure and recombination rates (Kent et al., 2017).

3.4.5. Local chromatin accessibility

The distribution of RTEs within active (euchromatin) and repressed (heterochromatin) chromatin states were investigated. Reference SVAs are the highest of the RTEs in heterochromatin regions, followed by non-reference L1s. In contrast, non-reference SVAs are the most predominant RTEs in euchromatin regions followed by reference Alus and non-reference L1s (Figure 18). The differential distribution of reference and non-reference SVAs in genomic regions of different chromatin accessibility, suggests the negative effect of SVA transposition on the fitness of its host. These observations are consistent with the continuous co-evolution of zinc finger genes with SVA elements to suppress SVA activity. The high fraction of reference Alus in euchromatin domains is consistent with its accumulation in intron regions. Non-reference L1 elements are more frequent in both euchromatin and heterochromatin domains compared to the reference L1s, however, they are more frequent in heterochromatin domains. These results are inconsistent with that of a recent study reporting that new L1 insertions are not affected by chromatin states (Sultana et al., 2019). The discrepancy between both studies may be due to this study using an average

percentage of L1s from different cell types that belong to the same group of tissue, while Sultana et al. (2019) investigated the insertion of *de novo* L1 insertions in HeLa S3 cells. Sultana et al. (2019) have also suggested that L1 integration is influenced by replication timing of the target region. Chromatin states of different cell types may potentially have a different effect in shaping the integration pattern of L1 insertions.

3.4.6. Study overview and concluding remarks

This chapter demonstrates the differential genomic distribution of reference RTEs that are fixed in the human genome, and non-reference RTEs that are polymorphic in the population. The current study builds on previous reports, supporting the role of selection in shaping the genomic distribution of endogenous elements, while revealing aspects of the interplay between integration preference and selection forces in shaping the distribution of polymorphic RTEs. Nevertheless, the genomic distribution of non-reference elements suggests that RTE activity has the potential to cause detrimental effects on genome function and increase susceptibility to multifactorial disorders, made evident by their ready occurrence in regions of functional relevance. In theory, RTE activity has the potential to affect all human diseases. Indeed, RTE activity has already been associated with several complex disorders such as cancer (Burns, 2017), ALS (Savage et al., 2019), schizophrenia and Alzheimer's disease (Guffanti et al., 2014; Terry et al., 2020). We hypothesise that SVA transposons in particular pose the largest detrimental effect on the human genome, based on the increased accumulation of non-reference SVAs in active regions in comparison to non-

reference L1 and Alu elements. This finding is significant since it demonstrates the importance of co-analysing the distribution of all active RTE elements simultaneously, in order to achieve a comprehensive understanding about the effect of all RTE activity on genome function as a whole.

Although this study analysed a larger sample size than any previous study, potentially analysing more of the most recent insertions of low and rare allelic frequencies, the heterogeneity of allele frequency may have masked some variations in the genomic distribution of RTE elements. Future in-depth analysis of the genomic distribution of RTEs separated into groups of similar allele frequencies is required to gain a deeper understanding regarding the interplay between selection and integration preference of the currently amplifying RTE families. In addition, a recent analysis of *de novo* SVA insertions in culture, similar to the recent analyses of *de novo* engineered L1s (Flasch et al., 2019; Sultana et al., 2019; Chen et al., 2020), is required to gain a better understanding about the integration preference of SVA elements in gene-rich regions.

4. RTEs as potential variants of disease

4.1. Introduction

Structural variants (SVs), defined as sequence variations (> 50bp) between individual genomes (Sudmant et al., 2015), are an important class of genomic variations that account for most base pair (bp) differences between the genomes of individuals within a population. SVs include, amongst other types, deletions, duplications, copy-number variations (CNVs), and insertions (Feuk et al., 2006; Stankiewicz and Lupski, 2010; Weischenfeldt et al., 2013; Escaramís et al., 2015). These types of genomic variations are known to influence gene expression and pathological traits in humans (Feuk et al., 2006; Weischenfeldt et al., 2013; Chiang et al., 2017). As such, SVs have been proposed as a potential source of genomic variants accountable for part of the missing heritability problem of complex traits.

4.1.1. The missing heritability of complex traits:

Complex traits are traits determined by multiple genetic and environmental factors. Such traits are more common than monogenic traits (i.e., traits influenced by a single allele/gene) and show a continuous range of phenotypic characteristics, influenced by the interplay between the genetic and environmental factors (Rowe and Tenesa, 2012; Barton et al., 2017). Examples of complex traits include natural hair colour, height, body mass index (BMI), and many diseases, including Autism spectrum disorders (ASDs), Parkinson's disease, and Alzheimer's disease. Genome-wide association studies (GWAS)

aim to identify associations between genetic variants, typically single-nucleotide polymorphisms (SNPs), and the observed phenotypic variation. However, for many complex traits, the cumulative effect size of all genome-wide significant SNPs ($P < 5x10^{-8}$) can only explain a small portion of the trait heritability as inferred from family-based studies, thus raising the missing heritability issue (Maher, 2008; Manolio et al., 2009; Rowe and Tenesa, 2012).

Two distinct models have been proposed for explaining the missing heritability of complex traits. The first model suggests that complex traits are controlled by many common variants, each contributing an infinitely small additive effect on the observed phenotype (Gibson, 2012; Hu et al., 2012; Weiner et al., 2017). This model proposes that variants with an effect size too small to be significant are usually missed from association studies, thus, heritability is not so much missing but rather hidden beneath the stringent significant level required by GWAS. Recent studies have indeed demonstrated that the majority of the missing heritability for several complex traits is recoverable from many variants of small effect sizes (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Wood et al., 2014; Abraham et al., 2016). The second model suggests that the missing heritability is largely caused by a smaller number of large-effect rare variants that are neither well tagged by local SNPs nor covered by commercial SNP arrays that are typically used in GWAS (Dickson et al., 2010; Gibson, 2012; Yang et al., 2015). It has been suggested that both models are likely to contribute by different degrees to the heritability of complex traits and disorders (Gibson, 2012; Agarwala et al., 2013). Recent studies have confirmed that integrating both common and rare variants in the heritability analysis does recover a larger proportion of the expected heritability for several traits (Weiner et al., 2017; Wainschtein et al., 2019). However, heritability estimates derived

from GWAS can be further enhanced by incorporating other types of genomic variations that continue to segregate in humans.

4.1.2. Structural variants and complex traits:

SVs are likely to have a greater impact on genome function than SNPs because of their larger genomic size, spanning hundreds and thousands of base pairs (bp). In addition to their large genomic size, SVs account for the majority of bp genetic variations in humans, therefore, integrating these variants in association studies could improve estimating the genetic variance of complex traits (Manolio et al., 2009). Indeed, numerous studies have implicated SV with complex human traits and diseases including obesity, cancer, cognitive ability, and psychiatric disorders plus many others (Stankiewicz and Lupski, 2010; Kawamura et al., 2011; Malhotra and Sebat, 2012; Lacaria et al., 2013; Weischenfeldt et al., 2013; Waddell et al., 2015; Carvalho et al., 2016; Cuccaro et al., 2016; Smoller, 2016; Shorter, 2017; Weiner et al., 2017). However, association studies of this sort have mainly focused on few SV types, such as copy number variants (CNVs) and megabase-scale deletions and duplications. As such, associations of other SV types, such as polymorphic RTE insertions, with complex human traits and diseases are not well described.

4.1.3. RTE insertions as SV:

Active RTE subfamilies provide an ongoing source of SVs in the human genome. Germline RTE insertions are capable of creating inter-individual insertional polymorphisms, defined as insertions showing variations in their presence or absence state at specific genomic loci across the genome of individuals within a

population (Wang et al., 2005; Mills et al., 2007; Huang et al., 2010; Rishishwar et al., 2015). Active RTEs are also capable of mediating other types of SV, including deletions and duplications. (Xing et al., 2009; Startek et al., 2015; Bourque et al., 2018). The repetitive nature of RTE elements in the human genome plus their high sequence homology with ancient RTE subfamilies that are no longer active makes it difficult to detect polymorphic insertions that contribute to human genetic diversity (Ewing, 2015; Rishishwar et al., 2017; Bourque et al., 2018). RTE detection studies post the development of computational detection tools enabled the genome-wide discovery of polymorphic RTEs. Such studies were able to show that RTE variants are a natural component of the human genome and can be present at both common (minor allele frequency (MAF) ≥ 0.01) and rare (MAF < 0.01) allele frequencies within a population (Rishishwar et al., 2017). A recent analysis of human genetic variations from phase 3 of the 1,000 genome project (1kGP) proposed that RTE insertions account for a large proportion of inter-individual genetic diversity, estimating that about 691 kilobases per individual genome are composed of polymorphic RTEs (Sudmant et al., 2015). Still, SVs mediated by RTE insertions have been frequently overlooked by previous association studies possible due to the complexity of their discovery within the human genome (Ewing, 2015; Goerner-potvin and Bourque, 2018). As such, the clinical significance of many RTE variants remains unknown.

4.1.4. RTE-mediated SVs and Complex traits:

Trait-associated SNPs (TAS) rarely have a direct causal effect on trait phenotype. Instead, TAS act as genomic markers that are co-inherited with the causing variant in the same haplotype (Frayling, 2014). This provides the opportunity for associating polymorphic RTE insertions with complex traits and diseases. Recent studies have shown that RTE insertions can be found in genomic regions that have been associated with complex traits. Sudmant et al. (2015) reported that GWAS loci are enriched for common SVs by up to threefold, however, the enrichment analysis combined many SV types including RTE-mediated SVs, and the enrichment was most pronounced for large deletions (>20 Kb). Alu elements are the most active and abundant class of RTEs by copy number in the human genome. A recent study investigating the association of these elements with complex traits reported a significant enrichment of Alus in GWAS risk loci (P-value = 0.013) (Payer et al., 2017). The previously mentioned studies plus others found that some polymorphic RTE insertions in GWAS loci are in linkage disequilibrium (LD) with the TASs (Sudmant et al., 2015; Hehir-Kwa et al., 2016; Payer et al., 2016; Wang et al., 2017). In addition, Wang *et al*. (2017) and Spirito et al., (2019) have shown that some RTEs in LD with TAS are associated with altered gene expression in a tissue-specific manner. These results suggest the RTE variants can potentially be the causative variant within some GWAS loci. Characterizing the association between RTE variants with complex human traits and diseases will improve as additional GWAS and RTE detection studies identify more TAS and RTE variants, respectively.

4.1.5. Study overview:

This study conducts an enrichment analysis of L1s, Alus, and SVAs in GWAS risk loci using an updated list of TAS from the NHGRI-EBI Catalog of published GWAS (www.ebi.ac.uk/gwas; Buniello et al., 2019), and a comprehensive in-house curated database of polymorphic RTE insertions reported in the literature up to April 2019. The total number of RTEs in the in-house curated database is 2-fold higher than the number reported in phase 3 of the 1kGP (Sudmant et al., 2015) and includes an additional 750, 718, and 35 common L1s, Alus, and SVAs, respectively. Although L1 and SVA elements are not as active or abundant as Alu elements, their larger size and genomic distribution are suggestive of their potentially harmful effect on genome function. The genomic distribution analysis of this study identified several L1 and SVA variants within gene regions, including genes critical for cellular maintenance, proper development, and neurophysiological processes. This study will also conduct a genome-wide screen for additional RTE polymorphisms in LD with TAS.

4.2. Methods

4.2.1. Overview of methods:

The in-house curated database of polymorphic RTE elements is used to identify RTE elements in linkage disequilibrium (LD) with SNPs that have been associated with the risk of various complex disorders through GWAS. As LD is population specific, this study focuses on variants identified in individuals of European descent, to maximize the use of GWAS data since the majority of the association studies have been conducted on cohorts of European descent (Evans and Cardon, 2005; Medina-Gomez et al., 2015). Note that European descent is defined here as Caucasian individuals of European ancestry. Polymorphic RTEs that overlap with the LD-blocks of genome-wide significant (GWS) ($P \leq 5x10^{-8}$) trait-associated SNPs (TAS) were identified. The enrichment of RTE in the LD-blocks of TAS was calculated by comparing the fraction of RTEs contained within a TAS LD-block with the fraction expected by chance using 1,000 sets of random LD-blocks that match the genomic features of the TAS. The non-random association between the occurrence of the RTE and the TAS, i.e., the likelihood of the two alleles being co-inherited in the same individual, was calculated using genotype data from the European population samples in phase 3 of the 1,000 genome project (1kGP) (Sudmant et al., 2015). The distribution of RTEs that are in LD with TAS within functional genomic regions was compared with the distribution of all polymorphic RTEs of the in-house curated database.

Note that RTE elements and LD-blocks located on the sex chromosomes or within the human leukocyte antigen (HLA) region were excluded from the analysis. RTEs and LD-blocks in these genomic regions were eliminated to reduce biases in the results that may arise due to the unequal effect of

evolutionary processes on the sex chromosomes (Johnson and Lachance, 2012) and the haplotype diversity of the HLA loci (Shiina et al., 2009).

4.2.2. Datasets:

4.2.2.1. Polymorphic retrotransposable element (RTE) insertions:

A comprehensive database of non-reference L1Hs, AluY, and SVA_E/F was curated in-house (Chapter 2: Database Curation). The database by individual contains the insertional profile of 2,987 nonrelated individuals and consists of 6,377 L1Hs, 18,698 AluYs, and 1,085 SVAs from the E and F human-specific subfamilies. RTEs overlapping with HLA regions were removed using the subtract tool of BEDtools version 2.25.0 (Quinlan, 2014). The HLA region coordinates were obtained from the genome reference consortium website (GRC; https://www.ncbi.nlm.nih.gov/grc/human/regions/MHC?asm=GRCh37).
Insertions on the sex chromosomes were removed using a conditional awk statement in UNIX based on the chromosome name column.

Low frequency and common RTEs (MAF ≥ 0.01) identified in samples of European descents were extracted using a conditional awk statement in UNIX based on the allele frequency column. In the case of missing allele frequency information, insertions identified in two or more unrelated individuals by more than one study were retained. This is because the chance of identifying a rare allele in the same location by more than one RTE detection method in two or more unrelated individuals is very slim considering that different methods produce different results as discussed in the database curation chapter. Alternatively, the AF for insertions identified in two or more unrelated individuals by the same study was inferred assuming heterozygosity and the total number of samples analysed

by the study. For example, if the study sample size was 10 and the insertion was identified in 2 individuals, the AF was inferred to be 0.1 and the insertion was retained.

4.2.2.2. Trait associated SNPs:

The trait associated SNPs (TAS) of the NHGRI-EBI Catalog of published genome-wide association studies (www.ebi.ac.uk/gwas) (Buniello et al., 2019) was downloaded from the UCSC genome browser tables (last updated: 2019-08-07) (https://genome.ucsc.edu/cgi-bin/hgTables) (Karolchik, 2004) using the hg19 genome assembly coordinates. Genome-wide significant (GWS) SNPs (P-value $\leq 5x10^{-8}$) identified in European cohorts were extracted using a conditional awk statement in UNIX.

4.2.2.3. SNP and RTE genotype files from the 1000 genome project:

Genotypes of SNPs from phase 3 (version 5a) release of the 1kGP (Sudmant et al., 2015) were obtained from the 1kGP ftb website (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). Two files in .vcf and .tbi format were obtained for each autosome (chr1-chr22). Sample names from the European super population were extracted into a text file from the 1kGP sample panel file (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/integrated_call_samples_v3.20130502.ALL.panel) using a conditional awk statement in UNIX. Vcftools version 0.1.11 ((C) Adam Auton 2009) was then used to extract SNP genotypes identified in the 503 European samples and convert the .vcf files to transposed ped (tped) and fam (tfam) files. The transposed files were converted

into plink PED/MAP files using PLINK version 1.07 (Purcell et al., 2007). A new dataset was created using the --maf plink filter which extracted all rare variants with a MAF below 0.01 (MAF < 0.01). The new dataset was then converted into plink binary format (.bed, .bim, .fam) using the --make-bed option. Duplicate markers were identified using the awk command and .bed file. Unique SNPs were saved into a text file and were filtered out using the --extract option. The process of downloading and formatting the vcf file was then repeated for each chromosome. The individual binary files for each autosome were then merged into one binary file using --merge-list option in PLINK version 1.07 (Purcell et al., 2007). All the commands used in the steps above from downloading the vcf file to creating the binary PLINK files for all autosomes were uploaded to github (https://github.com/RandaAli1/MyPhDproject/blob/master/RTE_enrichment_in_GWAS_loci/1kGPEuropeanGenotypes.txt). The merged binary files for all autosomes were then used to identify tagging SNPs for creating the LD-blocks.

Genotypes of genomic structural variations from phase 3 (volume 1) release of the 1kGP (Sudmant et al., 2015) were obtained from the 1kGP ftb website (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map) as .vcf and .tbi files. The vcf file contained 12 different structural variation types including 3,048 L1s, 12,748 Alus, and 835 SVAs (Sudmant et al., 2015). L1, Alu and SVA elements were extracted from the structural variations .vcf file using the grep command in UNIX. The new vcf file containing genotypes of the RTE elements was converted into a binary PLINK file format using the same steps mentioned above up to the removal of duplicate variants. The individual binary files for each autosome plus the binary file of the RTE variants were merged into one binary file using the --merge-list option in PLINK version 1.07 (Purcell et al., 2007). The new merged file was used for calculating LD between RTEs and TAS.

4.2.2.4. Functional regions file:

The Reference Sequence (RefSeq) genes (O'Leary et al., 2016) and GeneHancer (Fishilevich et al., 2017) enhancer files described in the genomic distribution chapter were used for the analysis of this chapter as well.

4.2.3. Method of data analyses

4.2.3.1. Overlapping RTEs with TAS LD-blocks

Linkage disequilibrium (LD) blocks for each TAS were defined using the left and right-most (5' and 3') tagging SNPs ($r^2 \geq 0.8$) (Figure 19). The tagging SNPs were identified using the --show-tags option and --tag-r2 filters of PLINK version 1.90b6.21 (Purcell et al., 2007). LD-blocks of TAS that lacked a 5' or a 3' tagging SNP were arbitrarily extended by half the median of the LD-blocks defined by 5' and 3' tagging SNPs using the slope tool and -l and -r options of BEDtools version 2.25.0 (Quinlan, 2014). Arbitrary LD blocks for untagged TAS were created using the slope tool and -d option of BEDtools version 2.25.0 (Quinlan, 2014). The list of LD-blocks were reduced using the merge tool of BEDtools version 2.25.0 (Quinlan, 2014) to create non-overlapping LD-blocks. LD-blocks overlapping the HLA region (GRC; https://www.ncbi.nlm.nih.gov/grc/human/regions/MHC?asm=GRCh37) were removed using the subtract tool of BEDtools version 2.25.0 (Quinlan, 2014). LD-blocks on the sex chromosomes were removed using a conditional awk statement in UNIX. The final list of LD-blocks were overlapped with the list of polymorphic RTEs using the intersect tool of BEDtools version 2.25.0 (Quinlan, 2014).

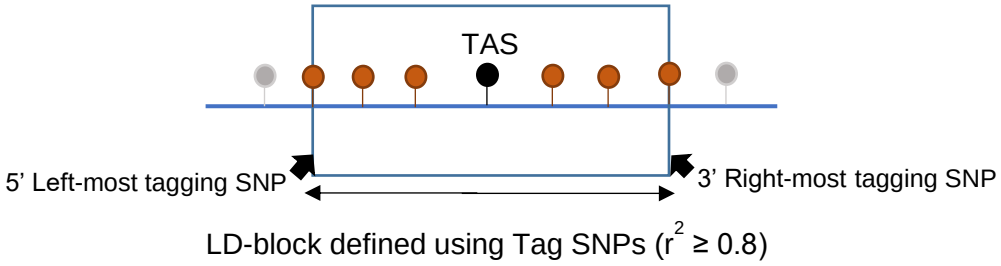


Figure 19: Creating LD-blocks around each trait associated SNP (TAS) using tagging SNPs with an r2 ≥ 0.8

4.2.3.2. Enrichment of RTE variants in GWAS risk loci

The enrichment of overlap between RTEs and TAS LD-blocks was investigated by creating a thousand set of random SNPs that match the genomic features of the TAS such as allele frequency, number of SNPs in LD, distance to nearest gene and gene density, were generated using SNPsnap default settings (https://data.broadinstitute.org/mpg/snpsnap/) (Pers et al., 2015). The TAS were matched in this way to control for the biased distribution of TAS in the genome, which are reportedly enriched in gene regions (Hindorff et al., 2009).

Non-overlapping LD-blocks for each matching SNPs set were generated as previously described using an integrated R script (https://github.com/RandaAli1/MyPhDproject/tree/master/RTE_enrichment_in_GWAS_loci). The script loops through each set of the 1,000 random SNPs sets and create a merged list of LD-blocks for it, which it then exports into a file using the write.table() function in R version 3.4.0 (R Core Team, 2012). The overlap between each of the 1,000 sets of random LD-blocks and polymorphic RTEs was identified using the intersect tool of BEDtools version 2.25.0 (Quinlan, 2014) via a loop function in R version 3.4.0 (R Core Team, 2012). The loop function

intersects each set of random LD blocks with the list of polymorphic RTE elements (L1, Alu, or SVA) and calculates the number of LD blocks containing a polymorphic RTE (https://github.com/RandaAli1/MyPhDproject/tree/master/RTE_enrichment_in_GWAS_loci). A density plot representing the percentages of LD blocks overlapped by a polymorphic RTE were created in R version 3.4.0 (R Core Team, 2012) for each RTE type. The empirical P-value was then calculated to compare the percentage of TAS LD-blocks overlapping with each RTE type against the percentage distribution of random LD-blocks (North et al., 2002).

4.2.3.3. LD between RTEs and TAS:

An RTE in the genomic region of a TAS that is potentially the causative variant within the TAS haplotype is expected to be in LD with the TAS. LD analysis was only possible for RTEs identified by the 1kGP (Sudmant et al., 2015) as it is the only study within the in-house curated database that provides an accessible source of SNP and RTE genotypes for each sample. LD analysis between RTEs in risk regions and the corresponding TAS was conducted using genotype data of European samples (n=503) from phase 3 of the 1kGP (Sudmant et al., 2015) using PLINK version 1.90b6.21 (Purcell et al., 2007). RTE elements from the 1kGP that overlapped with a TAS LD-block were extracted using a conditional awk statement. LD between the set of 1kGP RTEs and TASs were calculated using the --ld-snp-list command and the --ld-window-r2 filter using PLINK version 1.90b6.21 (Purcell et al., 2007). A moderate r^2 threshold of 0.6 (Tian et al., 2017) was set in order to capture as many associations between RTEs and TAS as possible. RTEs in LD with TAS were matched with the trait information using the merge by SNP function in R version 3.4.0 (R Core Team, 2012). RTEs in LD with

a TAS (r^2 > 0.6) were overlapped with gene and enhancer regions using the intersect tool of BEDtools version 2.25.0 (Quinlan, 2014). The difference between the distributions of RTEs in LD with TAS in gene regions compared to all non-reference RTEs was compared using the chi-squared goodness of fit test.

4.3. Results

4.3.1. TAS Linkage disequilibrium blocks (LD-blocks):

A total of 158,654 trait associated SNPs (TASs) were downloaded from the NHGRI-EBI Catalog of published genome-wide association studies (www.ebi.ac.uk/gwas) (August 2019 & earlier). From this set, 50,171 SNPs were extracted that met the following criteria:

1. Significantly association with a trait at a P-value level ≤ $5x10^{-8}$ and,
2. Identified in a European population.

Linkage disequilibrium blocks were identified for 64.24% of the 50,171 TASs (n=32,229). The generated LD blocks ranged from 3 bp to 499.964 kilobases (Kb) in size, with a median of 54.096 Kb. The remaining 17,942 SNPs lacked a 5' or 3' tagging SNP (r^2≥0.8). As such, LD blocks were created for these untagged TASs by extending the region either side of the TAS by half of the median (n=27.048) of the generated LD blocks.

A total of 48,911 LD-blocks located on autosomes and not interrupting the HLA region were extracted. The 48,911 LD blocks were reduced to 10,905 non-overlapping LD-blocks using BEDtools intersect. The 48,911 non-reduced LD blocks generated are available in github (https://github.com/RandaAli1/MyPhDproject/blob/master/RTE_enrichment_in_GWAS_loci/allEurGWS_TAS_LDblocks_nosexchrNoHLA_22082019.bed).

4.3.2. Overlapping RTEs with TAS LD-blocks:

A total of 5,240 RTEs (1,160 L1s, 3,825 Alus, and 255 SVAs) were extracted from the in-house curated database by individual that met the following criteria:

1. Identified in samples of European descents.
2. Low frequency and common RTEs identified with a MAF ≥ 0.01.
3. Are autosomal RTE insertions.
4. Do not overlap with the HLA region.

The extracted RTEs were overlapped with the 10,905 TAS LD-blocks using BEDtools intersect. Of the 5,240 RTEs, a total of 2,063 (425 L1s, 1,523 Alus, and 115 SVAs) overlapped the genomic regions of 19,296 TAS contained within the reduced TAS LD-blocks, of which 16,212 were unique TAS (Table 16). More than one-third of polymorphic L1s and Alus and half of the polymorphic SVAs were found within a TAS LD-block. The proportion of class-specific RTEs in a TAS LD-block is significantly higher for SVA elements compared to both L1 and Alu variants (P-value = 0.024, Fisher exact test).

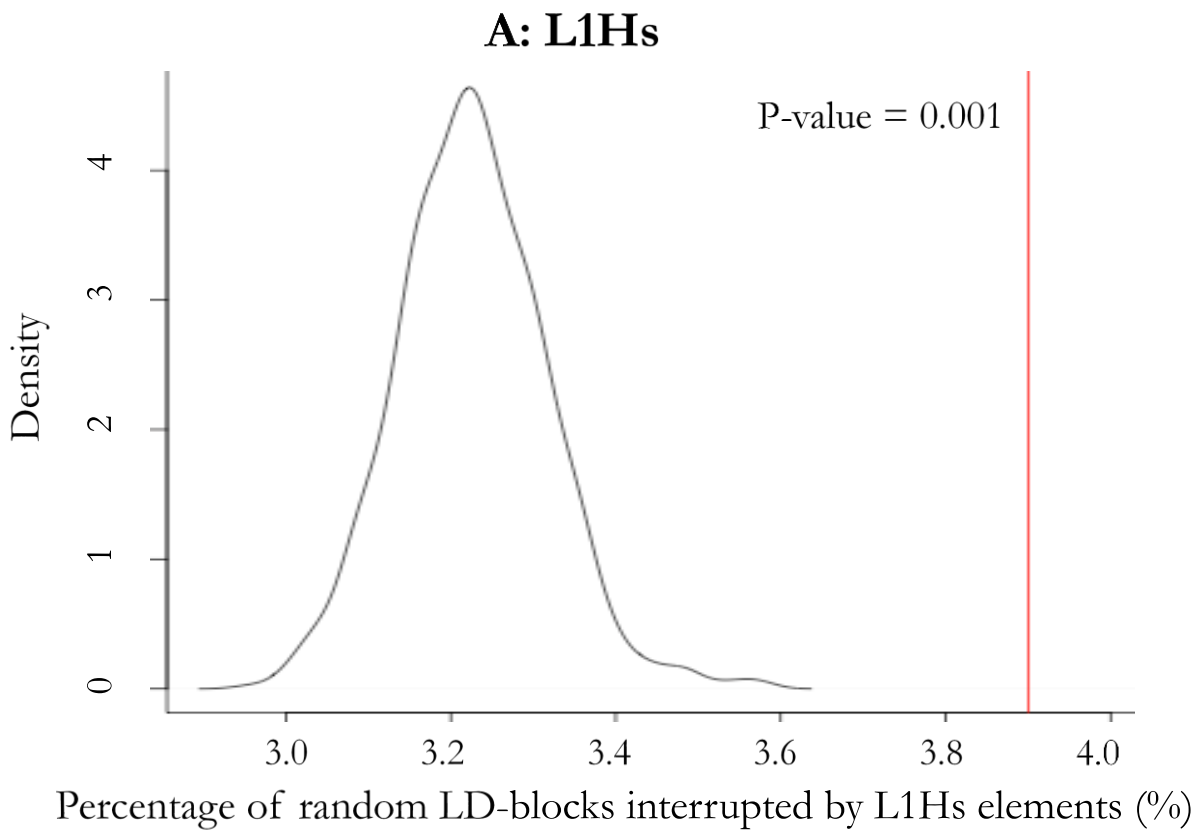
Table 16: The overlap between polymorphic RTEs and the LD-blocks of genome-wide significance ($P \leq 5x10^{-8}$) TASs. LD blocks were generated using tagging SNPs ($r^2 \geq 0.8$) in PLINK. Polymorphic RTEs were curated in-house. Abbreviations: RTE: Retrotransposable elements; TAS: Trait associated SNP; LD: Linkage disequilibrium.

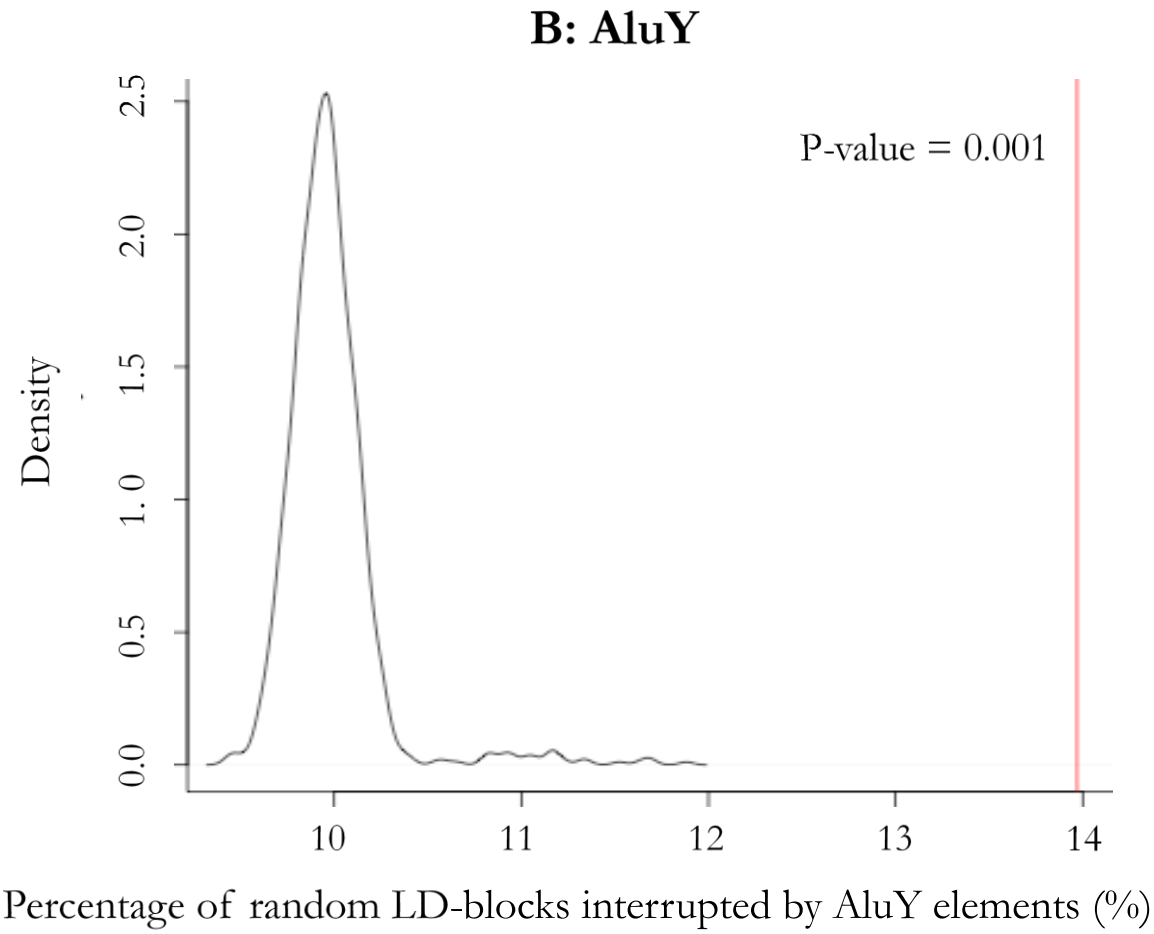
	L1Hs	**AluY**	**SVA_E/F**
Total RTEs in in-house curated database by individual	6,377	18,698	1,085
Common autosomal RTEs identified in European samples, excluding RTEs in the HLA region	1160	3825	255
Common RTEs intersecting a TAS LD-block	425 (36.64%)	1523 (39.82%)	115 (45.1%)
Number of TAS with an RTE within its LD-block (TAS n= 48,911)	4,189	13,407	1,700

4.3.3. Enrichment of RTEs in TAS LD-blocks:

The enrichment of RTEs in TAS LD-blocks was investigated by comparing the observed number of overlaps with the values expected by chance using 1,000 sets of random LD-blocks using SNPs that mirror the genomic properties of the TAS. A total of 42,766 were matched by SNPsnap (Pers et al., 2015) from the 48,911 autosomal TAS list. Similar to the TAS LD-blocks, the random LD-blocks were also reduced to non-overlapping LD-blocks using BEDtools intersect. The generated LD blocks for the 1,000 random sets ranged from 1536 to 19,746 non-overlapping LD-blocks, with an average of 17,388 LD-blocks. As such, the percentage of overlap was calculated to standardize the varying numbers of LD-blocks from each of the random sets.

On average, 3% of the random LD-blocks were interrupted by L1s, 10% were interrupted by Alus, and 0.7% were interrupted by SVAs. In comparison, 4% of the 10,905 TAS LD-blocks were interrupted by L1s, 14% were interrupted by Alus, and 1% were interrupted by SVA elements. The percentage of TAS LD-blocks interrupted by RTE elements is significantly more than that observed for random LD-blocks (Empirical P-values = $1x10^{-3}$ for L1s and Alus, and $3x10^{-3}$ for SVAs) (Figure 20).


A: L1Hs
P-value = 0.001
Density
0
1
2
3
4
3.0
3.2
3.4
3.6
3.8
4.0
Percentage of random LD-blocks interrupted by L1Hs elements (%)


B: AluY
P-value = 0.001
Density
0.0
0.5
1. 0
1.5
2.0
2.5
10
11
12
13
14
Percentage of random LD-blocks interrupted by AluY elements (%)

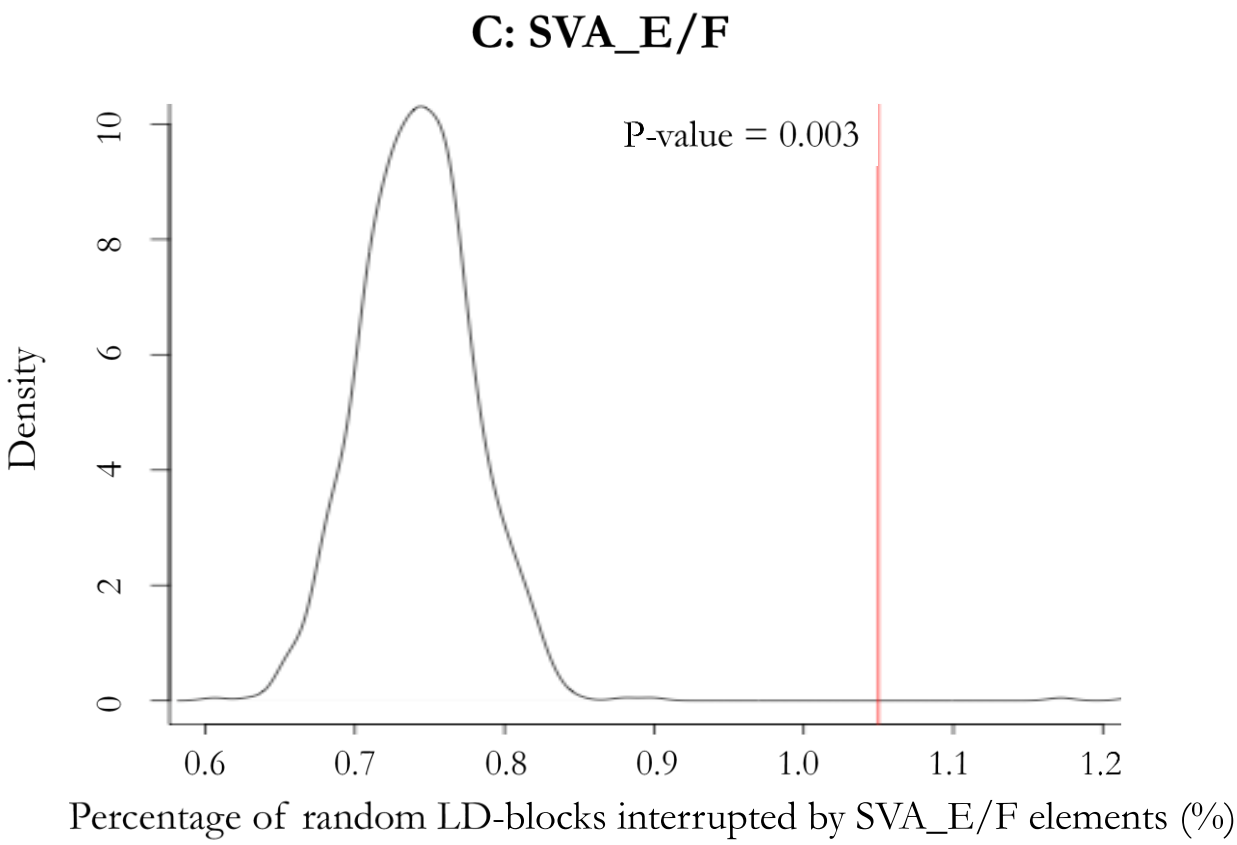


Figure 20: Density plot for the enrichment of polymorphic RTEs at GWAS risk loci. The frequency distribution of random LD-blocks intersected by polymorphic RTEs (black curve) was compared with the frequency observed for TAS LD-blocks intersected by L1Hs (3.90%; A), AluY (13.97%; B), and SVA_E/F (1.05%; C) (red vertical line). The 1,000 sets of random LD-blocks matched the genomic properties of risk SNPs LD-blocks. The empirical P-value was used to calculate the statistical significance of polymorphic RTE enrichment at GWAS risk loci.

4.3.4. LD between RTEs and TAS:

LD analysis between genomic variants is calculated using genotype information of the variants within a specified population. The 1kGP study is the only study within the in-house curated database with accessible genotype information of SNPs and RTEs within the same sample. In addition, about 80% of the Alus and SVAs and a third of the L1s interrupting a TAS LD-blocks (Table 16) were also identified by the 1kGP (Sudmant et al., 2015). As such, the 1kGP genotype data was utilised for calculating LD between RTEs and TASs using a moderate r^2 threshold of 0.6 (Tian et al., 2017). A total of 137 RTEs (16 L1, 113 Alus, and 7 SVAs) are in LD ($r^2 > 0.6$) with 429 TAS, including redundant TAS associated with the same phenotype or trait (Appendix 2). Forty-nine RTEs (8 L1s, 38 Alus, and 3 SVAs) of the 136 total are in strong LD ($r^2 \geq 0.8$) with 157 TAS that have been associated with a variety of traits, including brain-related disorders such as Parkinson's disease, Progressive supranuclear palsy, Schizophrenia, Depression, and anxiety (Table 17). In cases of redundant TAS, the TAS with the highest r^2 of association was included in the table. Where both TAS had the same r^2 value, the TAS reported with the most significant P-value of association with the trait is retained in table 17.

Table 17: Polymorphic RTEs in LD ($r^2 \geq 0.8$) with GWAS TAS. The r^2 values are to the nearest decimal place. Abbreviations: RTE: Retrotransposable element; TAS: Trait associated SNP; LD: Linkage disequilibrium; GWAS: Genome-wide association studies. Note: Table includes strongest LD/GWAS signal when there are multiple RTE-TAS associations of the same phenotype.

Chr	RTE Start	RTE Name	TAS (rs#)	Trait	P-value of TAS	r^2
6	141414904	ALU_5647	rs113803678	Body mass index	4.00E-08	0.9
12	24868717	LINE1_2249	rs61914312	Hair color	3.00E-10	0.9
9	11329329	ALU_7283	rs2152261	Menarche (age at onset)	2.00E-13	0.9
12	56753252	ALU_9228	rs2066819	Inflammatory skin disease	5.00E-17	0.9
1	174484646	ALU_604	rs140581634	Feeling miserable	2.00E-08	0.9
12	120130849	ALU_9540	rs17442937	Red cell distribution width	4.00E-08	0.9
17	44153977	SVA_706	rs62055546	Alcohol consumption (drinks per week)	8.00E-25	0.8
17	44153977	SVA_706	rs17563986	Cognitive ability	5.00E-12	0.8
17	44153977	SVA_706	rs62057061	Depressed affect	2.00E-22	0.8
17	44153977	SVA_706	rs12150229	Ease of getting up in the morning	4.00E-09	0.8
17	44153977	SVA_706	rs62057107	Educational attainment	5.00E-38	0.8
17	44153977	SVA_706	rs55657917	Experiencing mood swings	3.00E-20	0.8
17	44153977	SVA_706	rs56303031	Heel bone mineral density	6.00E-24	0.8
17	44153977	SVA_706	rs17563683	Hemoglobin concentration	2.00E-28	0.8
17	44153977	SVA_706	rs62055701	Irritable mood	6.00E-13	0.8
17	44153977	SVA_706	rs1864325	Lumbar spine bone mineral density	5.00E-11	0.8
17	44153977	SVA_706	rs79412431	Lung function	3.00E-49	0.8
17	44153977	SVA_706	rs12373124	Male-pattern baldness	5.00E-10	0.8
17	44153977	SVA_706	rs241036	Menarche (age at onset)	7.00E-13	0.8
17	44153977	SVA_706	rs55657917	Negative Feelings	7.00E-29	0.8
17	44153977	SVA_706	rs76761706	Neuroticism	7.00E-32	0.8
17	44153977	SVA_706	rs2942168	Parkinson's disease	1.00E-28	0.8
17	44153977	SVA_706	rs55657917	Physical activity measurement	5.00E-12	0.8
17	44153977	SVA_706	rs4606752	Reticulocyte count	1.00E-17	0.8
17	44153977	SVA_706	rs8072451	Subcortical brain region volumes	1.00E-08	0.8
17	44153977	SVA_706	rs75022332	Worry	2.00E-08	0.8
9	16682313	ALU_7311	rs12335424	Height	2.00E-21	0.8
15	76826019	ALU_10819	rs506000	Estimated glomerular filtration rate	2.00E-15	0.8
17	44153977	SVA_706	rs117124984	Daytime nap	3.00E-13	0.8
17	44153977	SVA_706	rs1981997	Interstitial lung disease	9.00E-14	0.8
17	44153977	SVA_706	rs8070723	Progressive supranuclear palsy	2.00E-118	0.8
17	44153977	SVA_706	rs2106786	Red blood cell count	3.00E-36	0.8
17	44153977	SVA_706	rs62061733	Eosinophil counts	3.00E-29	0.8
12	56753252	ALU_9228	rs59917308	Height	3.00E-32	0.8
17	44153977	SVA_706	rs112010353	Self-reported math ability	2.00E-08	0.8
2	210260754	ALU_1947	rs1080278	Lung function	1.00E-19	0.8
17	44153977	SVA_706	rs1991556	Sleep duration	3.00E-09	0.8
16	75655176	ALU_11116	rs61537885	Smoking Status	8.00E-09	0.8
17	44153977	SVA_706	rs80103986	Hand grip strength	1.00E-09	0.8
4	134596423	LINE1_967	rs12507927	Educational attainment	3.00E-11	0.8
17	44153977	SVA_706	rs17652520	Medication use (anilides)	8.00E-13	0.8
1	169524859	LINE1_164	rs6128	Blood protein levels	2.00E-26	0.8
17	44153977	SVA_706	rs62063281	Number of sexual partners	4.00E-15	0.8
17	44153977	SVA_706	rs62063281	Osteoarthritis (hip)	5.00E-12	0.8
1	163639693	ALU_559	rs12564153	Lung function	1.00E-09	0.8
5	109051004	ALU_4562	rs4388249	Schizophrenia	2.00E-08	0.8
17	44153977	SVA_706	rs62064364	Macular thickness	4.00E-35	0.8
10	106566893	ALU_8208	rs61867293	Depression	7.00E-10	0.8

(Table 17 continues)

21	33050849	ALU_12379	rs17660708	LDL cholesterol	1.00E-10	0.8
17	44153977	SVA_706	rs9303525	Intracranial volume	8.00E-15	0.8
20	26190974	ALU_12132	rs6051320	Lung function	2.00E-08	0.8
11	54958589	ALU_8580	rs77584654	Height	5.00E-17	0.8
7	18273084	ALU_5868	rs1528683	Lung function	2.00E-17	0.8
12	77965056	ALU_9355	rs17788937	Pathological Myopia	2.00E-12	0.8
11	43877448	ALU_8559	rs1061810	Type 2 diabetes	4.00E-10	0.8
8	110101605	ALU_7037	rs28499085	Pulse pressure	3.00E-13	0.8
14	92619420	SVA_615	rs34016308	Myopia	4.00E-14	0.8
5	40041345	LINE1_1097	rs10053502	Pathological Myopia	1.00E-16	0.8
1	174484646	ALU_604	rs75650221	Chronotype	4.00E-18	0.8
5	25233926	ALU_4154	rs111257433	General risk tolerance	5.00E-10	0.8
17	44153977	SVA_706	rs62062288	Risk-taking tendency	1.00E-29	0.8
6	56387576	ALU_5205	rs4288197	Heel bone mineral density	5.00E-17	0.8
11	49282683	ALU_8572	rs7103270	HDL cholesterol and physical activity interaction	7.00E-12	0.8
4	76993824	ALU_3412	rs7693693	Blood protein levels	2.00E-17	0.8
1	219558910	ALU_810	rs75128958	Heel bone mineral density	1.00E-08	0.8
1	219558910	ALU_810	rs75128958	Lung function	2.00E-23	0.8
2	30669993	ALU_1087	rs28538173	Eosinophil counts	3.00E-09	0.8
6	96009421	ALU_5389	rs80268500	Blood protein levels	2.00E-12	0.8
16	80848077	ALU_11145	rs34018670	Monocyte count	5.00E-09	0.8
1	119553366	LINE1_122	rs3790553	Male-pattern baldness	4.00E-19	0.8
17	44153977	SVA_706	rs76640332	Lymphocyte percentage of white cells	5.00E-13	0.8
3	193354185	LINE1_769	rs11925699	Educational attainment	3.00E-08	0.8
1	180857564	ALU_629	rs1043069	Systolic blood pressure	5.00E-15	0.8
9	4237141	ALU_7235	rs2224492	Intraocular pressure	4.00E-16	0.8
15	47507342	LINE1_2640	rs6493265	Educational attainment	2.00E-17	0.8
15	47507342	LINE1_2640	rs12914084	Neuroticism	3.00E-08	0.8
2	652672	ALU_958	rs13021737	Body mass index	8.00E-40	0.8
2	652672	ALU_958	rs12995480	C-reactive protein levels	1.00E-10	0.8
2	652672	ALU_958	rs12714415	Heel bone mineral density	4.00E-09	0.8
2	652672	ALU_958	rs6752706	Lung function	2.00E-13	0.8
2	652672	ALU_958	rs5017302	Menarche (age at onset)	5.00E-38	0.8
2	652672	ALU_958	rs13396935	Smoking status	4.00E-13	0.8
8	109135936	SVA_389	rs617117	Macular thickness	2.00E-09	0.8
12	28163331	ALU_9104	rs1838564	Breast size/Breast Cancer	1.00E-12	0.8
4	22043212	ALU_3139	rs62301574	Insomnia	1.00E-08	0.8
14	39875097	LINE1_2544	rs34983854	Systolic blood pressure	2.00E-11	0.8
6	140417842	ALU_5637	rs62429521	Insomnia	2.00E-09	0.8
14	55795871	ALU_10325	rs10146637	White blood cell count	4.00E-11	0.8
6	97017683	ALU_5395	rs11153071	Systolic blood pressure	3.00E-15	0.8
8	109135936	SVA_389	rs392783	Hair color	2.00E-23	0.8
15	49609604	ALU_10695	rs11632038	Lung adenocarcinoma	5.00E-10	0.8
11	49282683	ALU_8572	rs77828979	Intraocular pressure	6.00E-12	0.8
11	49282683	ALU_8572	rs11040595	Systolic blood pressure	1.00E-11	0.8
2	30669993	ALU_1087	rs829636	Eczema	6.00E-09	0.8
10	34571038	ALU_7901	rs610493	Height	2.00E-10	0.8
8	71914591	ALU_6806	rs2639935	Lung function	3.00E-08	0.8

4.3.5. The distribution of RTEs in LD with TAS in functional genomic regions:

The distribution of RTEs in LD ($r^2 > 0.6$) with TAS in functional genomic regions was compared with the distribution of all non-reference RTE elements to investigate the potential effect of RTEs in LD with TAS on gene function and regulation. About two-thirds (59.9%) of the RTEs in LD with TAS were found in gene regions (Table 18). A total of 10 L1s overlapped with intronic regions of genes, of which 4 were on the sense strand of the gene. Note that RTE insertions on the sense strand of genes have a greater impact on gene function and regulation as RTEs contain internal promoters and 3'-poly(A) tails that can interfere with gene transcription and post-transcriptional regulation (Guffanti et al., 2014; Elbarbary et al., 2016). Of the 113 Alus in LD with TAS, 64 were found in the intron regions of genes, 3 interrupted the 5' or 3' untranslated regions (UTR), and 2 Alus were less than 10 Kb upstream of gene regions. Twenty-nine of the 67 Alu elements in gene regions were on the sense strand of the gene. Five SVA elements interrupted intronic genomic regions, one of which was on the sense strand of the gene. Nevertheless, the distribution of L1 and SVA elements in LD with TAS in gene regions is not significantly different from the distribution of all non-reference RTEs in these genomic regions (P-value = 0.135 and 0.426, chi-squared test). However, Alu elements are significantly more frequent in gene regions compared with the distribution of all non-reference AluY elements (P-value = 0.0143, chi-squared test).

Table 18: A list of the overlap between gene regions and RTEs in LD ($r^2 > 0.6$) with TAS. The highlighted RTEs are on the sense strand of the gene. *Long non-coding RNA. Abbreviations: RTE: Retrotransposable element; TAS: Trait associated SNP; LD: Linkage disequilibrium.

Chr	Start_RTE	End_RTE	RTE Strand	RTE	Gene name	Gene Strand	Gene region
1	169524859	169524860	+	LINE1_164	F5	-	Intron
2	144010793	144010794	+	LINE1_410	ARHGAP15	+	Intron
3	55788580	55788581	+	LINE1_590	ERC2	-	Intron
3	85576571	85576572	-	LINE1_629	CADM2	+	Intron
3	193354185	193354186	-	LINE1_769	OPA1	+	Intron
6	46310306	46310307	+	LINE1_1293	RCAN2	-	Intron
6	46310306	46310307	+	LINE1_1293	LOC101926915*	+	Intron
7	8019027	8019028	+	LINE1_1448	GLCCI1	+	Intron
9	94058487	94058488	-	LINE1_1863	AUH	-	Intron
14	39875097	39875098	+	LINE1_2544	FBXO33	-	Intron
15	47507342	47507343	-	LINE1_2640	SEMA6D	+	Intron
1	174484646	174484647	-	ALU_604	RABGAP1L	+	Intron
1	180857564	180857565	+	ALU_629	XPR1	+	UTR
1	198243300	198243301	-	ALU_726	NEK7	+	Intron
1	227502452	227502453	+	ALU_841	CDC42BPA	-	Intron
1	232587774	232587775	+	ALU_865	SIPA1L2	-	Intron
2	11353711	11353712	+	ALU_1002	ROCK2	-	Intron
2	98582157	98582158	+	ALU_1385	TMEM131	-	Intron
2	141534074	141534075	-	ALU_1585	LRP1B	-	Intron
2	181880746	181880747	-	ALU_1805	UBE2E3	+	Intron
2	198763462	198763463	-	ALU_1894	PLCL1	+	Intron
3	42898420	42898421	-	ALU_2319	ACKR2	+	Intron
3	152053972	152053973	+	ALU_2814	MBNL1	+	Intron
3	157962934	157962935	-	ALU_2845	RSRC1	+	Intron
3	158089835	158089836	-	ALU_2846	RSRC1	+	Intron
3	168885760	168885761	+	ALU_2909	MECOM	-	Intron
4	76993824	76993825	+	ALU_3412	ART3	+	Intron
5	25233926	25233927	+	ALU_4154	LINC02211	+	Intron
5	109051004	109051005	-	ALU_4562	MAN2A1	+	Intron
6	45260479	45260480	+	ALU_5132	SUPT3H	-	Intron
6	53167475	53167476	+	ALU_5181	ELOVL5	-	Intron
6	56387576	56387577	-	ALU_5205	DST	-	Intron
6	66163982	66163983	-	ALU_5227	EYS	-	Intron
6	74504855	74504856	+	ALU_5280	CD109	+	Intron
6	96009421	96009422	-	ALU_5389	MANEA-AS1	-	Intron
6	97017683	97017684	-	ALU_5395	FHL5	+	Intron
6	139294734	139294735	-	ALU_5628	REPS1	-	Intron
6	163013855	163013856	-	ALU_5742	PRKN	-	Intron

(Table 18 continues)

7	18273084	18273085	-	ALU_5868	HDAC9	+	Intron
7	33195329	33195330	-	ALU_5942	BBS9	+	Intron
7	78146522	78146523	-	ALU_6127	MAGI2	-	Intron
7	91751552	91751553	+	ALU_6200	CYP51A1	-	Intron
7	119259819	119259820	-	ALU_6325	LINC02476	-	UTR
8	8920127	8920128	+	ALU_6560	ERI1	+	Intron
8	13975433	13975434	-	ALU_6584	SGCZ	-	Intron
8	63344481	63344482	+	ALU_6774	NKAIN3	+	Intron
8	100782579	100782580	+	ALU_6981	VPS13B	+	Intron
8	110101605	110101606	-	ALU_7037	TRHR	+	Intron
9	4237141	4237142	+	ALU_7235	GLIS3	-	Intron
9	16682313	16682314	-	ALU_7311	BNC2	-	Intron
9	117928281	117928282	+	ALU_7672	DEC1	+	Intron
10	3569025	3569026	+	ALU_7750	LOC105376360	+	Intron
10	11984965	11984966	+	ALU_7788	UPF2	-	Intron
10	34571038	34571039	+	ALU_7901	PARD3	-	Intron
10	46074893	46074894	-	ALU_7934	MARCH8	-	Intron
10	65356114	65356115	-	ALU_8023	REEP3	+	Intron
10	106566893	106566894	-	ALU_8208	SORCS3	+	Intron
11	43877448	43877449	-	ALU_8559	HSD17B12	+	UTR
11	65984338	65984339	-	ALU_8622	PACS1	+	Intron
12	26697612	26697613	-	ALU_9091	ITPR2	-	Intron
12	28417298	28417299	-	ALU_9107	CCDC91	+	Intron
12	28438612	28438613	-	ALU_9108	CCDC91	+	Intron
12	41847723	41847724	+	ALU_9169	PDZRN4	+	Intron
12	56753252	56753253	-	ALU_9228	STAT2	-	Intron
12	71525479	71525480	+	ALU_9320	TSPAN8	-	Intron
12	120130849	120130850	+	ALU_9540	CIT	-	Intron
13	46647748	46647749	-	ALU_9748	CPB2	-	Intron
13	46647748	46647749	-	ALU_9748	CPB2-AS1	+	Intron
13	62588972	62588973	-	ALU_9847	LINC00358	-	Intron
14	55795871	55795872	-	ALU_10325	FBXO34	+	Intron
14	60741400	60741401	+	ALU_10351	PPM1A	+	Intron
15	49609604	49609605	+	ALU_10695	GALK2	+	Intron
15	73983319	73983320	-	ALU_10812	CD276	+	Intron
15	76826019	76826020	+	ALU_10819	SCAPER	-	Intron
16	75655176	75655177	-	ALU_11116	ADAT1	-	Intron
17	46505002	46505003	+	ALU_11333	SKAP1	-	Intron
18	53146075	53146076	-	ALU_11714	TCF4-AS1	+	Intron
18	53146075	53146076	-	ALU_11714	TCF4	-	Intron
20	1546228	1546229	+	ALU_12014	SIRPB1	-	Intron
21	33050849	33050850	+	ALU_12379	SCAF4	-	Intron
2	174952231	174952232	-	SVA_134	OLA1	-	Intron
6	153429856	153429857	+	SVA_315	RGS17	-	Intron
9	33130564	33130565	+	SVA_401	B4GALT1	-	Intron
14	92619420	92619421	-	SVA_615	CPSF2	+	Intron
17	44153977	44153978	+	SVA_706	KANSL1	-	Intron
Upstream of Gene							
6	140417842	140417843	+	ALU_5637	LOC100507477	+	3 kb Upstrea
16	80848077	80848078	-	ALU_11145	CDYL2	-	10 kb Upstre

Twenty Alu elements and 1 SVA element interrupted enhancer regions (Table 19), however, the distribution of RTEs in LD with TAS in enhancer regions is not significantly different from the distribution of all non-reference RTEs (P-value = 0.069 and 0.908, respectively; chi-squared test).

Table 19: List of RTEs in LD with TAS that overlap with enhancer regions. Enhancer regions and names are based on data from the GeneHancer database of enhancers (Fishilevich et al., 2017). Abbreviations: RTE: Retrotransposable element; TAS: Trait associated SNP.

Chr	RTE start	RTE end	RTE name	Enhancer Start	Enhancer End	GeneHancer name
1	78607067	78607068	ALU_276	78606737	78609541	GH01J078141
1	198243300	198243301	ALU_726	198241174	198243363	GH01J198272
1	227502452	227502453	ALU_841	227501737	227507587	GH01J227314
1	232587774	232587775	ALU_865	232586521	232589442	GH01J232450
2	30669993	30669994	ALU_1087	30669326	30671855	GH02J030446
4	76993824	76993825	ALU_3412	76993459	76994847	GH04J076072
5	56109723	56109724	ALU_4294	56109413	56115758	GH05J056814
6	53167475	53167476	ALU_5181	53164103	53169223	GH06J053299
6	139294734	139294735	ALU_5628	139291022	139295452	GH06J138969
7	38209213	38209214	ALU_5970	38208802	38210018	GH07J038169
7	50473286	50473287	ALU_6027	50472233	50475870	GH07J050404
7	120538086	120538087	ALU_6336	120536766	120540791	GH07J120896
9	16682313	16682314	ALU_7311	16682312	16684265	GH09J016682
10	3569025	3569026	ALU_7750	3567620	3571242	GH10J003525
10	34571038	34571039	ALU_7901	34568588	34571868	GH10J034279
10	106566893	106566894	ALU_8208	106565146	106570717	GH10J104805
12	56753252	56753253	ALU_9228	56752000	56755336	GH12J056358
12	120130849	120130850	ALU_9540	120128986	120132955	GH12J119691
16	75655176	75655177	ALU_11116	75653160	75658011	GH16J075619
17	46505002	46505003	ALU_11333	46503573	46508375	GH17J048426
9	33130564	33130565	SVA_401	33127963	33132882	GH09J033127

4.4. Discussion

4.4.1. RTEs overlap and enrichment in GWAS risk loci

This study adopted an approach for identifying polymorphic RTEs that have the potential to implicate human health and influence individual predisposition to disease by screening for RTEs in risk regions previously identified by GWAS. Over one-third of the common L1s and Alus (36.64% and 39.82%, respectively) and more than half of the SVAs (57.27%) analysed in this study overlapped with TAS LD-blocks (Table 16). These RTEs were significantly enriched in risk loci compared with random genomic regions (Figure 20). The enrichment of structural variants (SVs) mediated by RTE insertions in risk loci is in line with previous reports in the literature. Sudmant et al. (2015) reported a 1.5-fold enrichment of TAS in the genomic region surrounding SVs, of which some were RTEs. However, the majority of the enrichment signal seemed to be attributed to large SV (>20 Kb) (Sudmant et al., 2015). A more recent study reported a significant enrichment of polymorphic Alu elements in GWAS risk regions (P-value = 0.013) using a set of 13,572 polymorphic Alus and 3,242 TAS LD-blocks ($P < 10^{-9}$) (Payer et al., 2017). These results are in line with the enrichment of Alus in TAS LD-blocks observed in this study. The significant enrichment of L1 and SVA elements in the genomic regions of TAS reported in this study (Figure 20) adds to previous reports in the literature and suggests the RTEs in risk loci are candidate causative variants with the potential to impact functional genomic regions.

4.4.2. LD analysis in comparison to previous reports in the literature

Polymorphic RTEs with a potential functional impact in risk regions are likely in LD with the TAS of that region. RTEs that are not in LD with the TAS are potentially not on the risk haplotype and are unlikely candidates of causation. The association analysis identified 429 RTE-TAS LD associations ($r^2 > 0.6$) using GWS TAS that have been identified with a P-value $\leq 5x10^{-8}$, and RTE variants of low and common AF (MAF ≥ 0.01) that have been identified in European populations. Of these, 157 TAS are in strong LD ($r^2 \geq 0.8$) with 49 RTEs (Table 17). Previous studies conducting similar analysis report a total of 164 RTE-TAS LD pairs ($r^2 > 0.6$) of which 41 RTE-TAS associations were replicated by this study (Table 20; Appendix 3). Some previously reported RTE-TAS associations not replicated by this study were either identified in non-European cohorts, were not GWS, or were within the HLA region. Other RTE-TAS associations may have been missed because of different studies using genotype data of different samples, thus affecting the r^2 of the LD calculation (Wray, 2005; Medina-Gomez et al., 2015). Another reason for missing some of the previously identified RTE-TAS associations might be because some of the LD-blocks of this study were arbitrarily extended using the median of the LD-blocks defined by tagging SNPs. This may have under-estimated the size of the LD-block of some TASs which may have resulted in missing some RTEs from the association analysis, especially since the LD analysis in this study was mainly calculated for RTEs interrupting a TAS LD-block. Nevertheless, this study adds a list of 354 new RTE-TAS associations to previous reports in the literature.

Table 20: Comparing the list of polymorphic RTEs in LD ($r^2 > 0.6$) with GWAS TAS identified by previous similar studies with the list of RTE-TAS associations identified by the current study. Abbreviations: RTE: Retrotransposable element; TAS: Trait associated SNP; LD: Linkage disequilibrium. GWS: Genome-wide significant, defined as TAS with a P-value ≤ $5x10^{-8}$.

Study reference [PMID]	Total number of TAS in LD with at least one RTE identified by the referenced study	Number of TAS in LD with RTEs that overlap with the current study	Number of discrepancies (reason)
Sudmant et al. (2015) [26432246]	6	5	1 (TAS not GWS)
Hehir-Kwa et al. (2016) [27708267]	43	13	16 (TAS not GWS) 10 (TAS not identified in European cohort) 2 (TAS in HLA region) 2 (other)
Payer et al. (2017) [28465436]	64	16	3 (TAS not GWS) 5 (TAS not identified in European cohort) 40 (other)
Wang et al. (2017) [28824558]	51	7	23 (TAS not GWS) 8 (TAS not identified in European cohort) 13 (other)

4.4.3. Distribution of RTEs in LD with TAS:

Of the RTEs in LD with TAS, 84 are in or near gene regions, including 36 RTEs that are on the sense strand of the gene (Table 18). In addition, 21 of the RTEs in LD with TAS are in enhancer regions (Table 19). The distribution of L1 and SVA elements in LD with TAS in functional genomic regions is not significantly different relative to all polymorphic L1s and SVAs in the curated database (P-value > 0.05). In contrast, polymorphic Alu elements in LD with TAS are significantly enriched in gene regions (P-value = 0.0143) in line with the results

of Payer et al. (2017). The relative enrichment of Alu elements in LD with TAS within genes relative to all polymorphic Alus is expected due to the enrichment of TAS in genes (Hindorff et al., 2009; Payer et al., 2017). The lack of enrichment of L1s and SVAs in LD with TAS within genes relative to all polymorphic L1 and SVA elements may have been missed due to their small sample size of L1s (n=16) and SVAs (n =7) in LD with TAS, which is likely to increase type II error (Banerjee et al., 2009). Nevertheless, given that RTE elements have been known to interfere with human gene expression through a variety of mechanisms (Chuong et al., 2017; Bourque et al., 2018), RTEs in LD with TAS that also interrupt gene regions are likely candidates of causation with the potential to interfere with gene function and expression.

4.4.4. RTEs as causative variants affecting gene expression:

A study by Wang et al. (2017) confirmed the potential of RTEs in LD with TAS as candidate causative variant. Wang et al. (2017) identified 437 RTEs identified in European samples and associated with disease. These RTEs were also located within tissue-specific enhancers. Expression quantitative trait loci (eQTL) analysis was then performed using human B-cells and the set of RTE-TAS associations within enhancers of blood and immune tissue. By using this method, the Wang et al. (2017) research group were able to identify seven RTEs in LD with TAS and associated with the expression of genes of importance to the trait phenotype.

An example of an RTE from the list of RTE-TAS associations reported in this study with the potential to act as a causative variant is SVA_706, located on chromosome 17q21.31 on the opposite strand of the KANSL1 gene. This SVA is in LD with 90 TAS linked with a variety of traits including Parkinson's disease,

progressive supranuclear palsy (PSP), cognitive ability, depression, and anxiety (Table 17).

A recent study by Spirito et al*.,* (2019) reported the significant association between SVA_706 and the expression of a number of genes including KANSL1 gene and its antisense transcript KANSL1-AS1, Corticotropin Releasing Hormone Receptor 1 (CRHR1) gene, and Leucine-Rich Repeat Containing 37A2 (LRRC37A2) gene. These genes have been previously associated with a number of physiological processes including PSP, depression, and anxiety (Liu et al., 2013; Allen et al., 2016; Ferrari et al., 2017). These findings confirm that RTE variants do have the potential to impact human health and can influence individual predisposition to disease.

4.4.5. Study overview and concluding remarks

This chapter discusses the candidacy of polymorphic RTE insertions as potential causative variants in GWAS risk loci. Polymorphic L1s, Alus, and SVAs with common allele frequencies in populations of European descents were found to be overrepresented at GWAS risk loci, more than expected by chance. In addition, hundreds of RTEs located within risk loci were found to be in LD with TASs. Thus the findings of the study at hand build on previous reports in the literature, supporting the potential contribution of RTEs as risk variants associated with complex diseases. It is noteworthy that association does not necessarily equate to causation, therefore the RTE-TAS associations identified in this study are merely candidate causative variants requiring future investigations to uncover whether their existence within risk haplotypes does have an observed effect on genome function that can be related to the trait phenotype.

5. General discussion and future directions

Transposable elements (TEs) make up a huge percentage of the human genome (~45%) and have been a component of our DNA throughout the history of human evolution (Lander et al., 2001). TEs were first discovered in the genome of maize by Barbara McClintock in the 1950s (Ravindran, 2012). However, the effect of TE activity on genome function in humans was not realised until the late 1980s, following the first reported case of a genetic disease, namely Haemophilia, caused by an L1 insertion into exon 14 of the factor VIII gene in two unrelated individuals (Kazazian, 1988). Since then, over 124 monogenic diseases caused by recent TE activity have been reported in the literature (Hancks and Kazazian, 2016). Despite this, the general effect of TE activity on genome function and its potential contribution towards multifactorial traits and disorders remains an open question. This thesis set out to investigate the effect of the ongoing transposition activity from the active L1, Alu, and SVA subfamilies on genome function, including investigating the candidacy of RTE-mediated genomic variants as potential causative variants that can influence individual predisposition to complex traits and diseases. To this end, a comprehensive database of known non-reference RTEs that are polymorphic in the human genome was curated from peer-reviewed journal articles scattered through the literature. The database was utilised to investigate the genomic distribution of non-reference RTEs in comparison to ancient RTE insertions that are fixed in the genome of all humans. The results of the distribution analysis show that polymorphic RTE variants are found in active genomic regions more frequently than ancient RTEs that are fixed in the human genome. These results suggest that the activity of RTE elements do harm genome function and thus are subjects of negative selection. As such,

RTE variants that are potentially deleterious and disease-causing were sought after in the next step of the analysis.

Firstly, the enrichment of RTEs in close proximity to risk regions was investigated. It was found that L1s, Alus, and SVAs were significantly enriched in GWAS risk loci, compared with random genomic regions of similar properties. We next tried to identify RTE variants that have the potential to be the causative variant within the risk haplotype, by calculating LD between the RTE variant and the TAS. This analysis resulted in a list of 354 RTE-TAS associations, each of which has the potential to be the causative variant within the haplotype of the TAS. This thesis employed a constructed framework for identifying RTE variants of relevance to disease risk by utilising public data that is available in the literature. The next sections will discuss each component of this thesis in more detail, including study limitations and proposed future studies.

5.1. Database curation

Identifying RTE insertions from the recently evolved and active subfamilies of L1s, Alus, and SVAs has been a difficult task, due to their repetitive nature and high sequence homology with endogenous RTEs that have been fixed in the human genome throughout human evolution. The development of short-read NGS technology and numerous RTE detection tools have simplified and replaced early labour-intensive methods and drastically increased the scalability of RTE detection. Since then, more and more research groups have shown a great interest in RTE discovery as numerous studies characterising polymorphic RTE insertions have been published. These studies have collectively reported thousands of RTE variants, of which only a fraction have been systematically

organised into online databases, while the majority remain scattered in the literature. As such, there was a need for a comprehensive database of RTE variants, to serve as a resource that could be used to facilitate large-scale genomic studies, including population genetics and association studies. This study reports a comprehensive list of L1, Alu, and SVA variants that have been reported in the literature up to April 2019. The current version of the curated database includes the insertional profiles of 3,360 samples and 39,798 RTEs, obtained from a total of 45 studies.

In contrast to previous online databases, namely dbRIP (http://dbrip.brocku.ca/; Wang et al., 2006) and the euL1db (http://eul1db.unice.fr; Mir et al., 2014), the curated database applied quality control measures to ensure the removal of potential false positives. In addition, the database of this study only includes germline variants that have been identified in healthy human samples, which is in contrast to the dbRIP and the euL1db databases (Wang et al., 2006; Mir et al., 2014), that include somatic insertions or insertions identified in pathological human samples. Nevertheless, 52% of dbRIP records and 46% of the germline L1Hs reported in the euL1Hs database overlapped with or were within 200bp of the RTE data included in the curated database of this study.

Overall, the curated database provides a comprehensive resource of known germline polymorphic RTE insertions that can be a vital resource in the study of the physical and pathological impact of recent RTE activity on the human genome.

5.1.1. Reflections and limitations

It was noted during the database curation process that different studies use various nomenclature to describe RTE insertions. Examples of terms (other than RTEs) that have been used by some studies to describe L1, Alu, and SVA insertions include: mobile element insertions, retrotransposons, and retroelements. Although the scientific community recognises that all these terms are synonyms for one another, PubMed searches for the different terms produce a different number of results. There are ways of getting around this, for example, by using MeSH terms or advanced information retrieval functions in PubMed to formulate a query. However, a recent study analysing PubMed user sessions has reported that over 94% of information queries were performed by inexperienced users, defined as users who do not utilise PubMed advanced information retrieval functions (Yoo and Mosa, 2015). As such, many scientists conducting a PubMed search of non-reference RTE insertions may misidentify several known non-reference elements. This could result in future RTE detection studies misidentifying known non-reference insertions as novel insertions, and further illustrating the importance and the need for a comprehensive and accessible database of RTE elements.

Variations within the RTE database of this study are bound to exist owing to differences within the methods applied by each study, including variations within the algorithmic design of the different RTE detection tools, and the applied quality control measures for RTE calling. Such variations have likely had different effects on the accuracy, precision, and false discovery rate of the included studies, thus limiting the consistency of the curated database. Nevertheless, every effort was made to minimise the potential effects of this shortfall, for example, by applying

a minimum supporting reads threshold to eliminate false positives, and by merging insertions located within 200bp of each other, to accommodate for variations within the precision of the different RTE detection tools.

5.2. Genomic distribution of RTEs

The genomic location of an RTE insertion can determine its potential impact on genome function and its ability to be expressed and mobilised. RTEs are capable of introducing insertional mutagens that can affect genome function in a variety of mechanisms, including interfering with gene expression, alternative splicing, and post-transcriptional regulation (Guffanti et al., 2014; Bourque et al., 2018; Savage et al., 2019). In addition, RTEs can impact genomic stability, by inducing DNA breakage during their retrotransposition or by causing post-transposition deletions and duplications via NAHR, which can occur due to the repetitive nature and high sequence homology between the different subfamilies of RTEs (Startek et al., 2015; Nazaryan-petersen et al., 2016; Bourque et al., 2018). RTEs can also induce epigenetic modifications by attracting chromatin modification complexes as part of the host defence mechanisms to suppress RTE activity (Jacobs et al., 2014; Garcia-Perez et al., 2016). Therefore, characterising the integration site of RTEs is essential for understanding the potential impact of their activity on genome function and integrity, including the extent of their contribution to human health and disease.

Previous studies investigating the genomic distribution of endogenous RTEs have reported the accumulation of L1 elements in AT-rich, low activity genomic regions, while Alu and SVA elements accumulate in GC-rich, high-activity regions (Smit, 1999; Lander et al., 2001; Wang et al., 2005). Nevertheless, endogenous

RTEs are mostly fixed in the human genome and are thought to have a neutral effect on genome function. The differential distribution of Alu and SVA elements, in comparison to the distribution of L1s despite their transposition via the L1 machinery, has been attributed to post-integrational processes that have differentially reshaped the genomic landscape of each element, mainly by purifying selection and ectopic recombination (Medstrand et al., 2002; Wang et al., 2005; Kvikstad and Makova, 2010; Costantini et al., 2012). Polymorphic RTE insertions that are not part of the human reference genome have not been subjects of selection pressures to the same extent as the fixed endogenous elements. Accordingly, analysing the genomic distribution of polymorphic insertions, in comparison with the landscape of fixed RTEs, is one of the methods that can be used to gain insight into the interplay between RTE activity and its potential impact on genome function and integrity. Previous studies conducting similar analysis have provided some insight into the potential effects of the continuous RTE activity on genome function, but provided limited information about integration site preference, due to the low number of polymorphic insertions recovered from an even lower number of samples (Ovchinnikov et al., 2001; Boissinot et al., 2004; Wang et al., 2005; Cordaux et al., 2006; Ewing and Kazazian, 2010). The small number of samples analysed by previous studies means that the majority of the recovered insertions were likely common, with an allele frequency >1%. In short, retrotransposition is induced *in vitro* using an engineered RTE construct, and the distribution of the recovered *de novo* insertions is then compared with that of endogenous elements (Flasch et al., 2019; Sultana et al., 2019; Chen et al., 2020). At the moment, this method has only been applied for investigating the integration preference of L1 elements.

This study compared the genomic distribution of endogenous RTEs with the distribution of polymorphic non-reference insertions that were curated in-house. Near one-third of the in-house curated database (32%) is composed of singleton insertions, that are more likely to provide information about the integration site preference of active RTEs, due to their younger evolutionary age compared with common polymorphic insertions. As expected, the results for the genomic distribution of L1 elements from this study were intermediate between those reported by previous studies using non-reference insertions and the more recent studies using engineered L1 constructs. Thus, it was deduced that the distribution results of non-reference insertions from this study represent aspects of the integration site preference of L1, Alu, and SVA elements.

The initial integration of non-reference RTEs is not completely random. Non-reference L1 and Alu elements are similarly distributed in AT-rich low-activity regions, suggesting that both of these elements share similar integration site preferences. A recent study has reported that the most influential factors in determining the integration site preference of active L1s include the specificity of the L1 machinery and pre-existing biases within the human genome (Sultana et al., 2019). These influential factors are also likely to be true for determining the initial integration site of Alus, especially since the retrotransposition of mobile Alu elements is reliant on the L1 machinery. In contrast, SVA elements accumulate in GC-rich high-activity regions, suggesting the role of different influential factors in determining the integration site preference of SVAs, which are yet to be determined.

Overall, the differential distribution of reference vs. non-reference RTEs displays aspects of the interplay between the integration site preference and the effect of

post-integrational processes in modulating the genomic landscape of RTE insertions. In addition, the more frequent occurrence of non-reference RTE insertions in functional genomic regions, in comparison with reference insertions, suggests the potential negative impact of RTE activity on genome function and integrity. This study provides the most recent analysis for the genomic distribution of Alu and SVA elements that have not been studied as extensively as L1 elements.

5.2.1. Study limitations

The non-reference SVA elements analysed by this study may not be representative of the true integration site preference of SVA elements. A recent pedigree-based study investigating the rate of new RTE elements has estimated the occurrence of 1 new L1 and SVA element per 63 births (Feusier et al., 2019). However, the number of non-reference SVA insertions recovered so far is a fraction of the number of known non-reference L1 elements, suggesting shortfalls of SVA discovery within the human genome. In addition, recent germline SVA insertions that have been identified in the population may have already been exposed to strong and rapid post-insertional selection via compositional matching. The compositional matching hypothesis suggests the removal of DNA elements, particularly repetitive elements with a sequence composition that does not match the isochore in which they are found (Pavlíček et al., 2001; Hackenberg et al., 2005). These possibilities limit the reliability of conclusions drawn by this study for the integration site preference of SVA elements.

Finally, the distribution analysis of this study provided some great insight into the potential impact of RTE activity on genome function. However, aspects of the integration site preference may have been masked by the grouping of

polymorphic RTEs of different allelic frequencies, thereby limiting the scope of the conclusions drawn.

5.3. Correlation between RTEs and TASs

RTEs that remain capable of transposing within the human genome are an important source of genomic diversity within human populations, yet their contribution to complex traits and diseases remains an open question. RTEs are often masked in genomic and association studies, despite their continuous contribution towards creating SV in humans, and their various effects on genome function and regulation. Previous studies may have neglected RTE insertions, due to their repetitive nature and high sequence homology, which makes them difficult to characterise and study. Many studies filter out repetitive regions including TEs from WGS data before conducting downstream analyses (Goerner-potvin and Bourque, 2018). As such, previous association studies have mainly relied on SNPs, as they are ubiquitous in the human genome, occurring on average once every 300 base pairs (Nelson et al., 2004). However, the collective effect size of GWS SNPs could only explain a fraction of the trait heritability for many complex traits, thus the missing heritability issue was raised (Manolio et al., 2009; Rowe and Tenesa, 2012). Soon after the missing heritability issue was raised, more and more studies started to incorporate other types of SVs, however, these studies were mainly focused on CNVs or mega base deletions and duplications (Stankiewicz and Lupski, 2010; Lacaria et al., 2013; Waddell et al., 2015). Studies have only recently begun investigating the contribution of RTEs to complex traits and diseases, thus the impact of RTE variants as potential causative candidates that can influence individual predisposition to disease, remains largely uncharacterised.

A recent study investigating the potential implications of polymorphic RTE insertions on human health has reported a significant enrichment of polymorphic Alu elements in GWAS risk loci (Payer et al., 2017). Also, recent studies have reported that polymorphic RTE insertions can be in strong LD with TAS, suggesting their potential candidacy as causative variants within the risk haplotype (Sudmant et al., 2015; Hehir-Kwa et al., 2016; Payer et al., 2016; Wang et al., 2017). In addition, some RTE variants in LD with TAS were also shown to be associated with altered gene expression of nearby genes in a tissue-specific manner (Wang et al., 2017; Spirito et al., 2019). This study aimed to build upon previous findings by investigating the proximity and enrichment of polymorphic RTEs in GWAS risk loci, and by identifying additional RTEs in LD with TAS that could be potential candidates of causation within the risk haplotype. The analysis of this study is focussed on exploring the effect of common variants from the three active classes of non-LTR RTEs, namely L1s, Alus, and SVAs, and their potentially uncharacterised contribution as underlying causative variants of complex traits.

L1 and SVA elements were found to be significantly enriched in GWAS risk loci, suggesting the detection of the potential functional impact of polymorphic RTE insertions at multiple GWAS risk regions. The significant enrichment of Alu elements in GWAS risk loci is in line with the results of Payer et al., 2016. RTE elements were readily identified in GWAS risk loci. This study also identified 354 new RTE-TAS associations, of which 157 are in strong LD (r2 > 0.8; Table 17). Most significantly, a non-reference SVA element (SVA_706) was found to be in strong LD with multiple TAS over a large genomic region. Upon further investigation, it was found that this SVA is located within a large haplotype in European populations that have been extensively studied in neurological

diseases, yet its contribution to disease risk has not yet been fully elucidated upon (Boettger et al., 2012; Li et al., 2014; Koolen et al., 2016). Interestingly, Spirito et al. (2019) reported strong associations between this SVA element and the expression of several genes that have been implicated with neurological conditions such as depression and anxiety. Altogether, these results suggest that SVA_706 is likely the missing variant that could potentially explain the contribution of this large haplotype to the risk of developing various neurological diseases, thus its contribution should be validated with follow-up wet-lab studies.

Overall, the results of our study and previous studies show RTEs can be causative variants in potentially all complex traits and disorders, and should be routinely incorporated in association studies to address aspects of the missing heritability issue.

5.3.1. Study limitations

Estimating the LD-block size for TAS that lacked tagging SNPs, may have introduced type II error during the LD-analysis, resulting in the misidentification of some RTE-TAS associations that were reported in previous studies. The LD-analysis was limited to RTEs identified by the 1kGP (Sudmant et al., 2015) as it is the only study within the curated database that included accessible genotype information of SNPs and RTE elements. The LD-analysis was also limited to variants identified in individuals of European descent, as most GWAS have been conducted predominantly on European cohorts. Finally, some of the RTEs in LD with TAS reported in this study may not be associated with disease risk. Nevertheless, the list of RTE-TAS associations reported in this study identifies some RTEs that are potential candidates of causation. The contribution of the

RTE variant as a potential risk variant from the reported list can only be confirmed by wet-lab experiments.

5.4. Summary of key findings

This study investigated the genomic landscape of recent RTE insertions and the potential contribution of polymorphic RTEs as causative variants of disease in GWAS risk loci, in an effort to understand the impact of RTE insertions and their role in complex human disorders. To this end, a comprehensive database of polymorphic RTE insertions was curated using online databases and peer-reviewed journal articles. The curated database is a handy resource for future studies in various fields, including population genetics and cancer genomics, made available at GitHub (https://github.com/RandaAli1/MyPhDproject/tree/master/MyAnalysis/RTE_files)

Through investigating the genomic distribution of RTEs, we have found that transposable elements are readily located in active genomic regions, and thus have the potential to influence gene function and regulation in a variety of mechanisms. Of the three RTE types investigated in this study, SVA elements accumulated the most in active regions, suggesting that SVAs are likely to have the most negative impact on genome function. This finding is significant as it highlights the importance of conducting more research on SVA elements, since they are the least favoured RTEs to study due to their complex structure, which makes them difficult to detect in the human genome.

Finally, we investigated the potential association of RTEV with disease susceptibility and found a significant enrichment of RTEs in GWAS risk loci, plus over 400 RTEs in LD (r2 > 0.6) with various TAS. This finding suggests that SV mediated by RTE insertions do have the potential to impact complex human traits

and are likely causative variants in some GWAS risk loci. Although the RTE-TAS associations identified in this study are purely candidate variants of causation requiring future validation studies, the result of this analysis highlights the importance of including RTEV in future association studies to better the understanding of complex human traits and disorders.

5.5. Future directions

RTEs provide a continuous source of genetic variation in humans, including some variants that can be involved with the aetiology of complex traits. Following the results of this study, it was concluded that SVA elements have the highest potential impact on the function of the human genome relative to other classes of active RTEs, namely L1 and Alu elements. This conclusion was based on the accumulation of non-reference SVA elements in highly active genomic regions, and the significantly higher proportion of their occurrence in risk loci, compared to both L1 and Alu variants. Nevertheless, the integration site preference analysis in this study was limited by the reliance on polymorphic germline insertions of various allele frequencies that may not be representative of the actual integration site preferences of SVAs. As such, future studies characterising the integration sites of experimentally induced *de novo* SVA elements in cultured cells are required to confirm the conclusions drawn by the current study, and better understand the contribution of various genomic features in favouring or restricting the mobilisation of SVA elements. In addition, this study identified a list of RTE variants that are potential candidates of causation at GWAS risk loci. These RTE-TAS associations require future experimental validation studies, to confirm the causal relationship between RTE variants and numerous multifactorial traits in humans. An example of this would be investigating the association between

candidate RTE variants and the expression of nearby genes with functional effects that can be linked to the trait phenotype. Lastly, the inclusion of ethnic cohorts other than European in future GWAS could allow upcoming studies to expand the RTE-TAS associations, which could lead to a better understanding of disease etiology.

References

Abraham, G. *et al.* (2016) 'Genomic prediction of coronary heart disease', *European Heart Journal*, 37(43), pp. 3267–3278. doi: 10.1093/eurheartj/ehw450.

Abrusán, G. and Krambeck, H. J. (2006) 'The distribution of L1 and Alu retroelements in relation to GC content on human sex chromosomes is consistent with the ectopic recombination model', *Journal of Molecular Evolution*, 63(4), pp. 484–492. doi: 10.1007/s00239-005-0275-0.

Achanta, P. *et al.* (2016) 'Somatic retrotransposition is infrequent in glioblastomas', *Mobile DNA*, 7(1), pp. 1–9. doi: 10.1186/s13100-016-0077-5.

Agarwala, V. *et al.* (2013) 'Evaluating empirical bounds on complex disease genetic architecture', *Nature Genetics*. Nature Publishing Group, 45(12), pp. 1418–1427. doi: 10.1038/ng.2804.

Akagi, K., Li, J. and Symer, D. E. (2013) 'How do mammalian transposons induce genetic variation? A conceptual framework: The age, structure, allele frequency, and genome context of transposable elements may define their wide-ranging biological impacts', *BioEssays*, 35(4), pp. 397–407. doi: 10.1002/bies.201200133.

Allen, M. *et al.* (2016) 'Data Descriptor : Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases', pp. 1–10.

Andersson, R. *et al.* (2014) 'Shiori Maeda 5, 6 , Yutaka Negishi 5,6 , Christopher J. Mungall 11', *Nature*, 507(7493), pp. 455–461. doi: 10.1038/nature12787.

Aneichyk, T. *et al.* (2018) 'Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly Article Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly', *Cell*, 172(5), pp. 897–909. doi: 10.1016/j.cell.2018.02.011.

Arokium, H. *et al.* (2014) 'Deep sequencing reveals low incidence of endogenous LINE-1 retrotransposition in human induced pluripotent stem cells', *PLoS ONE*, 9(10), p. e108682. doi: 10.1371/journal.pone.0108682.

Auton, A. (2009) 'VCFtools'. Available at [http://vcftools.sourceforge.net/man_latest.html]

Bailey, J. A. *et al.* (2000) 'Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis', *Proceedings of the National Academy of Sciences of the United States of America*, 97(12), pp. 6634–9. doi: 10.1073/pnas.97.12.6634.

Baillie, J. K. *et al*. (2011) 'Somatic retrotransposition alters the genetic landscape of the human brain', Nature, 479(7374), pp. 534–537. doi: 10.1038/nature10531.

Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. (2009) 'Hypothesis testing, type I and type II errors', *Ind Psychiatry J*, 18(2), pp. 127–131. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996198/.

Barton, N. H., Etheridge, A. M. and Véber, A. (2017) 'The infinitesimal model: Definition, derivation, and implications', *Theoretical Population Biology*. Elsevier Inc., 118, pp. 50–73. doi: 10.1016/j.tpb.2017.06.001.

Batzer, M. A. *et al.* (1996) 'Standardized nomenclature for Alu repeats', *Journal of Molecular Evolution*, 42(1), pp. 3–6. doi: 10.1007/BF00163204.

Beck, C. R. *et al.* (2010) 'LINE-1 Retrotransposition Activity in Human Genomes', *Cell*, 141(7), pp. 1159–1170. doi: 10.1016/j.cell.2010.05.021.LINE-1.

Beck, C. R. *et al.* (2011) 'LINE-1 Elements in Structural Variation and Disease', *Annual Review of Genomics and Human Genetics*, 12(1), pp. 187–215. doi: 10.1146/annurev-genom-082509-141802.

Belle, E. M. S., Webster, M. T., & Eyre-Walker, A. (2005). Why are young and old repetitive elements distributed differently in the human genome? Journal of Molecular Evolution, 60(3), 290–296. https://doi.org/10.1007/s00239-004-0020-0

Bennett, E. A. *et al*. (2004) 'Natural Genetic Variation Caused by Transposable Elements in Humans', 951(October), pp. 933–951. doi: 10.1534/genetics.104.031757.

Bennett, E. A. *et al.* (2008) 'Active Alu retrotransposons in the human genome', *Genome Research*, 18, pp. 1875–1883. doi: 10.1101/gr.081737.108.7.

Bestor, T. H. and Bourc'his, D. (2004) 'Transposon silencing and imprint establishment in mammalian germ cells', *Cold Spring Harbor Symposia on Quantitative Biology*, 69, pp. 381–387. doi: 10.1101/sqb.2004.69.381.

Bire, S. and Rouleux-bonnin, F. (2012) 'Transposable Elements as Tools for Reshaping the Genome: It Is a Huge World After All!'. *Methods in Molecular Biology*, 859. doi: 10.1007/978-1-61779-603-6_1.

Boeke JD, Garfinkel DJ, Styles CA, Fink GR. (1985) 'Ty elements transpose through an RNA intermediate', *Cell*, 40(3), pp. 491–500. doi: 10.1016/0092-8674(85)90197-7.

Boettger LM, Handsaker RE, Zody MC, M. S. (2012) 'Structural haplotypes and recent evolution of the human 17q21.31 region', *Nature Genetics*, 44(8), pp. 881–885. doi: 10.1038/ng.2334.

Boissinot, S., Entezam, a, & Furano, a V. (2001). Selection against deleterious LINE-1-containing loci in the human lineage. Mol. Biol. Evol., 18(6), 926–935. https://doi.org/10.1093/oxfordjournals.molbev.a003893

Boissinot, S. *et al.* (2004) 'The insertional history of an active family of L1 retrotransposons in humans', *Genome Research*, 14(7), pp. 1221–1231. doi: 10.1101/gr.2326704.

Boissinot, S. and Furano, A. (2005) 'The recent evolution of human L1 retrotransposons', *Cytogenetic and Genome Research*, 110, pp. 402–406. doi: 10.1159/000084972.

Bourgeois, Y. and Boissinot, S. (2019) 'On the population dynamics of junk: A review on the population genomics of transposable elements', *Genes*, 10(6). doi: 10.3390/genes10060419.

Bourque, G. *et al.* (2018) 'Ten things you should know about transposable elements', *Genome biology*. Genome Biology, 19(199), pp. 1–12. doi: https://doi.org/10.1186/s13059-018-1577-z.

Bowen NJ. Jordan IK. (2002) 'Transposable Elements and the Evolution of Eukaryotic Complexity', *Curr. Issues Mol. Biol*, 4(3), pp. 65–76. doi: 10.1016/0168-9525(89)90039-5.

Bragg, D. C., Mangkalaphiban, K., Vaine, C. A., Kulkarni, N. J., Shin, D., Yadav, R., Ozelius, L. J. (2017). Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. Proceedings of the National Academy of Sciences, 114(51), E11020–E11028. https://doi.org/10.1073/pnas.1712526114

Brandler, W. M. *et al.* (2016) 'Frequency and Complexity of *De Novo* Structural Mutation in Autism', *The American Journal of Human Genetics*. The American Society of Human Genetics, 98(4), pp. 667–679. doi: 10.1016/j.ajhg.2016.02.018.

Brouha, B. et al. (2003) 'Hot L1s account for the bulk of retrotransposition in the human population', Proceedings of the National Academy of Sciences, 100(9), pp. 5280–5285. doi: 10.1073/pnas.0831042100.

Buniello, A. *et al.* (2019) 'The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019', *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D1005–D1012. doi: 10.1093/nar/gky1120.

Burns, K. H. (2017) 'Transposable elements in cancer', *Nature Reviews Cancer*. Nature Publishing Group, 17(7), pp. 415–424. doi: 10.1038/nrc.2017.35.

Callinan, P. A. *et al.* (2005) 'Alu retrotransposition-mediated deletion', *Journal of Molecular Biology*, 348(4), pp. 791–800. doi: 10.1016/j.jmb.2005.02.043.

Cardelli, M., Marchegiani, F. and Provinciali, M. (2012) 'Alu insertion profiling: Array-based methods to detect Alu insertions in the human genome', *Genomics*. Elsevier Inc., 99(6), pp. 340–346. doi: 10.1016/j.ygeno.2012.03.005.

Carreira, P. E. *et al.* (2016) 'Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme', *Mobile DNA*, 7(1), p. 21. doi: 10.1186/s13100-016-0076-6.

Carter, A. B. *et al*. (2004) 'Genome-wide analysis of the human Alu Yb-lineage', Human Genomics, 1(3), pp. 167–178.

Carvalho, C. M. B. and Lupski, J. R. (2016) 'Mechanisms underlying structural variant formation in genomic disorders.', *Nature reviews. Genetics*. Nature Publishing Group, 17(4), pp. 224–38. doi: 10.1038/nrg.2015.25.

Chen, C., Ara, T. and Gautheret, D. (2009) 'Using Alu elements as polyadenylation sites: A case of retroposon exaptation', *Molecular Biology and Evolution*, 26(2), pp. 327–334. doi: 10.1093/molbev/msn249.

Chen, H., Hey, J. and Chen, K. (2015) 'Inferring very recent population growth rate from population-scale sequencing data: Using a large-sample coalescent estimator', *Molecular Biology and Evolution*, 32(11), pp. 2996–3011. doi: 10.1093/molbev/msv158.

Chen, D. *et al*. (2020) 'Human L1 transposition dynamics unraveled with functional data analysis', Molecular Biology and Evolution, 37(12), pp. 3576–3600. doi: 10.1093/molbev/msaa194.

Chénais, B. (2016) 'Transposable Elements in Cancer and Other Human Diseases Transposable Elements in Cancer and Other Human Diseases', (April). doi: 10.2174/1568009615666150317122506.

Chiang, C. *et al.* (2017) 'The impact of structural variation on human gene expression', *Nature Publishing Group*. Nature Publishing Group, 49(5), pp. 692–699. doi: 10.1038/ng.3834.

Chuong, E. B., Elde, N. C. and Feschotte, C. (2017) 'Regulatory activities of transposable elements: from conflicts to benefits', *Nat Rev Genet*, 18(2), pp. 71–86. doi: 10.1038/nrg.2016.139.

Conley, A. B. and Jordan, I. K. (2012) 'Cell type-specific termination of transcription by transposable element sequences', Mobile DNA, 3(1), pp. 1–13. doi: 10.1186/1759-8753-3-15.

Cordaux, R. *et al.* (2006) 'Recently integrated Alu retrotransposons are essentially neutral residents of the human genome', *Gene*, 373(1–2), pp. 138–144. doi: 10.1016/j.gene.2006.01.020.

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. Nature Reviews Genetics, 10(10), 691–703. https://doi.org/10.1038/nrg2640

Cost, G. J. *et al.* (2002) 'Human L1 element target-primed reverse transcription in vitro', *EMBO Journal*, 21(21), pp. 5899–5910. doi: 10.1093/emboj/cdf592.

Cost, G. J., & Boeke, J. D. (1998). Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry, 37(51), 18081–18093. https://doi.org/10.1021/bi981858s

Costantini, M., Auletta, F., & Bernardi, G. (2012). The Distributions of ‘“ New ”’ and ‘“ Old ”’ Alu Sequences in the Human Genome : The Solution of a ‘“ Mystery ”’ Research article. 29(2001), 421–427. https://doi.org/10.1093/molbev/msr242

Coufal, N. G. *et al.* (2011) ‘Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(51), pp. 20382–20387. doi: 10.1073/pnas.1100273108.

Cuccaro, D. *et al.* (2016) ‘Copy number variants in Alzheimer’s disease’, *Journal of Alzheimer’s Disease*, 55(1), pp. 37–52. doi: 10.3233/JAD-160469.

David, M., Mustafa, H. and Brudno, M. (2013) ‘Detecting Alu insertions from high-throughput sequencing data’, *Nucleic Acids Research*, 41(17), pp. 1–13. doi: 10.1093/nar/gkt612.

De Almeida, R. A. *et al*. (2016) ‘Non-coding RNAs and disease: The classical ncRNAs make a comeback’, Biochemical Society Transactions, 44(4), pp. 1073–1078. doi: 10.1042/bst20160089.

De Fazio, S. *et al.* (2011) ‘The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements’, *Nature*, 480(7376), pp. 259–263. doi: 10.1038/nature10547.

Dewannieux, M., Esnault, C. and Heidmann, T. (2003) ‘LINE-mediated retrotransposition of marked Alu sequences’, *Nature Genetics*, 35(1), pp. 41–48. doi: 10.1038/ng1223.

Dickson, S. P. *et al.* (2010) ‘Rare Variants Create Synthetic Genome-Wide Associations’, *PLoS Biology*, 8(1). doi: 10.1371/journal.pbio.1000294.

Diehl, A. G., Ouyang, N. and Boyle, A. P. (2020) ‘Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes’, *Nature Communications*. Springer US, 11(1), pp. 1–18. doi: 10.1038/s41467-020-15520-5.

Dolgin, E. S., & Charlesworth, B. (2008). The effects of recombination rate on the distribution and abundance of transposable elements. Genetics, 178(4), 2169–2177. https://doi.org/10.1534/genetics.107.082743

Dombroski, B. A. *et al*. (1991) ‘Isolation of an active human transposable element’, Science, 254(5039), pp. 1805–1808. doi: 10.1126/science.1662412.

Doolittle WF, Sapienza C. (1980) ‘Selfish genes, the phenotype paradigm and genome evolution’, *Nature*, 284(5757), pp. 601–3. doi: 10.1038/284601a0.

Doyle, G. A. *et al.* (2017) 'Analysis of LINE-1 elements in DNA from postmortem brains of individuals with schizophrenia', *Neuropsychopharmacology*. Nature Publishing Group, 42(13), pp. 2602–2611. doi: 10.1038/npp.2017.115.

Ecker, E. D. and Skelly, A. C. (2010) 'Conducting a winning literature search', *Evid Based Spine Care J*, 1(1), pp. 9–14. doi: 10.1055/s-0028-1100887.

Elbarbary, R. A., Lucas, B. A. and Maquat, L. E. (2016) 'Retrotransposons as regulators of gene expression', *Science*, 351(6274). doi: 10.1126/science.aac7247.

Erwin, J. A., Marchetto, M. C., & Gage, F. H. (2014). Mobile DNA elements in the generation of diversity and complexity in the brain. Nature Reviews. Neuroscience, 15(8), 497–506. https://doi.org/10.1038/nrn3730

Erwin, J. A. *et al.* (2016) 'L1-associated genomic regions are deleted in somatic cells of the healthy human brain', *Nature Neuroscience*, 19(12), pp. 1583–1591. doi: 10.1038/nn.4388.

Escaramís, G., Docampo, E. and Rabionet, R. (2015) 'A decade of structural variants: Description, history and methods to detect structural variation', *Briefings in Functional Genomics*, 14(5), pp. 305–314. doi: 10.1093/bfgp/elv014.

Evans, D. M. and Cardon, L. R. (2005) 'A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations', *American Journal of Human Genetics*, 76(4), pp. 681–687. doi: 10.1086/429274.

Evrony, G. D. *et al.* (2012) 'Single-neuron sequencing analysis of l1 retrotransposition and somatic mutation in the human brain', *Cell*. Elsevier, 151(3), pp. 483–496. doi: 10.1016/j.cell.2012.09.035.

Evrony, G. D. *et al.* (2015) 'Cell Lineage Analysis in Human Brain Using Endogenous Retroelements', *Neuron*, 85(1), pp. 49–59. doi: 10.1016/j.neuron.2014.12.028.

Evrony, G. D. *et al*. (2016) 'Resolving rates of mutation in the brain using single-neuron genomics', eLife, 5(FEBRUARY2016), pp. 1–32. doi: 10.7554/eLife.12966.

Ewing, A. D. and Kazazian, H. H. (2010) 'High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes', *Genome Research*, 20(215), pp. 1262–1270. doi: 10.1101/gr.106419.110.

Ewing, A. D. and Kazazian, H. H. (2011) 'Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans', *Genome Research*, 21(6), pp. 985–990. doi: 10.1101/gr.114777.110.

Ewing, A. D. (2015) 'Transposable element detection from whole genome sequence data', *Mobile DNA*. Mobile DNA, 6(1), p. 24. doi: 10.1186/s13100-015-0055-3.

Faulkner, G. J. and Billon, V. (2018) ‘L1 retrotransposition in the soma: A field jumping ahead’, *Mobile DNA*. Mobile DNA, 9(1), pp. 1–18. doi: 10.1186/s13100-018-0128-1.

Ferrari, R. *et al.* (2017) ‘Genetic architecture of sporadic frontotemporal dementia and overlap with Alzheimer’s and Parkinson’s diseases’, *Neurol Neurosurg Psychiatry*, 88(2), pp. 152–164. doi: 10.1136/jnnp-2016-314411.Genetic.

Feschotte, C. (2008) ‘The contribution of transposable elements ot the evolution of regulatory networks’, *Nature Reviews Genetics*, 9(5), pp. 397–405. doi: 10.1038/nrg2337.

Feuk, L. *et al.* (2006) ‘Structural variants: changing the landscape of chromosomes and design of disease studies.’, *Human molecular genetics*, 15(1), pp. 57–66. doi: 10.1093/hmg/ddl057.

Feusier, J. *et al*. (2017) ‘Discovery of rare, diagnostic AluYb8/9 elements in diverse human populations’, Mobile DNA. Mobile DNA, 8(1), pp. 21–23. doi: 10.1186/s13100-017-0093-0.

Feusier, J. *et al.* (2019) ‘Pedigree-based estimation of human mobile element retrotransposition rates’, *Genome Research*, 29(10), pp. 1567–1577. doi: 10.1101/gr.247965.118.

Fishilevich, S. *et al.* (2017) ‘GeneHancer: genome-wide integration of enhancers and target genes in GeneCards’, *Database : the journal of biological databases and curation*, 2017, pp. 1–17. doi: 10.1093/database/bax028.

Flasch, D. A. *et al*. (2019) ‘Genome-wide *de novo* L1 Retrotransposition Connects Endonuclease Activity with Replication’, Cell, 177(4), pp. 837-851.e28. doi: 10.1016/j.cell.2019.02.050.

Frank J. Massey, J. (1951) ‘The Kolmogorov-Smirnov Test for Goodness of Fit’, *Journal of the American Statistical Association*, 46(253), pp. 68–78. doi: https://doi.org/10.2307/2280095.

Frayling, T. (2014) ‘Genome-wide association studies: the good, the bad and the ugly’, 14(4), pp. 428–431. Available at: www.ebi.ac.uk/fgpt/gwas.

Fuhrman, S. A. *et al.* (1981) ‘Analysis of transcription of the human alu family ubiquitous repeating element by eukaryotic RNA polymerase III’, *Nucleic Acids Research*, 9(23), pp. 6439–6456. doi: 10.1093/nar/9.23.6439.

Fullerton, S. M., Carvalho, A. B. and Clark, A. G. (2001) ‘Local rates of recombination are positively correlated with GC content in the human genome [4]’, Molecular Biology and Evolution, 18(6), pp. 1139–1142. doi: 10.1093/oxfordjournals.molbev.a003886.

Gao, F. and Keinan, A. (2016) ‘Explosive genetic evidence for explosive human population growth’, *Current Opinion in Genetics and Development*. Elsevier Ltd, 41, pp. 130–139. doi: 10.1016/j.gde.2016.09.002.

Garcia-Perez, J. L. *et al.* (2010) 'Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells', *Nature*. Nature Publishing Group, 466(7307), pp. 769–773. doi: 10.1038/nature09209.

Garcia-Perez, J. L., Widmann, T. J. and Adams, I. R. (2016) 'The impact of transposable elements on mammalian development', *Development*, 143(22), pp. 4101–4114. doi: 10.1242/dev.132639.

Gardner, E. J. *et al.* (2017) 'The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology', *Genome Research*, 27, pp. 1916–1929. doi: 10.1101/gr.218032.116.Freely.

Gardner EJ, Prigmore E, Gallone G, Danecek P, Samocha KE, Handsaker J, Gerety SS, Ironfield H, Short PJ, Sifrim A, Singh T, Chandler KE, Clement E, Lachlan KL, Prescott K, Rosser E, FitzPatrick DR, Firth HV, H. M. (2019) 'Contribution of Retrotransposition to Developmental Disorders', Nature Communications, 10(1). doi: 10.1038/s41467-019-12520-y.

Gasior, S. L. *et al.* (2006) 'The human LINE-1 retrotransposon creates DNA double-strand breaks', *Journal of Molecular Biology*, 357(5), pp. 1383–1393. doi: 10.1016/j.jmb.2006.01.089.

Gasior, S. L. *et al.* (2007) 'Characterization of pre-insertion loci of *de novo* L1 insertions', *Gene*, 390(1–2), pp. 190–198. doi: 10.1016/j.gene.2006.08.024.

Gazave, E. *et al.* (2013) 'Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect', *Genetics*, 195(3), pp. 969–978. doi: 10.1534/genetics.113.153973.

Gianfrancesco, O., Bubb, V. J., & Quinn, J. P. (2017). SVA retrotransposons as potential modulators of neuropeptide gene expression. Neuropeptides, 64, 3–7. https://doi.org/10.1016/j.npep.2016.09.006

Gianfrancesco, O., Geary, B., Savage, A. L., Billingsley, K. J., Bubb, V. J., & Quinn, J. P. (2019). The role of SINE-VNTR-Alu (SVA) retrotransposons in shaping the human genome. International Journal of Molecular Sciences, 20(23), 1–17. https://doi.org/10.3390/ijms20235977

Gibson, G. (2012) 'Rare and common variants: Twenty arguments', *Nature Reviews Genetics*. Nature Publishing Group, 13(2), pp. 135–145. doi: 10.1038/nrg3118.

Goerner-potvin, P. and Bourque, G. (2018) 'Computational tools to unmask transposable elements', *Nature Reviews Genetics*. Springer US, 19(November), pp. 688–704. doi: 10.1038/s41576-018-0050-x.

Goodier, J. L. (2016) 'Restricting retrotransposons: A review', *Mobile DNA*. Mobile DNA, 7(1). doi: 10.1186/s13100-016-0070-z.

Graham, T. and Boissinot, S. (2006) 'The genomic distribution of L1 elements: The role of insertion bias and natural selection', Journal of Biomedicine and Biotechnology, 2006, pp. 1–5. doi: 10.1155/JBB/2006/75327.

Gravel, S. *et al.* (2011) 'Demographic history and rare allele sharing among human populations', *Proceedings of the National Academy of Sciences of the United States of America*, 108(29), pp. 11983–11988. doi: 10.1073/pnas.1019276108.

Gregory, T. R. (2005) 'Synergy between sequence and size in Large-scale genomics', Nature Reviews Genetics. Nature Publishing Group, 6, p. 699. Available at: http://dx.doi.org/10.1038/nrg1674.

Grover, D. *et al*. (2004) 'Alu repeat analysis in the complete human genome: Trends and variations with respect to genomic composition', Bioinformatics, 20(6), pp. 813–817. doi: 10.1093/bioinformatics/bth005.

Gu, Z. *et al*. (2000) 'Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence', Gene, 259(1–2), pp. 81–88. doi: 10.1016/S0378-1119(00)00434-0.

Guffanti, G. *et al.* (2014) 'Transposable elements and psychiatric disorders', *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 165B(3). doi: 10.1002/ajmg.b.32225.

Guffanti, G. *et al.* (2016) 'LINE1 insertions as a genomic risk factor for schizophrenia: Preliminary evidence from an affected family', *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 171(4), pp. 534–545. doi: 10.1002/ajmg.b.32437.

Ha, H., Loh, J. W. and Xing, J. (2016) 'Identification of polymorphic SVA retrotransposons using a mobile element scanning method for SVA (ME-Scan-SVA)', *Mobile DNA*. Mobile DNA, 7(1), pp. 15–17. doi: 10.1186/s13100-016-0072-x.

Hackenberg, M. *et al*. (2005) 'The biased distribution of Alus in human isochores might be driven by recombination', Journal of Molecular Evolution, 60(3), pp. 365–377. doi: 10.1007/s00239-004-0197-2.

Han, K. *et al.* (2005) 'Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages', *Nucleic Acids Research*, 33(13), pp. 4040–4052. doi: 10.1093/nar/gki718.

Hancks, D. C. *et al.* (2011) 'Retrotransposition of marked SVA elements by human L1s in cultured cells', *Human Molecular Genetics*, 20(17), pp. 3386–3400. doi: 10.1093/hmg/ddr245.

Hancks, D. C., Kazazian, H. H. and Jr. (2016) 'Roles for retrotransposon insertions in human disease.', *Mobile DNA*. Mobile DNA, 7, p. 9. doi: 10.1186/s13100-016-0065-9.

Hehir-Kwa, J. Y. *et al.* (2016) 'A high-quality human reference panel reveals the complexity and distribution of genomic structural variants', *Nature Communications*, 7, pp. 1–10. doi: 10.1038/ncomms12989.

Helman, E. *et al.* (2014) 'Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing', *Genome Research*, 24(7), pp. 1053–1063. doi: 10.1101/gr.163659.113.

Henry L. Levin, John V. Moran. (2014). Dynamic interactions between transposable elements and their hosts. (August 2011). https://doi.org/10.1038/nrg3030

Higashino, S. *et al.* (2014) 'Polymorphic L1 retrotransposons are frequently in strong linkage disequilibrium with neighboring SNPs', *Gene*. Elsevier B.V., 541(1), pp. 55–59. doi: 10.1016/j.gene.2014.03.008.

Hindorff, L. A. *et al.* (2009) 'Potential etiologic and functional implications of genome-wide association loci for human diseases and traits', *PNAS*, 106(23), pp. 9362–7. doi: 10.1073/pnas.0903103106.

Hormozdiari, F. *et al*. (2011) 'Alu repeat discovery and characterization within human genomes', Genome Research, 21(6), pp. 840–849. doi: 10.1101/gr.115956.110.

Hormozdiari, Farhad *et al.* (2019) 'Functional disease architectures reveal unique biological role of transposable elements', *Nature Communications*. Springer US, 10(1). doi: 10.1038/s41467-019-11957-5.

Hu, Z., Wang, Z. and Xu, S. (2012) 'An infinitesimal model for quantitative trait genomic value prediction', *PLoS ONE*, 7(7), pp. 1–14. doi: 10.1371/journal.pone.0041336.

Huang, C. R. L. et al. (2010) 'Mobile interspersed repeats are major structural variants in the human genome', Cell. Elsevier Ltd, 141(7), pp. 1171–1182. doi: 10.1016/j.cell.2010.05.026.

Iskow, R. C. *et al.* (2010) 'Natural mutagenesis of human genomes by endogenous retrotransposons', *Cell*. Elsevier Ltd, 141(7), pp. 1253–1261. doi: 10.1016/j.cell.2010.05.020.

Jacobs, F. M. J., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A. D., Katzman, S., Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature, 516(7530), 242–245. https://doi.org/10.1038/nature13760

Jacques, P.-E., Jeyakani, J. and Bourque, G. (2013) 'The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements', *PLoS Genetics*, 9(5). doi: 10.1371/journal.pgen.1003504.

Jiang, T. *et al.* (2019) 'RMETL: Sensitive mobile element insertion detection with long read realignment', *Bioinformatics*, 35(18), pp. 3484–3486. doi: 10.1093/bioinformatics/btz106.

Jeggo, P. A. and Löbrich, M. (2007) 'DNA double-strand breaks: Their cellular and clinical impact?', *Oncogene*, 26(56), pp. 7717–7719. doi: 10.1038/sj.onc.1210868.

Johnson, N. and Joseph, L. (2012) ‘The genetics of sex chromosomes: evolution and implications for hybrid incompatibility’, *Bone*, 1256(1), pp. E1-22. doi: 10.1111/j.1749-6632.2012.06748.x.

Johnson, R. and Guigó, R. (2014) ‘The RIDL hypothesis: Transposable elements as functional domains of long noncoding RNAs’, *Rna*, 20(7), pp. 959–976. doi: 10.1261/rna.044560.114.

Jordan, I. K. *et al.* (2003) ‘Origin of a substantial fraction of human regulatory sequences from transposable elements’, *Trends in Genetics*, 19(2), pp. 68–72. doi: 10.1016/S0168-9525(02)00006-9.

Jung, I. (2017) ‘Some facts that you might be unaware of about the P-value’, Archives of Plastic Surgery, 44(2), pp. 93–94. doi: 10.5999/aps.2017.44.2.93.

Jurka, J. and Smith, T. (1988) ‘A fundamental division in the Alu family of repeated sequences’, *Proceedings of the National Academy of Sciences of the United States of America*, 85(13), pp. 4775–4778. doi: 10.1073/pnas.85.13.4775.

Jurka, J. *et al*. (2004) ‘Duplication, coclustering, and selection of human Alu retrotransposons’, Proceedings of the National Academy of Sciences of the United States of America, 101(5), pp. 1268–1272. doi: 10.1073/pnas.0308084100.

Karolchik, D. (2004) ‘The UCSC Table Browser data retrieval tool’, *Nucleic Acids Research*, 32(90001), pp. 493D – 496. doi: 10.1093/nar/gkh103.

Kawamura, Y. *et al.* (2011) ‘A genome-wide CNV association study on panic disorder in a Japanese population’, *Journal of Human Genetics*. Nature Publishing Group, 56(12), pp. 852–856. doi: 10.1038/jhg.2011.117.

Kazazian HH Jr, Wong C, Youssoufian H, Scott AF, Phillips DG, A. S. (1988) ‘Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man’, *Nature*, 332(6160), pp. 164–6.

Kazazian, H. H. (2004) ‘Mobile elements: drivers of genome evolution.’ *Science*, 303(5664), pp. 1626–32. doi: 10.1126/science.1089670.

Keinan, A. and Clark, A. G. (2012) ‘Recent explosive human population growth has resulted in an excess of rare genetic variants.’ *Science*, 336(6082), pp. 740–3. doi: 10.1126/science.1217283.

Kelley, D. and Rinn, J. (2012) ‘Transposable elements reveal a stem cell-specific class of long noncoding RNAs’, *Genome biology*, 13(11), p. R107. doi: 10.1186/gb-2012-13-11-r107.

Kent, T. V., Uzunović, J. and Wright, S. I. (2017) ‘Coevolution between transposable elements and recombination’, Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1736). doi: 10.1098/rstb.2016.0458.

Kidwell MG, Lisch DR (2001) ‘Perspective: transposable elements, parasitic dna, and genome evolution’, *Evolution*, 55(1), pp. 1–24. doi: 10.1111/j.0014-3820.2001.tb01268.x.

Kleckner, N. (1981) ‘Transposable elements in prokaryotes.’ *Annu Rev Genet*, 15, pp. 341–404. doi: 10.1146/annurev.ge.15.120181.002013.

Kloosterman, W. P. *et al.* (2015) ‘Characteristics of *de novo* structural changes in the human genome’, *Genome Research*, 25(6), pp. 792–801. doi: 10.1101/gr.185041.114.

Kong, A. *et al.* (2010) ‘Fine-scale recombination rate differences between sexes, populations and individuals’, *Nature*. Nature Publishing Group, 467(7319), pp. 1099–1103. doi: nature09525 [pii]\r10.1038/nature09525.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Stefansson, K. (2002). A high-resolution recombination map of the human genome. Nature Genetics, 31(3), 241–247. https://doi.org/10.1038/ng917

Konkel, M. K. *et al.* (2015) ‘Sequence analysis and characterization of active human alu subfamilies based on the 1000 genomes pilot project’, *Genome Biology and Evolution*, 7(9), pp. 2608–2622. doi: 10.1093/gbe/evv167.

Koolen, D. A. *et al.* (2016) ‘The Koolen-de Vries syndrome: A phenotypic comparison of patients with a 17q21.31 microdeletion versus a KANSL1 sequence variant’, *European Journal of Human Genetics*, 24(5), pp. 652–659. doi: 10.1038/ejhg.2015.178.

Kornienko, A. E. et al. (2013) ‘Gene regulation by the act of long non-coding’, BMC Biology, 11(59), pp. 1–14.

Kuhn, A. *et al.* (2014) ‘Linkage disequilibrium and signatures of positive selection around LINE-1 retrotransposons in the human genome’, *Proceedings of the National Academy of Sciences of the United States of America*, 111(22), pp. 8131–6. doi: 10.1073/pnas.1401532111.

Kurnosov, A. A., Ustyugova, S. V and Nazarov, V. I. (2015) ‘The Evidence for Increased L1 Activity in the Site of Human Adult Brain Neurogenesis’, *PLoS ONE*, 10(2), p. e0117854. doi: 10.1371/journal.pone.0117854.

Kvikstad, E. M. and Makova, K. D. (2010) ‘The (r) evolution of SINE versus LINE distributions in primate genomes : Sex chromosomes are important’, *Genome Research*, 20, pp. 600–613. doi: 10.1101/gr.099044.109.1.

Kwon, Y., Choi, Y., Eo, J., Noh, Y., Gim, J., Jung, Y., Kim, H. (2013). Structure and Expression Analyses of SVA Elements in Relation to Functional Genes. 11(3), 142–148.

Lacaria, M., Gu, W. and Lupski, J. R. (2013) ‘A functional role for structural variation in metabolism’, *Adipocyte*, 2(1), pp. 55–57. doi: 10.4161/adip.22031.

Lall, A. (2015) 'Data streaming algorithms for the Kolmogorov-Smirnov test', *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, pp. 95–104. doi: 10.1109/BigData.2015.7363746.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860–921. https://doi.org/10.1038/35057062

Lee, J. *et al.* (2012) 'Human Genomic Deletions Generated by SVA-Associated Events', *Comparative and Functional Genomics*, 2012. doi: 10.1155/2012/807270.

Lev-Maor, G. *et al.* (2008) 'Intronic Alus influence alternative splicing', *PLoS Genetics*, 4(9), pp. 1–12. doi: 10.1371/journal.pgen.1000204.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

Li, R. *et al*. (2009) 'SOAP2: An improved ultrafast tool for short read alignment', Bioinformatics, 25(15), pp. 1966–1967. doi: 10.1093/bioinformatics/btp336.

Li, Y. *et al.* (2014) 'An Epigenetic Signature in Peripheral Blood Associated with the Haplotype on 17q21.31, a Risk Factor for Neurodegenerative Tauopathy', *PLoS Genetics*, 10(3). doi: 10.1371/journal.pgen.1004211.

Liu, Z. *et al.* (2013) 'Negative life events and corticotropin-releasing-hormone receptor1 gene in recurrent major depressive disorder', *Scientific Reports*, 3, pp. 1–5. doi: 10.1038/srep01548.

Loewe, L. (2008) Negative selection. Nature Education 1(1):59. Available at: [https://www.nature.com/scitable/topicpage/negative-selection-1136/] (Accessed: 22 June 2019).

Lowe, C. B. and Haussler, D. (2012) '29 Mammalian Genomes Reveal Novel Exaptations of Mobile Elements for Likely Regulatory Functions in the Human Genome', *PLoS ONE*, 7(8). doi: 10.1371/journal.pone.0043128.

Lu, H., Giordano, F. and Ning, Z. (2016) 'Oxford Nanopore MinION Sequencing and Genome Assembly', Genomics, Proteomics and Bioinformatics. Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China, 14(5), pp. 265–279. doi: 10.1016/j.gpb.2016.05.004.

Lutz, S. M. *et al.* (2003) 'Allelic Heterogeneity in LINE-1 Retrotransposition Activity', *The American Journal of Human Genetics*, 73(6), pp. 1431–1437. doi: 10.1086/379744.

Lynch, V. J. *et al.* (2011) 'Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals', *Nature Genetics*. Nature Publishing Group, 43(11), pp. 1154–1159. doi: 10.1038/ng.917.

Lynch, V. J. *et al.* (2015) 'Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy', *Cell Reports*, 10(4), pp. 551–561. doi: 10.1016/j.celrep.2014.12.052.

Maher, B. (2008) 'Personal genomes: The case of the missing heritability', *Nature*, 456(7218), pp. 18–21. doi: 10.1038/456018a.

Malhotra, D. and Sebat, J. (2012) 'CNVs: Harbingers of a rare variant revolution in psychiatric genetics', *Cell*. Elsevier Inc., 148(6), pp. 1223–1241. doi: 10.1016/j.cell.2012.02.039.

Mamedov, I. Z. *et al*. (2005) 'Whole-genome experimental identification of insertion/deletion polymorphisms of interspersed repeats by a new general approach.', Nucleic acids research, 33(2). doi: 10.1093/nar/gni018.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, *et al.* (2009) 'Finding the missing heritability of complex diseases', *Nature*, 461(7265), pp. 747–753. doi: 10.1038/nature08494.

McClintock, B. (1956) 'Intranuclear systems controlling gene action and mutation.' *Brookhaven Symp Biol*, 8, pp. 58–74.

McLeod, S. A. (2019, May 17). Z-score: definition, calculation and interpretation. Simply Psychology. https://www.simplypsychology.org/z-score.html

Medina-Gomez, C. *et al.* (2015) 'Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study', *European Journal of Epidemiology*, 30(4), pp. 317–330. doi: 10.1007/s10654-015-9998-4.

Medstrand, P., Lagemaat, L. N. Van De and Mager, D. L. (2002) 'Retroelement Distributions in the Human Genome : Variations Associated With Age and Proximity to Genes Distributions of Retroelements in Different', *Cold Spring Harbor Press*, 12, pp. 1483–1495. doi: 10.1101/gr.388902.5.

Mills, R. E. *et al*. (2007) 'Which transposable elements are active in the human genome?', Trends in Genetics, 23(4), pp. 183–191. doi: 10.1016/j.tig.2007.02.006.

Mir, A. A., Philippe, C. and Cristofari, G. (2014) 'euL1db: The European database of L1HS retrotransposon insertions in humans', Nucleic Acids Research. doi: 10.1093/nar/gku1043.

Moran, J. V. *et al.* (1996) 'High frequency retrotransposition in cultured mammalian cells', *Cell*, 87(5), pp. 917–927. doi: 10.1016/S0092-8674(00)81998-4.

Mugal, C. F., Weber, C. C., & Ellegren, H. (2015). GC-biased gene conversion links the recombination landscape and demography to genomic base composition: GC-biased gene conversion drives genomic base composition across a wide range of species. BioEssays, 37(12), 1317–1326. https://doi.org/10.1002/bies.201500058

Munoz-Lopez, M. and Garcia-Perez, J. (2010) ‘DNA Transposons: Nature and Applications in Genomics’, *Current Genomics*, 11(2), pp. 115–128. doi: 10.2174/138920210790886871.

Myers, J. S. *et al.* (2002) ‘A Comprehensive Analysis of Recently Integrated Human Ta L1 Elements’, *The American Journal of Human Genetics*, 71(2), pp. 312–326. doi: 10.1086/341718.

Myers, S. *et al*. (2008) ‘A common sequence motif associated with recombination hot spots and genome instability in humans’, Nature Genetics, 40(9), pp. 1124–1129. doi: 10.1038/ng.213.

Nazaryan-Petersen, L. *et al.* (2016) ‘Germline Chromothripsis Driven by L1-Mediated Retrotransposition and Alu/Alu Homologous Recombination’, *Human Mutation*, 37(4), pp. 385–395. doi: 10.1002/humu.22953.

Nelson, M. R. *et al.* (2004) ‘Large-scale validation of single nucleotide polymorphisms in gene regions’, *Genome Research*, 14(8), pp. 1664–1668. doi: 10.1101/gr.2421604.

Nelson, M. G., Linheiro, R. S. and Bergman, C. M. (2017) ‘McClintock: An integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data’, *G3: Genes, Genomes, Genetics*, 7(8), pp. 2763–2778. doi: 10.1534/g3.117.043893.

Nigumann, P. *et al.* (2002) ‘Many human genes are transcribed from the antisense promoter of L1 retrotransposon’, *Genomics*, 79(5), pp. 628–634. doi: 10.1006/geno.2002.6758.

Ngamphiw, C., Tongsima, S. and Mutirangura, A. (2014) ‘Roles of intragenic and intergenic L1s in mouse and human’, PLoS ONE, 9(11), pp. 1–8. doi: 10.1371/journal.pone.0113434.

Nguyen, T. H. M. *et al*. (2018) ‘L1 Retrotransposon Heterogeneity in Ovarian Tumor Cell Evolution’, Cell Reports. ElsevierCompany., 23(13), pp. 3730–3740. doi: 10.1016/j.celrep.2018.05.090.

North, B. V., Curtis, D., & Sham, P. C. (2002). A note on the calculation of empirical P values from Monte Carlo procedures. *American journal of human genetics*, 71(2), 439–441. https://doi.org/10.1086/341527.

O’Leary, N. A. *et al*. (2016) ‘Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation’, Nucleic Acids Research, 44(D1), pp. D733–D745. doi: 10.1093/nar/gkv1189.

Orgel LE. Crick FH. (1980) ‘Selfish DNA: the ultimate parasite’, *Nature*, 284(5757), pp. 604–7. doi: 10.1038/284604a0.

Ostertag, E. M. *et al*. (2003) ‘SVA Elements Are Nonautonomous Retrotransposons that Cause Disease in Humans’, The American Journal of Human Genetics, 73(6), pp. 1444–1451. doi: 10.1086/380207.

Otieno, A. C. *et al*. (2004) ‘Analysis of the human Alu Ya-lineage’, Journal of Molecular Biology, 342(1), pp. 109–118. doi: 10.1016/j.jmb.2004.07.016.

Ovchinnikov, I. *et al.* (2001) ‘Genomic Characterization of Recent Human LINE-1 Insertions : Evidence Supporting Random Insertion Genomic Characterization of Recent Human LINE-1 Insertions : Evidence Supporting Random Insertion’, *Genome Research*, pp. 2050–2058. doi: 10.1101/gr.194701.

Paterson, A. L. *et al.* (2015) ‘Mobile element insertions are frequent in oesophageal adenocarcinomas and can mislead paired-end sequencing analysis’, *BMC Genomics*. BMC Genomics, 16(1), p. 473. doi: 10.1186/s12864-015-1685-z.

Patil, V. S., Zhou, R. and Rana, T. M. (2014) ‘Gene regulation by noncoding RNAs Veena’, Crit Rev Biochem Mol Biol, 25(3), pp. 289–313. doi: 10.3109/10409238.2013.844092.Gene.

Patrushev, L. I. and Kovalenko, T. F. (2014) ‘Functions of noncoding sequences in mammalian genomes’, Biochemistry (Moscow), 79(13), pp. 1442–1469. doi: 10.1134/S0006297914130021.

Pavlíček, A. *et al*. (2001) ‘Similar integration but different stability of Alus and LINEs in the human genome’, Gene, 276(1–2), pp. 39–45. doi: 10.1016/S0378-1119(01)00645-X.

Payer, L. M. *et al.* (2017) ‘Structural variants caused by Alu insertions are associated with risks for many human diseases’, *Proceedings of the National Academy of Sciences*, 114(20), pp. E3984–E3992. doi: 10.1073/pnas.1704117114.

Payer, L. M. *et al.* (2019) ‘Alu insertion variants alter mRNA splicing’, *Nucleic Acids Research*. Oxford University Press, 47(1), pp. 1–11. doi: 10.1093/nar/gky1086.

Perepelitsa-Belancio, V. and Deininger, P. (2003) ‘RNA truncation by premature polyadenylation attenuates human mobile element activity’, *Nature Genetics*, 35(4), pp. 363–366. doi: 10.1038/ng1269.

Pers, T. H., Timshel, P. and Hirschhorn, J. N. (2015) ‘SNPsnap: A Web-based tool for identification and annotation of matched SNPs’, *Bioinformatics*, 31(3), pp. 418–420. doi: 10.1093/bioinformatics/btu655.

Platt, R. N., Vandewege, M. W. and Ray, D. A. (2018) ‘Mammalian transposable elements and their impacts on genome evolution’, *Chromosome Research*. Chromosome Research, 26(1–2), pp. 25–43. doi: 10.1007/s10577-017-9570-z.

Pontis, J., Planet, E., Offner, S., Turelli, P., Duc, J., Coudray, A., Trono, D. (2019). Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. Cell Stem Cell, 24(5), 724-735.e5. https://doi.org/10.1016/j.stem.2019.03.012

Price, A. L., Eskin, E. and Pevzner, P. A. (2004) 'Whole-genome analysis of Alu repeat elements reveals complex evolutionary history', Genome Research, 14(11), pp. 2245–2252. doi: 10.1101/gr.2693004.

Purcell, S. *et al.* (2007) 'PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses', *The American Journal of Human Genetics*, 81(September), pp. 559–575. doi: 10.1086/519795.

Qin, S. *et al.* (2015) 'The role of transposable elements in the origin and evolution of microRNAs in human', *PLoS ONE*, 10(6), pp. 1–10. doi: 10.1371/journal.pone.0131365.

Quail, M. A. *et al.* (2012) 'A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers', *BMC Genomics*, 13(341), p. 13.

Quinlan, A. R. (2014) BEDTools: The Swiss-Army tool for genome feature analysis, Current Protocols in Bioinformatics. doi: 10.1002/0471250953.bi1112s47.

Quinn, J. P. and Bubb, V. J. (2014) 'SVA retrotransposons as modulators of gene expression', Mobile Genetic Elements, 4, p. e32102. doi: 10.4161/mge.32102.

Quinn, J. J. and Chang, H. Y. (2016) 'Unique features of long non-coding RNA biogenesis and function', Nature Reviews Genetics. Nature Publishing Group, 17(1), pp. 47–62. doi: 10.1038/nrg.2015.10.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raiz, J. *et al*. (2012) 'The non-autonomous retrotransposon SVA is trans -mobilized by the human LINE-1 protein machinery', Nucleic Acids Research, 40(4), pp. 1666–1683. doi: 10.1093/nar/gkr863.

Ravindran, S. (2012) 'Barbara McClintock and the discovery of jumping genes.' *Proceedings of the National Academy of Sciences of the United States of America*, 109(50), pp. 20198–20199. doi: 10.1073/pnas.1219372109.

Reilly, M. T. *et al.* (2013) 'The role of transposable elements in health and diseases of the central nervous system.', *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(45), pp. 17577–86. doi: 10.1523/JNEUROSCI.3369-13.2013.

Rhoads, A. and Au, K. F. (2015) 'PacBio Sequencing and Its Applications', Genomics, Proteomics and Bioinformatics. Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China, 13(5), pp. 278–289. doi: 10.1016/j.gpb.2015.08.002.

Rishishwar, L., Tellez Villa, C. E. and Jordan, I. K. (2015) 'Transposable element polymorphisms recapitulate human evolution.', *Mobile DNA*. Mobile DNA, 6, p. 21. doi: 10.1186/s13100-015-0052-6.

Rishishwar, L., Mariño-Ramírez, L. and Jordan, I. K. (2016) 'Benchmarking computational tools for polymorphic transposable element detection', Briefings in Bioinformatics, (April), p. bbw072. doi: 10.1093/bib/bbw072.

Roadmap Epigenomics Consortium *et al.* (2015) 'Integrative analysis of 111 reference human epigenomes', *Nature*, 518, pp. 317–330. doi: 10.1038/nature14248.

Rouchka, E. *et al.* (2010) 'Assessment of genetic variation for the LINE-1 retrotransposon from next generation sequence data', *BMC Bioinformatics*, 11(SUPPL. 9), p. S12. doi: 10.1186/1471-2105-11-S9-S12.

Rowe, S.J. and Tenesa, A. (2012) 'Human Complex Trait Genetics: Lifting the Lid of the Genomics Toolbox - from Pathways to Prediction', *Current Genomics*, 13(3), pp. 213–224. doi: 10.2174/138920212800543101.

Saleh, A., Macia, A. and Muotri, A. R. (2019) 'Transposable elements, inflammation, and neurological disease', Frontiers in Neurology, 10. doi: 10.3389/fneur.2019.00894.

Salvatore Mangiafico (2021). rcompanion: Functions to Support Extension Education Program Evaluation. R package version https://CRAN.R-project.org/package=rcompanion.

Savage, A. L., Bubb, V. J., Breen, G., & Quinn, J. P. (2013). Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. BMC Evolutionary Biology, 13(1). https://doi.org/10.1186/1471-2148-13-101

Savage, A. L. *et al.* (2019) 'Retrotransposons in the development and progression of amyotrophic lateral sclerosis', *Journal of Neurology, Neurosurgery and Psychiatry*, 90(3), pp. 284–293. doi: 10.1136/jnnp-2018-319210.

Schauer, S. N. *et al.* (2018) 'L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis', *Genome Research*, 28(5), pp. 639–653. doi: 10.1101/gr.226993.117.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) 'Biological insights from 108 schizophrenia-associated genetic loci', *Nature*, 511. doi: 10.1038/nature13595.

Scott, E. C. and Devine, S. E. (2017) 'The role of somatic L1 retrotransposition in human cancers', Viruses, 9(6), pp. 1–19. doi: 10.3390/v9060131.

Sheen, F. *et al.* (2000) 'Reading between the LINEs : Human Genomic Variation Induced by LINE-1 Retrotransposition', *Genome Research*, 10, pp. 1496–1508. doi: 10.1101/gr.149400.

Shen, L. *et al.* (1994) 'Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region', *Journal of Biological Chemistry*, 269(11), pp. 8466–8476. doi: 10.1016/s0021-9258(17)37217-4.

Shiina, T. *et al.* (2009) 'The HLA genomic loci map: Expression, interaction, diversity and disease', *Journal of Human Genetics*, 54(1), pp. 15–39. doi: 10.1038/jhg.2008.5.

Shin, W. *et al.* (2019) 'Novel discovery of line-1 in a korean individual by a target enrichment method', *Molecules and Cells*, 42(1), pp. 87–95. doi: 10.14348/molcells.2018.0351.

Shorter, E. (2017) 'Current research into the association between DNA copy number variation (CNV) and obesity', *Bioscience Horizons: The International Journal of Student Research*, 10, pp. 1–8. doi: 10.1093/biohorizons/hzx014.

Shukla, R. et al. (2013) 'Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma', Cell, 153(1), pp. 101–111. doi: 10.1016/j.cell.2013.02.032.

Smalheiser, NR. Torvik, VI. (2005) 'Mammalian microRNAs derived from genomic repeats', *Trends in Genetics*, 21(6), pp. 322–326. doi: 10.1016/j.tig.2005.04.008.

Smit, A. F. A. (1996) 'The origin of interspersed repeats in the human genome', Current Opinion in Genetics and Development, 6(6), pp. 743–748. doi: 10.1016/S0959-437X(96)80030-X.

Smit, A. F. (1999) 'Interspersed repeats and other mementos of transposable elements in mammalian genomes', Current Opinion in Genetics and Development, 9(6), pp. 657–663. doi: 10.1016/S0959-437X(99)00031-3.

Smoller, J. W. (2016) 'The Genetics of Stress-Related Disorders: PTSD, Depression, and Anxiety Disorders', *Neuropsychopharmacology*. Nature Publishing Group, 41(1), pp. 297–319. doi: 10.1038/npp.2015.266.

Song, M. and Boissinot, S. (2007) 'Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination', *Gene*, 390(1–2), pp. 206–213. doi: 10.1016/j.gene.2006.09.033.

Soriano, P., Meunier-Rotival, M. and Bernardi, G. (1983) 'The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes.', *Proceedings of the National Academy of Sciences of the United States of America*, 80(7), pp. 1816–20. doi: 10.1073/pnas.80.7.1816.

Spiegel, J., Adhikari, S., & Balasubramanian, S. (2020). The Structure and Function of DNA G-Quadruplexes. Trends in Chemistry, 2(2), 123–136. https://doi.org/10.1016/j.trechm.2019.07.002

Spirito, G. *et al.* (2019) 'Impact of polymorphic transposable elements on transcription in lymphoblastoid cell lines from public data', *BMC Bioinformatics*. BMC Bioinformatics, 20(Suppl 9), pp. 1–13. doi: 10.1186/s12859-019-3113-x.

Stankiewicz, P. Lupski, J. (2010) 'Structural variation in the human genome and its role in disease', *Annu Rev Med*, 61, pp. 437–455. doi: 10.1146/annurev-med-100708-204735.

Startek, M. *et al*. (2015) 'Genome-wide analyses of LINE-LINE-mediated nonallelic homologous recombination', Nucleic Acids Research, 43(4), pp. 2188–2198. doi: 10.1093/nar/gku1394.

Stewart, C. *et al.* (2011) 'A comprehensive map of mobile element insertion polymorphisms in humans', *PLoS Genetics*, 7(8). doi: 10.1371/journal.pgen.1002236.

Streva, V. A. *et al*. (2015) 'Sequencing, identification and mapping of primed L1 elements (SIMPLE) reveals significant variation in full length L1 elements between individuals', BMC Genomics, 16. doi: 10.1186/s12864-015-1374-y.

Su, M., Han, D., Boyd-kirkup, J., Yu, X., & Han, J. J. (2014). Report Evolution of Alu Elements toward Enhancers. CellReports, 7(2), 376–385. https://doi.org/10.1016/j.celrep.2014.03.011

Sudmant, P. H. *et al.* (2015) 'An integrated map of structural variation in 2,504 human genomes', *Nature*, 526(7571), pp. 75–81. doi: 10.1038/nature15394.

Sultana, T., van Essen, D., Siol, O., Bailly-Bechet, M., Philippe, C., Zine El Aabidine, A., Cristofari, G. (2019). The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. Molecular Cell, 74(3), 555-570.e7. https://doi.org/10.1016/j.molcel.2019.02.036

Sundaram, V. *et al.* (2014) 'innovation of gene regulatory networks Widespread contribution of transposable elements to the innovation of gene regulatory networks', *Genome Research*, 24, pp. 1963–1976. doi: 10.1101/gr.168872.113.

Tang, W., Mun, S., Joshi, A., & Han, K. (2018). Mobile elements contribute to the uniqueness of human genome with 15 , 000 human-specific insertions and 14 Mbp sequence increase. 25(July), 521–533. https://doi.org/10.1093/dnares/dsy022

Terry, D. M. and Devine, S. E. (2020) 'Aberrantly High Levels of Somatic LINE-1 Expression and Retrotransposition in Human Neurological Disorders', *Frontiers in Genetics*, 10(January), pp. 1–14. doi: 10.3389/fgene.2019.01244.

Tian, C. *et al.* (2017) 'Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections', *Nature Communications*. Springer US, 8(1). doi: 10.1038/s41467-017-00257-5.

The 1000 Genomes Project Consortium. (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. doi: 10.1038/nature15393.

The ENCODE Project Consortium (2012) 'An Integrated Encyclopedia of DNA Elements in the Human Genome', *Nature*, 489(7414), pp. 57–74. doi: 10.1038/nature11247.An.

Thung, D. T. *et al.* (2014) 'Mobster: accurate detection of mobile element insertions in next generation sequencing data', *Genome biology*, 15(488). doi: 10.1186/s13059-014-0488-x.

Touchon, M. and Rocha, E. P. C. (2007) 'Causes of insertion sequences abundance in prokaryotic genomes', *Molecular Biology and Evolution*, 24(4), pp. 969–981. doi: 10.1093/molbev/msm014.

Trizzino, M., Kapusta, A., & Brown, C. D. (2018). Transposable elements generate regulatory novelty in a tissue-specific fashion. BMC Genomics, 19(1), 1–12. https://doi.org/10.1186/s12864-018-4850-3

Tubio, J. M. C. *et al.* (2014) 'Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes', *Science*, 345(6196), p. 1251343. doi: 10.1126/science.1251343.

Ullu E, Tschudi C, .(1984) 'Alu sequences are processed 7SL RNA genes', *Nature*, 312(5990), pp. 171–172. doi: doi: 10.1038/312171a0.

Upton, K. R. *et al.* (2015) 'Ubiquitous L1 mosaicism in hippocampal neurons', *Cell*, 161(2), pp. 228–239. doi: 10.1016/j.cell.2015.03.026.

Viollet, S., Monot, C. and Cristofari, G. (2014) 'L1 retrotransposition: The snap-velcro model and its consequences', *Mobile genetic elements*, 4(1), pp. e28907–e28907. doi: 10.4161/mge.28907.

Visel, A. *et al.* (2007) 'VISTA Enhancer Browser - A database of tissue-specific human enhancers', *Nucleic Acids Research*, 35(SUPPL. 1), pp. 88–92. doi: 10.1093/nar/gkl822.

Waddell N, Pajic M, Patch AM, Chang DK. *et al*. (2015) 'Whole genomes redefine the mutational landscape of pancreatic cancer', *Nature*, 518(7540), pp. 125–131. doi: 10.1038/nature14169.Whole.

Wagstaff, B. J. *et al*. (2012) 'Rescuing Alu: Recovery of New Inserts Shows LINE-1 Preserves Alu Activity through A-Tail Expansion', PLoS Genetics, 8(8). doi: 10.1371/journal.pgen.1002842.

Wainschtein, P. *et al.* (2021) 'Recovery of trait heritability from whole genome sequence data', *bioRxiv*, p. 588020. doi: 10.1101/588020.

Wang, H., Xing, J., Grover, D., Hedges Kyudong Han, D. J., Walker, J. A., & Batzer, M. A. (2005). SVA elements: A hominid-specific retroposon family. Journal of Molecular Biology, 354(4), 994–1007. https://doi.org/10.1016/j.jmb.2005.09.085

Wang, J. Song, L. Grover, D. Azrak, S. Batzer, MA. Liang, P. (2006) 'db RIP: A Highly Integrated Databadse of Retrotransposon Insertion Polymorphisms in Humans', *Human Mutation*, 27(4), pp. 323–329. doi: 10.1002/humu.20307.

Wang, T. *et al.* (2007) 'Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53', *Proceedings of the National Academy of Sciences of the United States of America*, 104(47), pp. 18613–18618. doi: 10.1073/pnas.0703637104.

Warnefors, M., Pereira, V., & Eyre-Walker, A. (2010). Transposable elements: Insertion pattern and impact on gene expression evolution in hominids.

Molecular Biology and Evolution, 27(8), 1955–1962. https://doi.org/10.1093/molbev/msq084

Watkins, W. S. *et al*. (2020) 'The Simons Genome Diversity Project: A Global Analysis of Mobile Element Diversity', Genome Biology and Evolution, 12(6), pp. 779–794. doi: 10.1093/gbe/evaa086.

Wei, W. E. I. *et al.* (2001) 'Human L1 Retrotransposition: cis Preference versus trans Complementation', *Molecular and cellular Biology*, 21(4), pp. 1429–1439. doi: 10.1128/MCB.21.4.1429.

Weiner, DJ. Wigdor, EM. Ripke, S. *et al.* (2017) 'Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders', *Nature Genetics*, 49(7), pp. 978–985. doi: 10.1038/ng.3863.

Weischenfeldt, J. *et al.* (2013) 'Phenotypic impact of genomic structural variation: Insights from and for human disease', *Nature Reviews Genetics*. Nature Publishing Group, 14(2), pp. 125–138. doi: 10.1038/nrg3373.

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Witherspoon, D. J. *et al*. (2009) 'Alu repeats increase local recombination rates', BMC Genomics, 10(530). doi: 10.1186/1471-2164-10-530.

Witherspoon, D. J. *et al.* (2013) 'Mobile element scanning (ME-Scan) identifies thousands of novel Alu insertions in diverse human populations', *Genome Research*, 23(7), pp. 1170–1181. doi: 10.1101/gr.148973.112.

Wildschutte, J. H. *et al.* (2015) 'Discovery and characterization of Alu repeat sequences via precise local read assembly', *Nucleic Acids Research*, 43(21), pp. 10292–10307. doi: 10.1093/nar/gkv1089.

Wood, AR. Esko, T. Yang, J. Vedantam, S. *et al.* (2014) 'Defining the role of common variation in the genomic and biological architecture of adult human height', *Nature Genetics*, 46(11), pp. 1173–1186. doi: 10.1038/ng.3097.

Wray, G. A. *et al.* (2003) 'The evolution of transcriptional regulation in eukaryotes', *Molecular Biology and Evolution*, 20(9), pp. 1377–1419. doi: 10.1093/molbev/msg140.

Xing, J. *et al.* (2009) 'Mobile elements create structural variation : Analysis of a complete human genome', (801), pp. 1516–1526. doi: 10.1101/gr.091827.109.1516.

Yang, J. Bakshi, A. Zhu, Z. Hemani, G. *et al.* (2015) 'Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index', *Nature Genetics*, 47(10), pp. 1114–1120. doi: 10.1038/ng.3390.

Yang, F. and Wang, P. J. (2016) ‘Multiple LINEs of retrotransposon silencing mechanisms in the mammalian germline’, *Seminars in Cell and Developmental Biology*. Elsevier Ltd, 59, pp. 118–125. doi: 10.1016/j.semcdb.2016.03.001.

Yoo, I., & Mosa, A. S. (2015). Analysis of PubMed User Sessions Using a Full-Day PubMed Query Log: A Comparison of Experienced and Nonexperienced PubMed Users. *JMIR medical informatics*, *3*(3), e25. https://doi.org/10.2196/medinform.3740.

Yu, Q. *et al.* (2017) ‘Population-wide sampling of retrotransposon insertion polymorphisms using deep sequencing and efficient detection’, *Gigascience*, 6(9), pp. 1–11. doi: 10.1093/gigascience/gix066.

Zampella, J. G. *et al.* (2016) ‘A map of mobile DNA insertions in the NCI-60 human cancer cell panel’, *Mobile DNA*, 7(1), p. 20. doi: 10.1186/s13100-016-0078-4.

Zeng, L. *et al.* (2018) ‘ Genome-Wide Analysis of the Association of Transposable Elements with Gene Regulation Suggests that Alu Elements Have the Largest Overall Regulatory Impact ’, *Journal of Computational Biology*, 25(6), pp. 551–562. doi: 10.1089/cmb.2017.0228.

Zerbino, D. R. *et al.* (2015) ‘The Ensembl Regulatory Build’, *Genome Biology*, 16(1), pp. 1–8. doi: 10.1186/s13059-015-0621-5.

Zhang, W. *et al.* (2011) ‘Alu distribution and mutation types of cancer genes’, *BMC Genomics*, 12(157). doi: 10.1186/1471-2164-12-157

Zhang, L., Chen, J. and Zhao, Q. (2015) ‘Regulatory roles of Alu transcript on gene expression’, *Experimental Cell Research*. Elsevier, 338(1), pp. 113–118. doi: 10.1016/j.yexcr.2015.07.019.

Zhang, H. H. *et al.* (2020) ‘Horizontal transfer and evolution of transposable elements in vertebrates’, *Nature Communications*. Springer US, 11(1), pp. 1–10. doi: 10.1038/s41467-020-15149-4.

Appendix 1: Additional information about the 45 studies included in the curated database of this study

This supplementary table is an extension of table 4 (p. 36-37) and it includes a record of the following:

1. RTE Detection tool applied by the study
2. Information regarding the detection sensitivity/validation as reported by each respective study
3. Source of insertions data

Detection Sensitivity/Validation Key:

¶	|	PCR verification
P	|	Precision
R	|	Recall
S	|	Sensitivity
δ	|	Specificity
‡	|	False discovery rate
.avg	|	Average

Table 21: Studies included in the curated non-reference retrotransposable element (RTE) databases, including name of the detection tool used by each study, information about sensitivity/validation, and source file from which the insertions in the database got extracted from

#	Study ID	PMID	Detection tool	Reported Detection sensitivity	Data source
1	David et al., (2013)	23921633	Alu-detect	97%^{P}, 85% ^{R} 60% ¶	File S1
2	Sudmant et al., (2015)♦	26432246	MELT	83-96%^{S},4%‡	1KGP ftp. web
3	Witherspoon et al., (2013)‖	23599355	ME-Scan	95%^{S},44%¶,4.6%‡	Sup .txt file
4	Thung et al., (2014)♦	25348035	Mobster	90-99%^{S}, 91%¶,9%‡	File S2,S3,S4
5	Brandler et al., (2016)	27018473	Mobster	70% KNR (Fig. S7)	Tbl. S2
6	Baillie et al., (2011)♦⁝	22037309	RC-seq	100%¶ (Tbl. S6)	Tbl. S4 and S5
7	Shukla et al., (2013)♦⁝	23540693	RC-seq	98.5%¶ (Tbl. S5)	Tbl. S3, S4
8	Solyom et al., (2012)♦⁝	22968929	L1- & RC-seq	67.3%¶ (Tbl. S3)	Tbl. S1a.b, S2
9	Wildschutte et al., (2015)	26503250	RetroSeq	5%‡, 100%¶ (Tbl. S3)	Tbl S1a, S6
10	Evrony et al., (2015)♦	25569347	scTea	86-96%^{S}, 96%¶	Tbl. S3
11	Lee et al., (2012)♦⁝	22745252	Tea	100%¶ (Tbl. S7)	Tbl. S6, S8
12	Hormozdiari et al., (2011)	21131385	VariationHunter-2	98%¶ (Tbl. S4, S5)	Tbl. S1

13	Ha et al., (2016)	27478512	ME-Scan-SVA	89%^{S}, 55%¶ (Tbl.S2)	File S3
14	Stewart et al., (2011)♦⁝	21876680	SPANNER	95%¶, 4.5±0.8‡	Tbl. S1
15	Yu Q et al., (2017)	28938719	SID	86.7%$^{S.avg}$,96.9%¶ (Tbl. S7,S8)	Tbl. S11
16	Xing et al., (2009)‖	19439515	Computational	100%¶ (Tbl. S1)	Tbl. S3
17	Upton et al., (2015)♦	25860606	RC-seq	97.5%^{S}, 100%¶ (Tbl. S2)	Tbl. S2
18	Tubio et al., (2014)♦	25082706	TraFiC	73-83.9%^{S}, >99%^{δ} (Tbl. S4)	Tbl. S7
19	Shin et al., (2019)	30699287	Custom pipeline	92.5%¶ (Tbl.S4)	Tbl. S3
20	Schauer et al., (2018)	29643204	RC-seq	100%¶* (Tbl. S2)	Tbl. S2
21	Rouchka et al., (2010)	21044359	Computational	100%¶ (Tbl. 1)	Tbl. 1
22	Payer et al., (2017)	28465436	PCR of KNR (S3)	100%¶ (Dataset S3)	Dataset S3
23	Mir et al., (2014)♦⁝	25352549	euL1db	--	euL1db web
24	Kurnosov et al., (2015)♦	25689626	Computational	53%^{¶,avg*} (Tbl. S1)	Tbl. S1
25	Kuhn et al., (2014)♦	24847061	L1-seq	78%^{S}, 94%^{δ} (Fig. S2)	Dataset S2
26	Konkel et al., (2015)	26319576	Sanger sequencing of 343 Alu MEIs PCR validated by 1kGP (pilot phase)		Tbl. S1,S2
27	Kloosterman et al., (2015)	25883321	Mobster	77.6%^{S}, 100%¶ (Tbl. S2)	Tbl. S2
28	Iskow et al., (2010)♦⁝	20603005	L1-seq	89-97%¶* (Tbl.1)	Tbl. S1,S2,S3
29	Helman et al., (2014)♦⁝	24823667	TranspoSeq	99%^{S}, 83%¶*(Tbl. S1)	Tbl.S2
30	Hehir-Kwa et al., (2016)	27708267	Mobster	96%¶ (Data S1)	Nlgenome web
31	Feusier et al., (2017)	28770012	ME-Scan	58%¶ (Tbl. S2,S3,S6)	Tbl. S11
32	Ewing et al., (2015)♦	26260970	L1-seq	>93%^{S} (Tbl. S1)	Tbl. S3a-d
33	Ewing et al., (2011)♦⁝	20980553	Custom pipeline (Fig.3)	80.5% (Tbl. S1)	Tbl. S2
34	Ewing et al., (2010)♦⁝	20488934	Perl script	>80%^{δ}, 93±4%^{S}, ~7%‡ (S.pdf)	Tbl. S1
35	Evrony et al., (2012)♦⁝	23101622	Custom pipeline	81±6%^{S}, 94%¶	Tbl. S3
36	Erwin et al., (2016)♦	27618310	Machine learning-based	80%¶ (Tbl. S3)	Tbl. S3
37	Doyle et al., (2017)	28585566	L1-seq	50%¶ (Tbl.S2)	Tbl. S5
38	Cardelli et al., (2012)	22495107	2 AIP methods	67-90%^{S} (Tbl.1), 100%¶ (Tbl.2)	Tbl. 2
39	Beck et al., (2010)♦⁝	20602998	Experimental Strategy	Confirmed 18/68¶ in ABC13	Tbl. S2
40	Arokium et al., (2014)♦	25289675	Adapted L1-Seq	94%¶ (Tbl. S3)	Tbl. S2
41	Nguyen et al., (2018)	29949758	RC-seq	93.5%¶* (Tbl.S2)	Tbl. S2
42	Carreira et al., (2016)♦	27843499	RC-seq	>82%^{S} (Tbl. S1)	Tbl. S2
43	Achanta et al., (2016)	27843500	TIPseq	100%¶ (Tbl.2)	Tbl. 2
44	Streva et al., (2015)♦	25887476	SIMPLE	94%^{S}, 100%¶ (Tbl. S4)	Tbl. S2
45	Scott et al., (2016)♦	27197217	Adapted L1-Seq and MELT	84.4%¶ (Tbl. S5)	Data S2

Appendix 2: Full list of RTEs in LD with TAS

Table 22: Full list of RTEs in LD with TAS. This supplementary table is an expansion of table 17 (p.132-133) and it includes all RTEs in LD ($r^2 > 0.6$) with GWS TAS identified in cohorts of European descents. Note: When there are RTEs in LD with a TAS that have been identified by multiple GWAS, the table includes the strongest GWAS signal. Table ordered by r^2 value.

Chr	Start	RTE (1kGP name)	TAS	Trait	P-val	r^2	Trait_PMID
6	141414904	ALU_5647	rs113803678	Body mass index	4E-08	0.92	30595370
3	85576571	LINE1_629	rs62250759	Self-reported risk-taking behaviour	5E-12	0.92	29391395
12	24868717	LINE1_2249	rs61914312	Hair color	3E-10	0.91	30595370
1	174484646	ALU_604	rs140588606	Feeling miserable	6E-10	0.91	29500382
1	174484646	ALU_604	rs77417259	Feeling miserable	3E-10	0.91	29500382
9	11329329	ALU_7283	rs2152261	Menarche (age at onset)	2E-13	0.91	30595370
12	56753252	ALU_9228	rs2066819	Psoriasis	5E-17	0.88	23143594
1	174484646	ALU_604	rs140581634	Feeling miserable	2E-08	0.87	29500382
12	120130849	ALU_9540	rs17442937	Red cell distribution width	4E-08	0.85	30595370
17	44153977	SVA_706	rs12150229	Ease of getting up in the morning	4E-09	0.84	30804565
17	44153977	SVA_706	rs12150672	Red blood cell count	4E-09	0.84	28017375
17	44153977	SVA_706	rs12185268	Parkinson's disease	3E-14	0.84	21738487
17	44153977	SVA_706	rs12373124	Male-pattern baldness	5E-10	0.84	22693459
17	44153977	SVA_706	rs17563683	Hemoglobin concentration	2E-28	0.84	27863252
17	44153977	SVA_706	rs17563986	Cognitive ability	5E-12	0.84	29186694
17	44153977	SVA_706	rs17650842	Irritable mood	2E-12	0.84	29500382
17	44153977	SVA_706	rs1864325	Lumbar spine bone mineral density	5E-11	0.84	22504420
17	44153977	SVA_706	rs2214258	Neuroticism	2E-25	0.84	29255261
17	44153977	SVA_706	rs241036	Experiencing mood swings	8E-20	0.84	29500382
17	44153977	SVA_706	rs241036	Menarche (age at onset)	7E-13	0.84	30595370
17	44153977	SVA_706	rs2942168	Parkinson's disease	1E-28	0.84	21292315
17	44153977	SVA_706	rs4606752	Reticulocyte count	1E-17	0.84	27863252
17	44153977	SVA_706	rs55657917	Feeling hurt/Mood swings	7E-29	0.84	29500382
17	44153977	SVA_706	rs55657917	Accelerometer-based physical activity measurement (average acceleration)	5E-12	0.84	29899525
17	44153977	SVA_706	rs56303031	Heel bone mineral density	6E-24	0.84	30595370
17	44153977	SVA_706	rs56319902	Educational attainment (years of education)	6E-33	0.84	30038396
17	44153977	SVA_706	rs62055546	Alcohol consumption (drinks per week)	8E-25	0.84	30643258
17	44153977	SVA_706	rs62055701	Irritable mood	6E-13	0.84	29500382
17	44153977	SVA_706	rs62055935	Feeling nervous	2E-15	0.84	29500382
17	44153977	SVA_706	rs62057061	Depressed affect	2E-22	0.84	29942085
17	44153977	SVA_706	rs62057107	Educational attainment	5E-38	0.84	30038396

17	44153977	SVA_706	rs75022332	Worry too long after an embarrassing experience	2E-08	0.84	29500382
17	44153977	SVA_706	rs76761706	Neurociticism	7E-32	0.84	29500382
17	44153977	SVA_706	rs79412431	Lung function	3E-49	0.84	30804560
17	44153977	SVA_706	rs79857651	Experiencing mood swings	9E-20	0.84	29500382
17	44153977	SVA_706	rs8072451	Subcortical brain region volumes	1E-08	0.84	25607358
9	16682313	ALU_7311	rs12335424	Height	2E-21	0.84	30595370
15	76826019	ALU_10819	rs166906	Estimated glomerular filtration rate	5E-13	0.83	31152163
15	76826019	ALU_10819	rs506000	Estimated glomerular filtration rate	2E-15	0.83	31152163
17	44153977	SVA_706	rs112333322	Experiencing mood swings	4E-19	0.83	29500382
17	44153977	SVA_706	rs117124984	Daytime nap	3E-13	0.83	30804565
17	44153977	SVA_706	rs17577369	Feeling miserable	4E-12	0.83	29500382
17	44153977	SVA_706	rs17649553	Parkinson's disease	1E-68	0.83	28892059
17	44153977	SVA_706	rs17661015	Irritable mood	4E-12	0.83	29500382
17	44153977	SVA_706	rs17661015	Feeling hurt	3E-28	0.83	29500382
17	44153977	SVA_706	rs1981997	Interstitial lung disease	9E-14	0.83	23583980
17	44153977	SVA_706	rs35524223	Lung function (FEV1)	1E-13	0.83	28166213
17	44153977	SVA_706	rs56280951	Feeling miserable	8E-13	0.83	29500382
17	44153977	SVA_706	rs62055544	Feeling fed-up	5E-16	0.83	29500382
17	44153977	SVA_706	rs79301522	Neurociticism	1E-30	0.83	29500382
17	44153977	SVA_706	rs8070723	Progressive supranuclear palsy/Parkinson's disease	2E-118	0.83	21685912
17	44153977	SVA_706	rs919462	Male-pattern baldness	1E-26	0.83	29146897
17	44153977	SVA_706	rs111433752	Neuroticism	9E-12	0.83	27067015
17	44153977	SVA_706	rs17689882	Subcortical brain region volumes	8E-09	0.83	25607358
17	44153977	SVA_706	rs2106785	Irritable mood	3E-13	0.83	29500382
17	44153977	SVA_706	rs2106786	Red blood cell count	3E-36	0.83	27863252
17	44153977	SVA_706	rs365825	Parkinson's disease	4E-32	0.83	27182965
17	44153977	SVA_706	rs393152	Parkinson's disease	2E-16	0.83	19915575
17	44153977	SVA_706	rs62061733	Eosinophil counts	3E-29	0.83	30595370
17	44153977	SVA_706	rs62061733	Feeling hurt	2E-28	0.83	29500382
12	56753252	ALU_9228	rs59917308	Height	3E-32	0.83	30595370
17	44153977	SVA_706	rs112010353	Self-reported math ability	2E-08	0.83	30038396
2	210260754	ALU_1947	rs1080278	Lung function (FVC)	1E-19	0.83	30595370
17	44153977	SVA_706	rs1991556	Lung function (FVC)	1E-53	0.83	30595370
17	44153977	SVA_706	rs1991556	Sleep duration	3E-09	0.83	30531941
16	75655176	ALU_11116	rs61537885	Smoking initiation (ever regular vs never regular)	8E-09	0.82	30643251
17	44153977	SVA_706	rs80103986	Hand grip strength	1E-09	0.82	29313844
4	134596423	LINE1_967	rs12507927	Highest math class taken	3E-11	0.82	30038396
17	44153977	SVA_706	rs17652520	Medication use (anilides)	8E-13	0.82	31015401
17	44153977	SVA_706	rs17652520	Neuroticism	2E-27	0.82	30595370
1	174484646	ALU_604	rs75035127	Feeling miserable	7E-09	0.82	29500382
1	169524859	LINE1_164	rs6128	Blood protein levels	2E-26	0.82	29875488
12	56753252	ALU_9228	rs11575234	Inflammatory skin disease	2E-12	0.82	25574825
17	44153977	SVA_706	rs62063281	Number of sexual partners	4E-15	0.82	30643258
17	44153977	SVA_706	rs62063281	Osteoarthritis (hip)	5E-12	0.82	30664745
17	44153977	SVA_706	rs17665188	Experiencing mood swings	1E-18	0.82	29500382
17	44153977	SVA_706	rs62065453	Neuroticism	2E-24	0.82	29255261
17	44153977	SVA_706	rs62065453	Feeling nervous/Irritable mood	6E-15	0.82	29500382

1	163639693	ALU_559	rs12564153	Lung function (FEV1/FVC)	1E-09	0.82	30595370
12	56753252	ALU_9228	rs2066807	Psoriasis	5E-12	0.81	25903422
12	56753252	ALU_9228	rs2066807	Height	1E-13	0.81	20881960
5	109051004	ALU_4562	rs4388249	Schizophrenia	8E-09	0.81	28991256
17	44153977	SVA_706	rs62064364	Macular thickness	4E-35	0.81	30535121
10	106566893	ALU_8208	rs61867293	Depression	7E-10	0.81	29700475
21	33050849	ALU_12379	rs17660708	LDL cholesterol	1E-10	0.81	30275531
17	44153977	SVA_706	rs2732631	Macular thickness	3E-35	0.81	30535121
17	44153977	SVA_706	rs4471723	Feeling guilty	5E-09	0.81	29500382
17	44153977	SVA_706	rs8080583	Cognitive ability	1E-08	0.81	29186694
17	44153977	SVA_706	rs9303525	Intracranial volume	8E-15	0.81	22504418
20	26190974	ALU_12132	rs6051320	Lung function (FEV1/FVC)	2E-08	0.81	30595370
11	54958589	ALU_8580	rs77584654	Heel bone mineral density/ Height	5E-17	0.81	30595370
17	44153977	SVA_706	rs2732708	Feeling miserable	3E-11	0.80	29500382
17	44153977	SVA_706	rs2732708	Neuroticism	2E-23	0.80	29255261
17	44153977	SVA_706	rs62057151	Feeling worry	1E-09	0.80	29500382
7	18273084	ALU_5868	rs1528683	Lung function (FVC)	2E-17	0.80	30595370
12	77965056	ALU_9355	rs17788937	Myopia (pathological)	4E-15	0.80	23049088
17	44153977	SVA_706	rs2696532	Feeling guilty	4E-08	0.80	29500382
17	44153977	SVA_706	rs242559	General cognitive ability	1E-13	0.80	29844566
11	43877448	ALU_8559	rs1061810	Type 2 diabetes	4E-10	0.80	28566273
8	110101605	ALU_7037	rs28499085	Pulse pressure	3E-13	0.79	30224653
14	92619420	SVA_615	rs34016308	Myopia	4E-14	0.79	27182965
5	40041345	LINE1_1097	rs10053502	Myopia (pathological)	1E-16	0.79	23049088
17	44153977	SVA_706	rs58879558	Red blood cell count	3E-98	0.79	30595370
17	44153977	SVA_706	rs77804065	Feeling guilty	6E-10	0.79	29500382
17	44153977	SVA_706	rs77804065	Neuroticism	1E-31	0.79	29942085
1	174484646	ALU_604	rs75650221	Negative Feelings	3E-10	0.79	29500382
1	174484646	ALU_604	rs75650221	Ease of getting up in the morning	4E-18	0.79	30804565
5	25233926	ALU_4154	rs111257433	General risk tolerance	5E-10	0.79	30643258
17	44153977	SVA_706	rs62062288	Neurociticism	6E-32	0.79	29500382
17	44153977	SVA_706	rs62062288	Alcohol use disorder (total score)	5E-10	0.79	30336701
17	44153977	SVA_706	rs62062288	Negative Feelings	7E-10	0.79	29500382
17	44153977	SVA_706	rs62062288	Risk-taking tendency	1E-29	0.79	30643258
6	56387576	ALU_5205	rs4288197	Heel bone mineral density	5E-17	0.78	30595370
14	92619420	SVA_615	rs11160044	Spherical equivalent or myopia (age of diagnosis)	7E-11	0.78	29808027
11	49282683	ALU_8572	rs7103270	HDL cholesterol x physical activity interaction (2df test)	7E-12	0.78	30670697
4	76993824	ALU_3412	rs7693693	Blood protein levels	2E-17	0.78	29875488
1	219558910	ALU_810	rs75128958	Lung function (FEV1/FVC)	2E-23	0.78	30804560
1	219558910	ALU_810	rs75128958	Heel bone mineral density	1E-08	0.78	30595370
2	30669993	ALU_1087	rs28538173	Eosinophil counts	3E-09	0.78	30595370
6	96009421	ALU_5389	rs80268500	Blood protein levels	0E+00	0.78	29875488
16	80848077	ALU_11145	rs34018670	Monocyte count	5E-09	0.78	27863252
1	174484646	ALU_604	rs115073088	Chronotype	4E-12	0.78	30696823
1	119553366	LINE1_122	rs3790553	Male-pattern baldness	4E-19	0.78	30573740
12	56753252	ALU_9228	rs2066808	Psoriasis	6E-10	0.77	25574825
12	56753252	ALU_9228	rs36207871	Inflammatory skin disease	3E-12	0.77	25574825
3	193354185	LINE1_769	rs34023161	Highest math class taken	3E-08	0.77	30038396
17	44153977	SVA_706	rs76640332	Lymphocyte percentage of white cells	5E-13	0.77	27863252

12	56753252	ALU_9228	rs59626664	General risk tolerance	2E-09	0.77	30643258
3	193354185	LINE1_769	rs11925699	Educational attainment	3E-08	0.77	30038396
1	180857564	ALU_629	rs1043069	Systolic blood pressure	5E-15	0.77	30224653
9	4237141	ALU_7235	rs2224492	Intraocular pressure	4E-16	0.77	29235454
17	44153977	SVA_706	rs1378358	Negative Feelings	2E-11	0.76	29500382
17	44153977	SVA_706	rs1378358	Neurociticism	2E-27	0.76	29500382
17	44153977	SVA_706	rs199456	Macular thickness	3E-28	0.76	30535121
17	44153977	SVA_706	rs538628	Feeling nervous	9E-13	0.76	29500382
15	47507342	LINE1_2640	rs12914084	Neuroticism	3E-08	0.76	30643256
15	47507342	LINE1_2640	rs6493265	Educational attainment (years of education)	2E-17	0.76	30038396
2	652672	ALU_958	rs10189761	Obesity	6E-24	0.76	23563607
2	652672	ALU_958	rs12714415	Heel bone mineral density	4E-09	0.76	28869591
2	652672	ALU_958	rs12995480	C-reactive protein levels	1E-10	0.76	30388399
2	652672	ALU_958	rs13021737	Body mass index	4E-69	0.76	30108127
2	652672	ALU_958	rs13396935	Smoking status (ever vs never smokers)	4E-13	0.76	30643258
2	652672	ALU_958	rs4854344	Body mass index (joint analysis main effects and physical activity interaction)	9E-23	0.76	28448500
2	652672	ALU_958	rs5017302	Menarche (age at onset)	5E-38	0.76	30595370
2	652672	ALU_958	rs6548238	Body mass index	1E-18	0.76	19079261
2	652672	ALU_958	rs6725549	Body mass index	1E-74	0.76	26426971
2	652672	ALU_958	rs6748821	Obese vs. thin	8E-21	0.76	30677029
2	652672	ALU_958	rs6752706	Lung function (FEV1/FVC)	2E-13	0.76	30595370
2	652672	ALU_958	rs6755502	Waist/Hip circumference	2E-30	0.76	25673412
17	44153977	SVA_706	rs199443	Feeling fed-up	4E-13	0.76	29500382
17	44153977	SVA_706	rs199447	Neuroticism	2E-26	0.76	29942085
17	44153977	SVA_706	rs199533	Parkinson's disease	1E-14	0.76	19915575
8	109135936	SVA_389	rs617117	Macular thickness	2E-09	0.76	30535121
12	28163331	ALU_9104	rs1838564	Breast size	1E-12	0.76	27182965
4	22043212	ALU_3139	rs62301574	Insomnia	1E-08	0.76	30804565
14	39875097	LINE1_2544	rs34983854	Systolic blood pressure	2E-11	0.76	30224653
6	140417842	ALU_5637	rs62429521	Insomnia	2E-09	0.76	30804565
11	49282683	ALU_8572	rs658118	HDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df)	1E-49	0.76	30698716
14	55795871	ALU_10325	rs10146637	White blood cell count	4E-11	0.76	30595370
6	97017683	ALU_5395	rs11153071	Systolic blood pressure	3E-15	0.76	30595370
2	652672	ALU_958	rs4854349	Childhood body mass index	5E-22	0.76	26604143
8	109135936	SVA_389	rs392783	Hair color	2E-23	0.76	30595370
6	97017683	ALU_5395	rs11153018	Systolic blood pressure	1E-11	0.76	30578418
15	49609604	ALU_10695	rs11632038	Lung adenocarcinoma	5E-10	0.76	28604730
6	56387576	ALU_5205	rs112462597	Heel bone mineral density	6E-16	0.76	30048462
2	652672	ALU_958	rs62105306	Body mass index (adult)	3E-28	0.75	28430825
11	49282683	ALU_8572	rs11040595	Systolic blood pressure	1E-11	0.75	30578418
11	49282683	ALU_8572	rs77828979	Intraocular pressure	6E-12	0.75	29617998
2	30669993	ALU_1087	rs829636	Eczema	6E-09	0.75	30595370
4	134596423	LINE1_967	rs1157684	Self-reported math ability	3E-14	0.75	30038396
4	134596423	LINE1_967	rs981033	Self-reported math ability	1E-14	0.75	30038396
10	34571038	ALU_7901	rs610493	Height	2E-10	0.75	30595370
8	71914591	ALU_6806	rs2639935	Lung function (FEV1/FVC)	3E-08	0.75	30595370
6	97017683	ALU_5395	rs35410524	Systolic blood pressure	5E-10	0.75	27841878
5	56109723	ALU_4294	rs16886364	Breast cancer (early onset)	5E-12	0.75	24493630

15	47507342	LINE1_2640	rs11853760	Educational attainment (years of education)	5E-13	0.75	30038396
15	47507342	LINE1_2640	rs2860049	Educational attainment	2E-15	0.75	30038396
2	144010793	LINE1_410	rs62171698	Body mass index	1E-10	0.75	30595370
2	144010793	LINE1_410	rs6710871	Body mass index	3E-19	0.75	30108127
6	45260479	ALU_5132	rs10948222	Height	1E-20	0.75	25282103
1	174484646	ALU_604	rs41304550	Feeling miserable	4E-09	0.75	29500382
1	174484646	ALU_604	rs76785379	Feeling miserable	2E-09	0.75	29500382
2	652672	ALU_958	rs7561317	Weight	2E-18	0.75	19079260
2	652672	ALU_958	rs7561317	Body mass index	4E-17	0.75	19079260
12	28163331	ALU_9104	rs12371778	Breast size	1E-08	0.75	22747683
2	652672	ALU_958	rs62105303	Breast size	3E-08	0.74	27182965
3	42898420	ALU_2319	rs4683346	Granulocyte percentage of myeloid white cells	5E-19	0.74	27863252
17	44153977	SVA_706	rs9896243	Worry	2E-11	0.74	29942085
2	652672	ALU_958	rs12463617	Body mass index	3E-17	0.74	23669352
1	174484646	ALU_604	rs77560793	Body mass index	7E-13	0.74	30595370
6	46310306	LINE1_1293	rs10498767	Body mass index	2E-10	0.74	30595370
16	75655176	ALU_11116	rs4888444	Age of smoking initiation	9E-09	0.74	30643251
8	63344481	ALU_6774	rs16928927	Rapid automised naming of letters	2E-08	0.74	30741946
17	44153977	SVA_706	rs17688916	Sleep duration (long sleep)	1E-11	0.74	30846698
17	44153977	SVA_706	rs17688916	Feeling worry	3E-11	0.74	29500382
17	46505002	ALU_11333	rs7207826	Ovarian cancer	2E-17	0.74	28346442
2	198763462	ALU_1894	rs700655	Red blood cell count	3E-10	0.74	30595370
17	44153977	SVA_706	rs57222984	Snoring	3E-11	0.73	30804565
7	8019027	LINE1_1448	rs56195338	Eosinophil counts	7E-10	0.73	30595370
2	198763462	ALU_1894	rs700641	Morning vs. evening chronotype	5E-10	0.73	26835600
11	43877448	ALU_8559	rs11555762	Body mass index	5E-14	0.73	29273807
17	44153977	SVA_706	rs199441	Male-pattern baldness	1E-181	0.73	30573740
17	44153977	SVA_706	rs199441	Neuroticism	3E-20	0.73	29255261
17	44153977	SVA_706	rs199441	Feeling hurt/Mood swings	4E-24	0.73	29500382
2	198763462	ALU_1894	rs12472359	Morning person	1E-27	0.73	30696823
1	232587774	ALU_865	rs4649269	Hair color	2E-08	0.73	30595370
17	44153977	SVA_706	rs199525	Intracranial volume	2E-20	0.73	30818988
17	44153977	SVA_706	rs199525	Feeling guilty	4E-08	0.73	29500382
17	44153977	SVA_706	rs199525	Lung function (FEV1)	1E-09	0.73	30061609
3	152053972	ALU_2814	rs182314334	Prostate cancer	4E-11	0.73	29892016
3	182299013	ALU_2986	rs4484214	Chronotype	7E-10	0.73	30696823
3	182299013	ALU_2986	rs6443810	Morningness	2E-10	0.73	30804565
7	8019027	LINE1_1448	rs7804306	Blood Cell counts	2E-09	0.73	27863252
17	46505002	ALU_11333	rs9303542	Ovarian cancer	5E-15	0.73	25581431
5	43870854	ALU_4250	rs79904209	Lung function (FEV1/FVC)	3E-17	0.73	30595370
2	652672	ALU_958	rs6744646	Body mass index	4E-111	0.73	30595370
6	45260479	ALU_5132	rs2396502	Osteoarthritis (hip)	2E-12	0.73	30664745
9	118509752	ALU_7676	rs12344818	Height	5E-33	0.72	30595370
15	47507342	LINE1_2640	rs12903078	Neurociticism	4E-08	0.72	29500382
1	78607067	ALU_276	rs540742	Body mass index	7E-09	0.72	29273807
6	97017683	ALU_5395	rs9486719	Migraine	6E-21	0.72	27182965
5	87399827	ALU_4449	rs2217250	Feeling tense	1E-08	0.72	29500382
6	153429856	SVA_315	rs9479509	Diastolic blood pressure	1E-09	0.72	30224653
12	56753252	ALU_9228	rs808919	Blood protein levels	7E-16	0.72	30072576
2	652672	ALU_958	rs2867125	Body mass index	5E-75	0.72	26426971
2	652672	ALU_958	rs2867125	Type 2 diabetes	4E-10	0.72	30054458
2	652672	ALU_958	rs2903492	Body mass index	6E-15	0.72	23563607

2	652672	ALU_958	rs6711012	Obesity	3E-40	0.72	23563607
2	652672	ALU_958	rs6711254	Age of smoking initiation	1E-16	0.72	30643251
2	652672	ALU_958	rs6731872	Smoking initiation (ever regular vs never regular)	4E-31	0.72	30643251
2	652672	ALU_958	rs6731872	Alcohol consumption (drinks per week)	4E-09	0.72	30643251
2	652672	ALU_958	rs7567570	Smoking status	7E-12	0.72	30595370
2	652672	ALU_958	rs939584	Body mass index	6E-23	0.72	28892062
7	50473286	ALU_6027	rs80271829	Monocyte count	4E-15	0.72	27863252
6	97017683	ALU_5395	rs3798293	Pulse pressure	6E-09	0.72	30578418
17	44153977	SVA_706	rs199505	Depressed affect	1E-18	0.71	29942085
17	44153977	SVA_706	rs199515	Parkinson's disease	3E-17	0.71	22451204
17	44153977	SVA_706	rs70600	Irritable mood	3E-11	0.71	29500382
5	56109723	ALU_4294	rs1017226	Breast cancer (early onset)	6E-11	0.71	24493630
5	56109723	ALU_4294	rs16886397	Breast cancer (early onset)	4E-12	0.71	24493630
5	56109723	ALU_4294	rs16886448	Breast cancer (early onset)	2E-12	0.71	24493630
9	17062597	ALU_7312	rs112488223	Heel bone mineral density	5E-16	0.71	30595370
9	17062597	ALU_7312	rs79439080	Height	7E-24	0.71	30595370
11	65984338	ALU_8622	rs10896090	Bipolar disorder	2E-08	0.71	31043756
15	47507342	LINE1_2640	rs1563245	Neuroticism	5E-11	0.71	29255261
15	47507342	LINE1_2640	rs1563245	Well-being spectrum (multivariate analysis)	1E-08	0.71	30643256
8	76080146	ALU_6846	rs72656192	Height	5E-10	0.71	30595370
2	652672	ALU_958	rs66906321	Obesity (extreme)	1E-34	0.71	30677029
2	652672	ALU_958	rs2947411	Smoking initiation	5E-10	0.70	30617275
2	652672	ALU_958	rs2947411	Menarche (age at onset)	2E-19	0.70	25231870
17	44153977	SVA_706	rs17690703	Idiopathic pulmonary fibrosis	6E-09	0.70	24429156
5	56109723	ALU_4294	rs3822625	Breast cancer (early onset)	5E-12	0.70	24493630
10	65356114	ALU_8023	rs41274072	Reticulocyte count	4E-16	0.70	27863252
9	34703699	SVA_402	rs11574914	Rheumatoid arthritis	2E-15	0.70	24390342
12	28163331	ALU_9104	rs10771399	Breast cancer	5E-34	0.70	25751625
3	103419436	ALU_2577	rs57714592	Highest math class taken	1E-08	0.70	30038396
3	175673943	ALU_2951	rs57939424	Educational attainment (years of education)	2E-13	0.70	30038396
1	227502452	ALU_841	rs112779011	Hair color	7E-11	0.70	30595370
9	34703699	SVA_402	rs2812378	Chronic inflammatory diseases (ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis, ulcerative colitis) (pleiotropy)	6E-09	0.70	26974007
9	34703699	SVA_402	rs2812378	Rheumatoid arthritis (ACPA-positive)	5E-11	0.70	24532676
3	182299013	ALU_2986	rs7652216	Body mass index	3E-08	0.69	30595370
17	44153977	SVA_706	rs417968	Educational attainment (years of education)	6E-16	0.69	30595370
4	145395308	ALU_3782	rs13142879	Post bronchodilator FEV1	1E-11	0.69	26634245
4	145395308	ALU_3782	rs35440220	Post bronchodilator FEV1	2E-11	0.69	26634245
4	145395308	ALU_3782	rs7656246	Post bronchodilator FEV1	1E-11	0.69	26634245
4	145395308	ALU_3782	rs1813337	Post bronchodilator FEV1	2E-11	0.69	26634245
4	145395308	ALU_3782	rs1961266	Post bronchodilator FEV1	2E-11	0.69	26634245
4	145395308	ALU_3782	rs56304346	Post bronchodilator FEV1	6E-11	0.69	26634245
4	145395308	ALU_3782	rs7666298	Post bronchodilator FEV1	2E-11	0.69	26634245
7	91751552	ALU_6200	rs10644111	Breast cancer	3E-11	0.69	29059683
7	91751552	ALU_6200	rs6964587	Breast cancer	9E-11	0.69	29059683
17	44153977	SVA_706	rs4630591	Feeling fed-up	1E-17	0.69	29500382

13	46647748	ALU_9748	rs9526137	Blood protein levels	2E-15	0.69	29875488
13	46647748	ALU_9748	rs9526138	Blood protein levels	1E-109	0.69	29875488
6	70327733	ALU_5250	rs7761673	Body mass index	5E-11	0.69	30595370
13	46647748	ALU_9748	rs3742264	Blood protein levels	3E-83	0.69	28240269
12	28417298	ALU_9107	rs10843151	Waist circumference adjusted for BMI (joint analysis main effects and physical activity interaction)	1E-09	0.69	28448500
13	46647748	ALU_9748	rs9567617	Blood protein levels	5E-109	0.69	29875488
21	28221359	ALU_12341	rs162531	Lung function (FEV1/FVC)	8E-11	0.69	30595370
4	145395308	ALU_3782	rs1505770	Post bronchodilator FEV1	2E-11	0.68	26634245
4	145395308	ALU_3782	rs2130499	Post bronchodilator FEV1	2E-11	0.68	26634245
4	145395308	ALU_3782	rs55694701	Post bronchodilator FEV1	1E-11	0.68	26634245
4	145395308	ALU_3782	rs56268708	Post bronchodilator FEV1	2E-11	0.68	26634245
4	145395308	ALU_3782	rs62334742	Post bronchodilator FEV1	2E-11	0.68	26634245
13	46647748	ALU_9748	rs4942471	Blood protein levels	3E-115	0.68	30072576
7	91751552	ALU_6200	rs35417517	Breast cancer	8E-11	0.68	29059683
7	91751552	ALU_6200	rs35522438	Breast cancer	1E-09	0.68	29059683
7	91751552	ALU_6200	rs6465353	Educational attainment (years of education)	3E-08	0.68	30038396
3	114915094	ALU_2641	rs7643617	Menarche (age at onset)	3E-11	0.68	30595370
11	49282683	ALU_8572	rs11040204	Intraocular pressure	4E-14	0.68	29617998
11	49282683	ALU_8572	rs2202454	Medication use (diuretics)	8E-11	0.68	31015401
16	75655176	ALU_11116	rs117657830	Smoking initiation (ever regular vs never regular)	3E-09	0.68	30643251
8	8920127	ALU_6560	rs2953805	Neuroticism	2E-29	0.68	29255261
6	97017683	ALU_5395	rs11759769	Migraine	2E-12	0.68	23793025
17	44153977	SVA_706	rs4327090	Highest math class taken	1E-10	0.68	30038396
2	11353711	ALU_1002	rs56211149	Height	2E-13	0.68	30595370
7	120538086	ALU_6336	rs201852005	Heel bone mineral density	4E-14	0.68	30048462
7	120538086	ALU_6336	rs73427834	Heel bone mineral density	1E-47	0.68	30595370
8	8920127	ALU_6560	rs2921378	Neuroticism	1E-27	0.68	30643256
3	110271029	ALU_2610	rs1398346	Chronotype	4E-08	0.68	30696823
17	44153977	SVA_706	rs199529	Intraocular pressure	4E-08	0.68	30054594
18	53146075	ALU_11714	rs4801157	Depressed affect	2E-10	0.68	29942085
3	175673943	ALU_2951	rs72622559	Educational attainment	6E-13	0.68	30038396
6	66163982	ALU_5227	rs7449561	Educational attainment	1E-11	0.67	30038396
12	28417298	ALU_9107	rs11049611	Height	3E-32	0.67	25282103
7	38209213	ALU_5970	rs9801416	Height	2E-09	0.67	30595370
1	191477465	ALU_689	rs677325	Subjective well-being	1E-08	0.67	29292387
17	32629274	ALU_11283	rs9906695	Monocyte percentage of white cells	3E-11	0.67	27863252
17	32629274	ALU_11283	rs9909465	Granulocyte percentage of myeloid white cells	1E-09	0.67	27863252
11	49282683	ALU_8572	rs10839204	Medication use (agents acting on the renin-angiotensin system)	2E-09	0.67	31015401
9	4237141	ALU_7235	rs6476827	Intraocular pressure	5E-33	0.67	29617998
8	71914591	ALU_6806	rs7007887	Snoring	1E-13	0.67	30804565
14	60741400	ALU_10351	rs1887103	Hair color	3E-16	0.67	30595370
9	94058487	LINE1_1863	rs7048945	Male-pattern baldness	7E-10	0.67	30573740
10	46074893	ALU_7934	rs17157836	Lymphocyte counts	2E-09	0.67	27863252
10	46074893	ALU_7934	rs34731408	Neutrophil percentage of white cells	7E-10	0.67	27863252
10	46074893	ALU_7934	rs34897497	Mean corpuscular volume	1E-52	0.67	27863252
10	46074893	ALU_7934	rs35993099	Red blood cell count	2E-44	0.67	27863252

10	46074893	ALU_7934	rs76493570	Immature fraction of reticulocytes	4E-52	0.67	27863252
10	46074893	ALU_7934	rs901683	Red blood cell traits	2E-16	0.67	23222517
3	36839159	ALU_2289	rs75968099	Schizophrenia	1E-13	0.66	25056061
12	28417298	ALU_9107	rs10843164	Height	6E-12	0.66	23563607
2	198763462	ALU_1894	rs57862683	Morning person	1E-27	0.66	30696823
2	174952231	SVA_134	rs77345174	Blood protein levels	9E-15	0.66	30072576
17	44153977	SVA_706	rs7207400	Height	2E-30	0.66	30595370
17	44153977	SVA_706	rs7207400	Worry	2E-15	0.66	29942085
15	49609604	ALU_10695	rs10519227	Thyroid hormone levels	1E-11	0.66	23408906
3	157962934	ALU_2845	rs73030851	Blood protein levels	1E-300	0.66	30072576
15	49609604	ALU_10695	rs7167852	Lung function (FEV1/FVC)	4E-13	0.66	30595370
2	181880746	ALU_1805	rs4563182	Mean corpuscular hemoglobin	5E-11	0.66	30595370
2	181880746	ALU_1805	rs79719017	Risk-taking tendency (4-domain principal component model)	4E-08	0.66	30643258
6	53167475	ALU_5181	rs209489	Survival in colorectal cancer (distant metastatic)	8E-10	0.66	26586795
1	219558910	ALU_810	rs17525033	Lung function (FEV1/FVC)	2E-27	0.65	30595370
9	4237141	ALU_7235	rs736893	Glaucoma (primary angle closure)	1E-14	0.65	27064256
17	44153977	SVA_706	rs56214516	Medication use (antithrombotic agents)	1E-08	0.65	31015401
17	44153977	SVA_706	rs56214516	Feeling fed-up	1E-16	0.65	29500382
2	198763462	ALU_1894	rs1025549	Eczema	3E-19	0.65	30595370
2	198763462	ALU_1894	rs6738825	Crohn's disease	4E-09	0.65	21102463
9	117928281	ALU_7672	rs3833490	Blood protein levels	6E-224	0.65	29875488
10	46074893	ALU_7934	rs34285816	Red blood cell count	6E-49	0.65	30595370
10	46074893	ALU_7934	rs71494799	Mean corpuscular hemoglobin	9E-138	0.65	30595370
3	85576571	LINE1_629	rs1375561	Cognitive performance	2E-15	0.65	30038396
3	85576571	LINE1_629	rs73141547	Highest math class taken	1E-16	0.65	30038396
3	55788580	LINE1_590	rs6801405	Lung function (FEV1/FVC)	6E-12	0.65	30595370
1	194595960	ALU_707	rs13376197	Cerebrospinal fluid t-tau:AB1-42 ratio	3E-10	0.65	28641921
12	28438612	ALU_9108	rs1581630	Heel bone mineral density	7E-30	0.65	30048462
6	74504855	ALU_5280	rs10943130	Heel bone mineral density	1E-40	0.65	30048462
15	49609604	ALU_10695	rs17400427	Lung adenocarcinoma	6E-10	0.65	28604730
11	49282683	ALU_8572	rs113221947	Heel bone mineral density	2E-09	0.64	30595370
2	181880746	ALU_1805	rs10184839	Diastolic blood pressure	2E-13	0.64	30224653
10	106566893	ALU_8208	rs2864034	Self-reported math ability	2E-16	0.64	30038396
10	54466942	ALU_7961	rs12218358	Heel bone mineral density/Height	7E-58	0.64	30595370
7	33195329	ALU_5942	rs10232036	Height	5E-11	0.64	30595370
12	20473893	ALU_9052	rs11045171	HDL cholesterol	2E-18	0.64	30275531
10	11984965	ALU_7788	rs11819344	Mean corpuscular hemoglobin	4E-09	0.64	30595370
20	1546228	ALU_12014	rs3848788	Blood protein levels	1E-213	0.64	29875488
6	74504855	ALU_5280	rs6903575	Blood protein levels	7E-69	0.64	29875488
5	148096780	ALU_4747	rs58862611	Systolic blood pressure	4E-08	0.64	30595370
6	74504855	ALU_5280	rs10455097	Blood protein levels	3E-46	0.64	28240269
7	78146522	ALU_6127	rs62468583	Menarche (age at onset)	3E-08	0.64	30595370
8	13975433	ALU_6584	rs12675921	Intelligence	2E-08	0.64	29942086
8	100782579	ALU_6981	rs921313	Mean corpuscular hemoglobin	3E-12	0.64	30595370
6	163013855	ALU_5742	rs36007635	Body mass index	7E-14	0.64	30595370
2	98582157	ALU_1385	rs4851462	Diastolic blood pressure	6E-13	0.64	30224653
2	181880746	ALU_1805	rs10191559	Heel bone mineral density	9E-32	0.64	30595370
2	181880746	ALU_1805	rs10191559	Red blood cell count	4E-10	0.64	27863252
11	49282683	ALU_8572	rs61448762	Systolic blood pressure	4E-09	0.64	27841878

3	103419436	ALU_2577	rs6776198	Highest math class taken	1E-12	0.64	30038396
2	174952231	SVA_134	rs77998199	Reaction time	2E-10	0.64	29844566
6	74504855	ALU_5280	rs6909201	Blood protein levels	1E-115	0.63	30072576
10	3569025	ALU_7750	rs10795055	Waist-hip ratio	6E-10	0.63	30595370
9	4237141	ALU_7235	rs1570204	Intraocular pressure	7E-31	0.63	29617998
13	62588972	ALU_9847	rs17208030	Educational attainment (years of education)	4E-10	0.63	30595370
7	119259819	ALU_6325	rs73719951	Heel bone mineral density	7E-09	0.63	30595370
2	198763462	ALU_1894	rs10497813	Self-reported allergy	6E-10	0.63	23817569
11	55408899	ALU_8584	rs7104561	Medication use (agents acting on the renin-angiotensin system)	5E-08	0.63	31015401
4	145395308	ALU_3782	rs13142776	Post bronchodilator FEV1	2E-12	0.63	26634245
4	145395308	ALU_3782	rs1512283	Post bronchodilator FEV1	1E-10	0.63	26634245
4	145395308	ALU_3782	rs1960493	Post bronchodilator FEV1	1E-10	0.63	26634245
4	145395308	ALU_3782	rs34265962	Post bronchodilator FEV1	1E-12	0.63	26634245
4	145395308	ALU_3782	rs35937742	Post bronchodilator FEV1	1E-10	0.63	26634245
4	145395308	ALU_3782	rs7678427	Post bronchodilator FEV1	1E-12	0.63	26634245
4	145395308	ALU_3782	rs973796	Post bronchodilator FEV1	3E-12	0.63	26634245
1	198243300	ALU_726	rs1938376	Height	2E-08	0.63	30595370
4	145395308	ALU_3782	rs1512282	Post bronchodilator FEV1	4E-11	0.63	26634245
17	44153977	SVA_706	rs113322852	Neuroticism	3E-19	0.63	30643256
11	49282683	ALU_8572	rs7929717	Intraocular pressure	1E-15	0.63	29617998
17	44153977	SVA_706	rs183211	Ovarian cancer	2E-13	0.63	25581431
15	47507342	LINE1_2640	rs12442330	Neuroticism	2E-10	0.63	29942085
3	168885760	ALU_2909	rs9290361	Plateletcrit	1E-22	0.63	27863252
3	158089835	ALU_2846	rs1714510	Neuroticism	2E-08	0.62	29942085
9	33130564	SVA_401	rs10971420	IgG glycosylation patterns	2E-12	0.62	29535710
13	46647748	ALU_9748	rs1087	Thrombin-activatable fibrinolysis inhibitor levels	3E-29	0.62	29378355
10	54466942	ALU_7961	rs7088220	Heel bone mineral density	2E-54	0.62	30048462
11	49366217	ALU_8573	rs7929543	Type 2 diabetes	2E-09	0.62	30054458
10	106566893	ALU_8208	rs10400054	Highest math class taken	2E-11	0.62	30038396
10	106566893	ALU_8208	rs17118088	Highest math class taken	5E-18	0.62	30038396
11	65503489	ALU_8620	rs478304	Spherical equivalent or myopia (age of diagnosis)	1E-09	0.62	29808027
11	65503489	ALU_8620	rs478304	Acne (severe)	3E-11	0.62	24927181
17	44153977	SVA_706	rs56192752	Educational attainment (years of education)	9E-31	0.62	30038396
9	117928281	ALU_7672	rs35157100	Blood protein levels	3E-96	0.62	29875488
4	112628973	LINE1_928	rs9991259	Body mass index	1E-10	0.62	30595370
15	73983319	ALU_10812	rs8038465	Liver enzyme levels (gamma-glutamyl transferase)	1E-09	0.62	22001757
3	85576571	LINE1_629	rs55686445	Educational attainment	5E-12	0.62	27046643
3	85576571	LINE1_629	rs66568921	Educational attainment	6E-39	0.62	30038396
12	41847723	ALU_9169	rs1458156	Body mass index	3E-15	0.62	30595370
17	44153977	SVA_706	rs1879586	Serous invasive ovarian cancer	3E-12	0.62	28346442
17	44153977	SVA_706	rs1879586	Ovarian cancer	2E-19	0.62	28346442
17	44153977	SVA_706	rs916888	Lung function (FEV1)	4E-09	0.62	30061609
9	94058487	LINE1_1863	rs112679102	Balding type 1	1E-10	0.62	30595370
6	74504855	ALU_5280	rs9447004	Blood protein levels	2E-44	0.62	28240269
12	71525479	ALU_9320	rs7138300	Type 2 diabetes	6E-10	0.62	30054458
5	87399827	ALU_4449	rs55940342	Systolic blood pressure	1E-12	0.61	30595370
8	8920127	ALU_6560	rs4537305	Medication use (diuretics)	1E-14	0.61	31015401
12	28417298	ALU_9107	rs1511550	Height	3E-114	0.61	30595370

12	71525479	ALU_9320	rs11178649	Respiratory diseases	3E-09	0.61	30595370
12	71525479	ALU_9320	rs7955901	Type 2 diabetes	7E-09	0.61	22885922
6	139294734	ALU_5628	rs62441842	Height	3E-13	0.61	30595370
3	175673943	ALU_2951	rs66481714	Smoking initiation (ever regular vs never regular)	3E-10	0.61	30643251
9	17062597	ALU_7312	rs78817479	Heel bone mineral density	1E-12	0.61	30048462
12	26697612	ALU_9091	rs11613431	Educational attainment (years of education)	7E-09	0.61	30038396
5	87399827	ALU_4449	rs17286052	Blood pressure	7E-11	0.61	27841878
10	27929928	ALU_7862	rs1494204	Waist-hip ratio	4E-12	0.61	30595370
13	61734497	LINE1_2424	rs9563886	Insomnia	2E-08	0.61	30804565
3	36839159	ALU_2289	rs6550435	Bipolar disorder	2E-08	0.61	24618891
3	16796099	ALU_2173	rs7625399	Smoking cessation	9E-10	0.60	30643251
3	175673943	ALU_2951	rs9841807	Smoking initiation (ever regular vs never regular)	1E-08	0.60	30643251
3	36839159	ALU_2289	rs3732386	Schizophrenia	3E-11	0.60	28991256
2	141534074	ALU_1585	rs17515225	Motion sickness	3E-09	0.60	25628336

Appendix 3: RTE-TAS associations in the literature compared to the results of this study.

Table 23: List of RTEs in LD ($r^2 > 0.6$) with TAS identified by previous studies in the literature and how it compares to RTE-TAS associations identified by this study. Y = Yes. N = No

TAS	RTE type	r2	Study	PMID	Data source	Replicated?
rs2679073	ALU	0.80	Sudmant	26432246	Supplementary Table 10	N
rs2942168	SVA	0.84	Sudmant	26432246	Supplementary Table 10	Y
rs12185268	SVA	0.84	Sudmant	26432246	Supplementary Table 10	Y
rs1864325	SVA	0.84	Sudmant	26432246	Supplementary Table 10	Y
rs12373124	SVA	0.84	Sudmant	26432246	Supplementary Table 10	Y
rs1981997	SVA	0.83	Sudmant	26432246	Supplementary Table 10	Y
rs7534016	ALU	0.75	Wang	28824558	Supplementary Table S1	N
rs28588043	ALU	0.74	Wang	28824558	Supplementary Table S1	N
rs2820037	ALU	0.84	Wang	28824558	Supplementary Table S1	N
rs10189761	ALU	0.77	Wang	28824558	Supplementary Table S1	N
rs2681019	L1	0.75	Wang	28824558	Supplementary Table S1	N
rs10496262	ALU	0.72	Wang	28824558	Supplementary Table S1	N
rs2163349	ALU	0.77	Wang	28824558	Supplementary Table S1	N
rs7594648	L1	0.70	Wang	28824558	Supplementary Table S1	N
rs10865924	ALU	0.78	Wang	28824558	Supplementary Table S1	N
rs9841504	ALU	0.75	Wang	28824558	Supplementary Table S1	N
rs13077101	ALU	0.71	Wang	28824558	Supplementary Table S1	N
rs345013	ALU	0.82	Wang	28824558	Supplementary Table S1	N
rs7442317	ALU	0.75	Wang	28824558	Supplementary Table S1	N
rs10034228	L1	0.87	Wang	28824558	Supplementary Table S1	N
rs16886364	ALU	0.76	Wang	28824558	Supplementary Table S1	Y
rs4388249	ALU	0.80	Wang	28824558	Supplementary Table S1	Y
rs2523822	ALU	0.77	Wang	28824558	Supplementary Table S1	N
rs4530903	ALU	0.78	Wang	28824558	Supplementary Table S1	N
rs3077	SVA	0.81	Wang	28824558	Supplementary Table S1	N
rs10948222	ALU	0.75	Wang	28824558	Supplementary Table S1	Y
rs9357506	L1	0.76	Wang	28824558	Supplementary Table S1	N
rs11757063	ALU	0.72	Wang	28824558	Supplementary Table S1	N
rs12666612	ALU	0.79	Wang	28824558	Supplementary Table S1	N
rs1404697	ALU	0.82	Wang	28824558	Supplementary Table S1	N
rs16939046	ALU	0.72	Wang	28824558	Supplementary Table S1	N
rs11574914	SVA	0.70	Wang	28824558	Supplementary Table S1	Y
rs7028939	ALU	0.78	Wang	28824558	Supplementary Table S1	N
rs10768747	L1	0.88	Wang	28824558	Supplementary Table S1	N
rs11246602	ALU	0.70	Wang	28824558	Supplementary Table S1	N
rs12371778	ALU	0.81	Wang	28824558	Supplementary Table S1	N
rs11049611	ALU	0.71	Wang	28824558	Supplementary Table S1	Y
rs1979679	ALU	0.74	Wang	28824558	Supplementary Table S1	N
rs11575234	ALU	0.84	Wang	28824558	Supplementary Table S1	N
rs17788937	ALU	0.79	Wang	28824558	Supplementary Table S1	Y
rs8023445	ALU	0.78	Wang	28824558	Supplementary Table S1	N
rs1436958	ALU	0.71	Wang	28824558	Supplementary Table S1	N

rs12373124	SVA	0.85	Wang	28824558	Supplementary Table S1	Y
rs9303542	ALU	0.73	Wang	28824558	Supplementary Table S1	N
rs816535	ALU	0.81	Wang	28824558	Supplementary Table S1	N
rs12530	ALU	0.75	Wang	28824558	Supplementary Table S1	N
rs426736	ALU	0.77	Wang	28824558	Supplementary Table S1	N
rs426736	ALU	0.79	Wang	28824558	Supplementary Table S1	N
rs1585471	L1	0.79	Wang	28824558	Supplementary Table S1	N
rs2523822	ALU	0.84	Wang	28824558	Supplementary Table S1	N
rs225675	L1	0.71	Wang	28824558	Supplementary Table S1	N
rs12554999	L1	0.76	Wang	28824558	Supplementary Table S1	N
rs7947821	ALU	0.78	Wang	28824558	Supplementary Table S1	N
rs2250417	L1	0.72	Wang	28824558	Supplementary Table S1	N
rs1340490	ALU	0.70	Wang	28824558	Supplementary Table S1	N
rs6117615	ALU	0.80	Wang	28824558	Supplementary Table S1	N
rs816535	ALU	0.75	Wang	28824558	Supplementary Table S1	N
rs7534016	ALU	0.90	Hehir-Kwa	27708267	Supplementary Data 7	N
rs4085613	L1	0.93	Hehir-Kwa	27708267	Supplementary Data 7	N
rs4112788	L1	0.92	Hehir-Kwa	27708267	Supplementary Data 7	N
rs12185268	SVA	0.92	Hehir-Kwa	27708267	Supplementary Data 7	Y
rs12027542	ALU	0.96	Hehir-Kwa	27708267	Supplementary Data 7	N
rs2820037	ALU	0.91	Hehir-Kwa	27708267	Supplementary Data 7	N
rs2681019	L1	0.92	Hehir-Kwa	27708267	Supplementary Data 7	N
rs10496262	ALU	0.92	Hehir-Kwa	27708267	Supplementary Data 7	N
rs6741172	ALU	0.94	Hehir-Kwa	27708267	Supplementary Data 7	N
rs2163349	ALU	0.93	Hehir-Kwa	27708267	Supplementary Data 7	N
rs10865924	ALU	0.92	Hehir-Kwa	27708267	Supplementary Data 7	N
rs7442317	ALU	0.93	Hehir-Kwa	27708267	Supplementary Data 7	N
rs10034228	L1	0.95	Hehir-Kwa	27708267	Supplementary Data 7	N
rs1585471	L1	0.95	Hehir-Kwa	27708267	Supplementary Data 7	N
rs10053502	L1	0.95	Hehir-Kwa	27708267	Supplementary Data 7	Y
rs4388249	ALU	0.92	Hehir-Kwa	27708267	Supplementary Data 7	Y
rs2523822	ALU	0.93	Hehir-Kwa	27708267	Supplementary Data 7	N
rs1061235	ALU	0.93	Hehir-Kwa	27708267	Supplementary Data 7	N
rs12185268	SVA	0.94	Hehir-Kwa	27708267	Supplementary Data 7	N
rs10948222	ALU	0.92	Hehir-Kwa	27708267	Supplementary Data 7	Y
rs11757063	ALU	0.92	Hehir-Kwa	27708267	Supplementary Data 7	N
rs848353	ALU	0.94	Hehir-Kwa	27708267	Supplementary Data 7	N
rs1404697	ALU	0.96	Hehir-Kwa	27708267	Supplementary Data 7	N
rs12344488	ALU	0.90	Hehir-Kwa	27708267	Supplementary Data 7	N
rs12554999	L1	0.95	Hehir-Kwa	27708267	Supplementary Data 7	N
rs10771399	ALU	0.91	Hehir-Kwa	27708267	Supplementary Data 7	Y
rs1979679	ALU	0.92	Hehir-Kwa	27708267	Supplementary Data 7	N
rs1612141	L1	0.95	Hehir-Kwa	27708267	Supplementary Data 7	N
rs1436958	ALU	0.90	Hehir-Kwa	27708267	Supplementary Data 7	N
rs2679073	ALU	0.96	Hehir-Kwa	27708267	Supplementary Data 7	N
rs225212	ALU	0.92	Hehir-Kwa	27708267	Supplementary Data 7	N
rs12373124	SVA	0.93	Hehir-Kwa	27708267	Supplementary Data 7	Y
rs17577094	SVA	0.95	Hehir-Kwa	27708267	Supplementary Data 7	N
rs17649553	SVA	0.94	Hehir-Kwa	27708267	Supplementary Data 7	Y
rs1864325	SVA	0.94	Hehir-Kwa	27708267	Supplementary Data 7	Y

rs1981997	SVA	0.95	Hehir-Kwa	27708267	Supplementary Data 7	Y
rs2942168	SVA	0.94	Hehir-Kwa	27708267	Supplementary Data 7	Y
rs3077	SVA	0.94	Hehir-Kwa	27708267	Supplementary Data 7	N
rs3790672	SVA	0.91	Hehir-Kwa	27708267	Supplementary Data 7	N
rs393152	SVA	0.94	Hehir-Kwa	27708267	Supplementary Data 7	Y
rs8070723	SVA	0.95	Hehir-Kwa	27708267	Supplementary Data 7	Y
rs9303525	SVA	0.94	Hehir-Kwa	27708267	Supplementary Data 7	Y
rs12530	ALU	0.91	Hehir-Kwa	27708267	Supplementary Data 7	N
rs2300747	ALU	0.92	Payer	28465436	Dataset S3	N
rs1335532	ALU	1.00	Payer	28465436	Dataset S3	N
rs2779116	ALU	0.81	Payer	28465436	Dataset S3	N
rs857721	ALU	0.81	Payer	28465436	Dataset S3	N
rs857684	ALU	0.81	Payer	28465436	Dataset S3	N
rs426736	ALU	1.00	Payer	28465436	Dataset S3	N
rs1367228	ALU	0.77	Payer	28465436	Dataset S3	N
rs12463617	ALU	0.91	Payer	28465436	Dataset S3	Y
rs7561317	ALU	1.00	Payer	28465436	Dataset S3	Y
rs6548238	ALU	1.00	Payer	28465436	Dataset S3	Y
rs2867125	ALU	1.00	Payer	28465436	Dataset S3	Y
rs6711012	ALU	1.00	Payer	28465436	Dataset S3	Y
rs10189761	ALU	1.00	Payer	28465436	Dataset S3	Y
rs2903492	ALU	1.00	Payer	28465436	Dataset S3	Y
rs2667011	ALU	0.80	Payer	28465436	Dataset S3	N
rs6738825	ALU	0.92	Payer	28465436	Dataset S3	Y
rs16857609	ALU	0.95	Payer	28465436	Dataset S3	N
rs11177	ALU	0.91	Payer	28465436	Dataset S3	N
rs2251219	ALU	0.96	Payer	28465436	Dataset S3	N
rs4256159	ALU	0.93	Payer	28465436	Dataset S3	N
rs2712381	ALU	0.88	Payer	28465436	Dataset S3	N
rs2362965	ALU	1.00	Payer	28465436	Dataset S3	N
rs9877502	ALU	1.00	Payer	28465436	Dataset S3	N
rs2087160	ALU	0.72	Payer	28465436	Dataset S3	N
rs6825911	ALU	0.84	Payer	28465436	Dataset S3	N
rs10034228	ALU	0.96	Payer	28465436	Dataset S3	N
rs11748327	ALU	0.82	Payer	28465436	Dataset S3	N
rs9472155	ALU	0.95	Payer	28465436	Dataset S3	N
rs441460	ALU	0.96	Payer	28465436	Dataset S3	N
rs204247	ALU	1.00	Payer	28465436	Dataset S3	N
rs11759769	ALU	0.89	Payer	28465436	Dataset S3	Y
rs7809799	ALU	1.00	Payer	28465436	Dataset S3	N
rs4609139	ALU	0.85	Payer	28465436	Dataset S3	N
rs2293889	ALU	0.80	Payer	28465436	Dataset S3	N
rs13281615	ALU	0.77	Payer	28465436	Dataset S3	N
rs16901979	ALU	1.00	Payer	28465436	Dataset S3	N
rs10505483	ALU	1.00	Payer	28465436	Dataset S3	N
rs6983561	ALU	1.00	Payer	28465436	Dataset S3	N
rs10512248	ALU	0.96	Payer	28465436	Dataset S3	N

rs399593	ALU	1.00	Payer	28465436	Dataset S3	N
rs7089424	ALU	0.83	Payer	28465436	Dataset S3	N
rs10821936	ALU	0.83	Payer	28465436	Dataset S3	N
rs2638953	ALU	0.89	Payer	28465436	Dataset S3	N
rs10771399	ALU	0.92	Payer	28465436	Dataset S3	Y
rs10843164	ALU	0.96	Payer	28465436	Dataset S3	Y
rs2066808	ALU	1.00	Payer	28465436	Dataset S3	Y
rs2066807	ALU	1.00	Payer	28465436	Dataset S3	Y
rs17788937	ALU	0.83	Payer	28465436	Dataset S3	Y
rs975739	ALU	0.87	Payer	28465436	Dataset S3	N
rs4900384	ALU	0.78	Payer	28465436	Dataset S3	N
rs1456988	ALU	0.82	Payer	28465436	Dataset S3	N
rs10519227	ALU	0.76	Payer	28465436	Dataset S3	Y
rs7178424	ALU	0.78	Payer	28465436	Dataset S3	N
rs8038465	ALU	0.83	Payer	28465436	Dataset S3	Y
rs10852344	ALU	0.75	Payer	28465436	Dataset S3	N
rs3729639	ALU	1.00	Payer	28465436	Dataset S3	N
rs4351	ALU	0.89	Payer	28465436	Dataset S3	N
rs2665838	ALU	0.92	Payer	28465436	Dataset S3	N
rs2941551	ALU	0.92	Payer	28465436	Dataset S3	N
rs11658329	ALU	0.96	Payer	28465436	Dataset S3	N
rs4343	ALU	1.00	Payer	28465436	Dataset S3	N
rs4329	ALU	1.00	Payer	28465436	Dataset S3	N
rs9894429	ALU	0.81	Payer	28465436	Dataset S3	N
rs6015450	ALU	1.00	Payer	28465436	Dataset S3	N