TRANSPOSABLE ELEMENT POLYMORPHISMS AND HUMAN GENOME REGULATION

A Dissertation
Presented to
The Academic Faculty

by

Lu Wang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biological Sciences

Georgia Institute of Technology
December 2017

TRANSPOSABLE ELEMENT POLYMORPHISMS AND HUMAN GENOME REGULATION

Approved by:

Dr. I. King Jordan, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. John F. McDonald
School of Biological Sciences
Georgia Institute of Technology

Dr. Fredrik O. Vannberg
School of Biological Sciences
Georgia Institute of Technology

Dr. Victoria V. Lunyak
Aelan Cell Technologies
San Francisco, CA

Dr. Greg G. Gibson
School of Biological Sciences
Georgia Institute of Technology

Date Approved: November 6, 2017

To my family and friends

ACKNOWLEDGEMENTS

I am truly grateful to my advisor Dr. I. King Jordan for his guidance and support throughout my time working with him as a graduate student. I am fortunate enough to have him as my mentor, starting from very basic, well-defined research tasks, and guided me step-by-step into the exciting world of scientific research. Throughout my PhD training, I have been always impressed by his ability to explain complex ideas – sometimes brilliant ideas of his own – in short and succinct sentences in such a way that his students could easily understand. I am also very impressed and inspired by his diligence and passion for his work.

It is my great honor to have Dr. Greg Gibson, Dr. Victoria Lunyak, Dr. John McDonald, Dr. Fredrik Vannberg as my committee members. I really appreciate the guidance they provided me throughout my PhD study and the insightful thoughts they generously share with me during our discussions. Dr. Victoria Lunyak is also a long-time collaborator of the Jordan lab. I have always been impressed and inspired by her brilliant ideas for scientific research and great passion for science.

I am very grateful to my friends and colleagues from the Jordan lab, Dr. Andrew Conley, Dr. Ashan Huda, Dr. Jianrong Wang, Dr. Daudi Jjingo, Dr. Lee Katz, Eishita Tyagi, Dr. Lavanya Rishishwar, Emily Norris, Ying Sha, Evan Clayton, Aroon Chande, Junke Wang. I also appreciate the helpful discussions with Dr. Urko Martinez Marigorta and Biao Zeng from the department.

I am also grateful to have my friends from within and outside Georgia Tech to be here with me in Atlanta, exploring graduate school and life in Atlanta piece by piece. I feel very lucky to have spent time with you during school days and during weekends. Thank you for making my adventures in Atlanta so exciting and colorful.

Last but not least, I feel truly grateful to my family for all their love, understanding and support for my research and career.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

DNA-seq	DNA sequencing
eQTL	Expression Quantitative Trait Loci
ERV	Endogenous Retrovirus
GEUVADIS	Genetic European Variation in Health and Disease Project
GWAS	Genome-wide Association Studies
LD	Linkage Disequilibrium
LINE	Long Interspersed Element
LTR	Long Terminal Repeat
polyTE	Polymorphic Transposable Element
RNA-seq	RNA sequencing
SINE	Short Interspersed Element
SNP	Single Nucleotide Polymorphism
TE	Transposable Element
TF	Transcription Factor
1KGP	1000 Genomes Project

SUMMARY

A large proportion of the human genome, over 60% by estimation, is derived from transposable element (TE) sequences. Majority of these TE sequences in the human genome are retrotransposons – a type of TEs that replicates and inserts in the host genome via reverse transcription of RNA intermediates [1, 2]. TEs are known to contribute to the regulation of the human genome. Despite the fact that the majority of known TE-derived regulatory sequences correspond to relatively ancient insertions, which are fixed across human populations, there are several active families of retrotransposons, including the Alu [3, 4], LINE-1 (L1) [5, 6], and SVA [7, 8] retrotransposons that are capable of mobilizing via reverse transcription of RNA intermediates . Germline transposition of these elements generates polymorphisms between individuals, and somatic transposition generates cellular heterogeneity.

Given the disruptive potential of TE insertions, along with the regulatory potential of TE-derived sequences, one may expect a complex interplay between TE insertion polymorphisms and inter-individual differences in the expression of human genes, as well as TE activities and the host regulatory mechanisms. In the past, the host genome regulation on TE activities, as well as the impact of TE on host gene expressions, have been studied mostly in cell lines and model organisms. Due to the nature of the experimental approach, previous studies only evaluated a limited number genes that regulates TE activities. For the TE sequences that may potentially regulates host gene expression, majority of the previous studies have been focusing on fixed TE elements, *i.e.* elements that are no longer capable of transposition, and evaluated their impact on the human genome. Other studies that have

focused on polymorphic TE insertions were limited the investigation within model organisms or human cell lines.

Recently, with the growing number of whole genome sequences of healthy human individuals available at the population scale, it is now possible for the first time to systematically screen for modifiers of TE activities, as well as, evaluate the impact of TE on host genome regulation at the genome-wide scale. My dissertation is focusing on the role of polymorphic TEs in the regulation of the genes and how the human genome regulates TE activities. Specifically, I used the genome-wide association approach to evaluate how the host genome regulates TE activities, and the extent to which recent TE activity could lead to regulatory polymorphisms among populations.

Research Advance 1: Genome-wide association screens were performed in search for modifiers of human L1 activities. With integrated genotype profiles and gene expression profiles of matched individuals, the association analysis was performed to relate the SNP genotype profiles and gene, L1 transcript abundance levels. Full-length, intact L1 expression levels were quantified with covariates adjusted to control for confounding variables. The expression quantitative trait loci (eQTL) analysis was applied for the discovery of individual single nucleotide polymorphisms (SNPs) that are jointly associated with both L1 and gene expression. Putative L1 modifier genes were identified including 35% of 34 known L1 modifier genes, 24% putative transcription factor genes and 30% chromatin modifier genes.

Research Advance 2: Genome-wide association screens were performed to evaluate the impact of polyTE on human genome regulation. The locus-specific polyTE

insertion genotypes were related to B cell gene expression levels among 445 individuals from 5 human populations. My results showed that numerous human polyTE insertion sites correspond to both *cis* and *trans* expression quantitative trait loci (eQTL) with genes that are directly related to cell type-specific function in the immune system. A polyTE insertion loci was found to be associated with the expression level of a cell type specific transcription factor *PAX5* and its downstream target genes. The genome-wide significant associations indicate that human TE genetic variation can have important phenotypic consequences. Our results also suggested that TE-eQTL may be involved in transcriptional network rewiring and population-specific gene regulation.

Research Advance 3: Integrated genome-wide associations and other evidences was performed to evaluate the impact of polyTE on human disease and health outcomes. TE insertion polymorphisms were related to common health and disease phenotypes that have been previously interrogated through genome-wide association studies (GWAS) based on the linkage disequilibrium (LD) structure of the population. eQTL analysis was performed on the GWAS SNP-linked and enhancer co-localized polyTEs to identify polyTE insertions that are linked to gene expression changes in B-cells. Two polyTE loci that are co-located with cell type-specific enhancers were linked to common diseases phenotypes. Taken together, my results showed that polyTE could impact human health and disease phenotypes by causing changes in gene expression.

Research Advance 4: A comprehensive survey of recent research progress on evaluation of the impact of polyTE on human genome regulation. Genome-enabled approaches, including developed bioinformatics tools and high-throughput experimental approaches, for characterizing TE insertion variants were developed in recent years.

Several studies have applied these approaches to characterize polyTE insertion profiles at population level. The eQTL approaches, along analyses of linkage disequilibrium structures in the population have uncovered the connection between TE insertion polymorphisms with previously characterized genome-wide association study (GWAS) trait variants of common complex diseases. In summary, these genome-enabled population scale analysis approach shows great potential for evaluating the contribution of recent and ongoing TE activity to the variations among human individuals.

CHAPTER 1. INTRODUCTION

1.1 Transposable elements (TEs) – definitions and concepts

Transposable elements (TEs) are DNA sequences that capable of moving themselves to a new location in the host genome. There are two distinct mechanisms through which TEs can transpose [9]. One is the so called “copy-and-paste” mechanism, where the TE sequences in the host genome were first transcribed into RNA and then reverse transcribed to DNA and insert back to a new genomic location in the host genome. This type of TEs are called *retrotransposons*. Another way through which the TE sequences jumps in the host genome is the “cut-and-paste” mechanism. The type of TEs are transposes through this mechanism is called *DNA transposons*.

1.1.1 Retrotransposons

In order to facilitate their transposition in the host genome, some retrotransposons encode their own proteins that helps them transpose. This type of retrotransposons are known as the *autonomous retrotransposons*. Both the long terminal repeat (LTR) retrotransposons and the non-LTR encodes their own proteins, and the latter ones encode proteins with a poly A tail [10-12]. The retrotransposons that does not encode their own proteins are called *non-autonomous retrotransposons*. Not surprisingly, these non-autonomous retrotransposons that do not encode any proteins lack the molecular machinery that are required for transposition. In fact, the non-autonomous retrotransposons are found to rely on the reverse transcriptase and endonuclease that encoded by autonomous elements

to transpose [13, 14]. A successful transpositional event of retrotransposons will create at least one new copy of themselves in the host genome. Therefore, transposable element activities can be measured by their copy numbers in the host genomes. In addition to the direct observation of the copy numbers of retrotransposon in the host genome, the abundance of gene transcripts and protein products that encoded by the autonomous retrotransposons become a key measurement of the activities of TE activities.

1.1.2 TE compositions in the host genomes

TEs were initially discovered in maize by Barbara McClintock in the late 1940s [15]. As more genomes have been studied, TEs sequences were found to have broad existence in the genomes of almost all living organisms. Different host genomes usually have quite distinct profiles of TEs - both in terms of the types of TEs that exist in the host genome and the proportion that the host genome is comprised of TE sequences. In fact, in the maize genome where TEs were initially discovered, TEs comprised about 60% of its genome, whereas the proportion of TE sequences in other species, such as the yeast genome can be as low as 3% [16].

The human genome project was the first systematic characterization of transposable elements throughout the entire human genome [1]. By estimation, there are more than 50% of the human genome sequence that is derived from TE insertions, containing both retrotransposons and DNA transposons [1, 2]. Retrotransposons, specifically the long interspersed elements (LINEs), short interspersed elements (SINEs) and long terminal repeat (LTR) retrotransposons, are the major types of TEs in the human genome. LINEs

are the largest class of TEs in the human genome measured by bases, taking up over 20% of the genome. SINE is the largest class of TEs in the human genome measured by copy number, with over 1.5 million copies in the human genome. LINEs and SINEs takes up a total of 34% of the human genome. The remaining TEs in the human genome are ~0.5 million LTR, including endogenous retroviruses (ERVs) and ~0.3 million DNA transposons. Among all the LINEs, the LINE-1 (L1) is the most abundant class in the human genome, with over half million copies and 462.1 megabases (Mb) in total lengths [1]. Among all the SINEs in the human genome, Alu is youngest and the the largest subclass of SINEs [3, 4].

1.2 Functional Roles of TEs

Followed by the initial discovery of TEs, the “selfish DNA” hypothesis seemed to hold when little was known about any function that these widely spread repetitive sequences encodes and how they could be involved in various of basic biological processes that are necessary for the survival of their host. As genome parasites, it seems plausible that the only purpose of TEs in the host genome is simply to propagate themselves and colonize the host genome. However, accumulating evidences have been found to show that TE sequences have been widely recruited and integrated in the host genome regulatory machinery, providing functional regulatory elements [17]. In fact, TEs provide an abundant source of regulatory sequences in the host genomes [17, 18].

1.2.1 TE-derived cis-regulatory elements

TEs were originally studied for their impact on the expression of host gene due to the destructive effects that the new TE insertions have on the host genes or regulatory elements. Nevertheless, TEs actually provide an abundant source of cis-regulatory elements for the host genomes [17], including promoters [19-21], enhancers [22-26]. It has been shown that TE sequences were present in ~25% of experimentally validated human promoter [20]. Over 50 thousand ERV-derived sequences were found to initiate transcription in the human genome[19]. These observation is coherent with our knowledge about TEs from an evolutionary perspective. In order to survive and proliferate in the host genome, TEs carry cis regulatory sequences that mimic the host promoters. In addition to promoter sequences, TE can also function as enhancers and alternatively spliced exons and regulate host gene expression levels [22-27].

1.2.2 Trans-regulatory activities of TEs

In addition to its cis-regulatory activities, TEs were also found to contribute a number of trans-regulatory elements such as transcription terminators [28], small RNAs [29-31], chromatin boundary elements [32] and involved in regulatory networks [33]. There are 55 experimentally validated human microRNA (miRNA) genes were found to be derived from TEs. These TE-derived sequences could impact thousands of human genes that are regulated by these miRNAs [30]. A subclass of SINEs, the mammalian-wide interspersed repeats (MIRs) in the human genome were found to provide insulators that could organize the chromatin in different regulatory domains through enhancer-blocking and separate

active and repressive chromatin domains through chromatin barrier activity [32]. A set of closely related ERV sequences were found to have significant impact on human tumor suppressor gene p53 regulatory network by providing a near-perfect binding site for its target genes [33]. Therefore, TE-derived sequences can have a directly impact on host gene expression through regulating chromatin states, as well as have an indirect impact on host gene expressions via regulatory network rewiring.

1.3 Human genome regulation of TE activities

Human TE activities were initially discovered by their mutagenic effects. The repetitive sequences in TEs may introduce genome instabilities such as alterations and inversions in the host genome. TE insertions may also occur in genes or regulatory regions such that they disrupt function encoded by the gene and lead to deleterious effects to the host. In fact, there are many diseases that are found to be directly caused by TE insertions. For example, there are 96 genetic diseases that have been demonstrated to be caused by retrotransposon insertions, including cystic fibrosis (Alu), hemophilia A (L1) and X-linked dystonia-parkinsonism (SVA) [34, 35]. In addition, somatic mutations caused by human TE activity have been linked to a number of different kinds of cancer [36, 37]. Therefore, it is not surprising that the host genomes have evolved to have different suppression mechanisms to suppress TE activities. Specifically, for retrotransposons the host suppressions target both the transcriptional processes and post-transcriptional processes.

1.3.1 Transcriptional suppression of TE activities

Transcriptional suppression of TE activities mainly relies on the epigenetic silencing of TE sequences through different chromatin modifications, including DNA methylation, histone modification and chromatin remodeling [38]. TE activations have been observed in mice with deficiencies in their DNA methyltransferases (DNMTs) [39]. The cytosine methyltransferases are enzymes that transfer methyl groups to cytosine nucleotides of genomic DNA and maintain the methylation. Loss of function in these enzymes will cause the loss of methylation at the CpG island of TE promoter regions and thus lead to the derepression of TEs. In humans, hypomethylation of L1 DNA has been shown to be associated with elevated L1 transcriptional activities [40-42]. Histone modifications have also been shown to silence TEs in mice and preimplantation embryos [43]. Moreover, loss of function in genes involve in repressive histone modifications also lead to activation of TEs. Mutations in a methyltransferase gene for histone H3 lysine 9 (H3K9) methylation in mice result in the overexpression of TEs [44]. In humans, analysis on global histone modifications have found that H3K9 is enriched at human retrotransposons, suggesting the repression role of histone methylation in TE repression [44-46]. Proteins that involve in alteration of chromatin structures, such as condensation and packing, could also be critical for TE silencing [47]. For example, condensin II subunit, which participates in maintaining the structures of chromosomes, has been shown to repress retrotransposition in Drosophila. Nevertheless, epigenetic silencing is not the only mechanism through which the host cells suppress retrotranspositions. Other mechanisms at the post-transcription level act as a suppression to TE activities.

1.3.2 Post-transcriptional suppression of TE activities

In addition to transcriptional suppressions, there are also different mechanisms that suppresses TE activities at the post-transcriptional level. The RNAi pathway also plays an important role in the post-transcriptional silencing of TEs [48]. In the RNAi pathway, the dicer proteins cleave the dsRNAs of the TE transcripts and yield small interfering RNAs (siRNAs) [48]. The siRNAs and the RNA-degrading complexes forms the siRNA-guided transcript-cleavage complex RISC and the degrade TE transcripts that are complementary to the guide siRNA. It has been shown that many human siRNAs were generated from TEs and antisense transcripts-mediated RNAi regulate retrotranspositional activities [49]. In addition to RNAi, other mechanisms has also been shown to suppress retrotransposition It has been shown that in human cells, ribonucleoprotein (hnRNPL) interact with L1 RNA and down-regulates L1 retrotransposition [50]. The melatonin receptor 1 (MT1) was found to inhibit retrotranspositional activities of L1 through downregulation of the transcripts [51]. For L1s, the ORF1 protein (ORF1p) and ORF2 protein (ORF2p) binds the L1 mRNAs and forms a ribonucleoprotein (RNP). Therefore, addition to degrading RNA transcripts of L1s, some host mechanisms also targets the RNP formation and delivery to genomic DNA processes in order to inhibit retrotransopositions [52]. The innate immune system has also been found to play a role in post-transcriptional TE suppression. The autophagy signaling pathway has been shown to prevent new TE insertions by degrading RNA intermediate of TEs [53].

In summary, existing evidences show that the host genome regulates TE activities through a variety of mechanisms. Different mechanisms interact or interfere with each other and other host regulatory machineries in a complex way. Several genes were

demonstrated to be essential for the suppressing or silencing of TE activities in mammalian cells [22, 23]. However, genome-wide screens for such suppressor genes were only performed in a few organisms such as *S. cerevisiae* [6] and *C. elegans* [7]. Moreover, most of the current human-specific analysis on TE suppressors were performed under disease conditions such as cancers, which may not represent the full picture of the TE regulations by human genomes. Therefore, the question of how TE activities are regulated in human cells has yet to be systematically addressed at the genome-wide scale.

1.4 Active TEs in the human genome

While most of the TE sequences in the human genome are remnants of ancient insertional events. In other words, a large proportion of the TE sequences in the human genome are no longer capable of transposition. There are three families of retrotransposons that are currently active in the human genome: Alu [3, 4], L1 [5, 6] and SVA [7, 8]. Recent studies have also found a small number of human endogenous retrovirus K (HERV-K) elements are also active in the human genome [54].

A full-length L1 is ~6 kilobases (kb) in length and encodes its own RNA polymerase II promoter, and two proteins – the ORF1 and ORF2 proteins [55, 56]. The ORF1 proteins have the RNA-binding function and the ORF2 proteins can function as both the endonuclease and reverse-transcriptase [56]. These molecular functions collectively facilitate the transpositions, and thus make L1 the dominant autonomous retrotransposons in the human genome. In addition to its own transposition, the machinery that encoded by L1s also promotes the reverse transcriptional activities in the human genome, including the

transposition of other non-autonomous retrotransposons such as the Alu elements. While most of the L1 sequences accumulates mutations over time and gradually lost their mobility, there are ~145 full-length, intact L1s in the human genome [5, 57]. These L1s have their have their intact ORF1 and ORF2 sequences that encodes a functional protein and thus they have their full capacity to transpose. As a result, for these full-length, intact L1 elements, it is possible to measure their transcription levels with a relatively high confidence via either a targeted assay approach or whole transcriptome approach.

As a class of non-autonomous retrotransposons, SINEs do not encode any protein. Of the ~1.5 million copies of SINEs in the human genome, ~1.1 million copies of them are Alu, which is the most abundant [1]. Among all the SINEs in the human genome, Alu is youngest and the only active element. An Alu element is usually ~300 base pairs (bp) long and harbors their own RNA polymerase III promoter sequence [58]. Since Alu sequences do not carry any RNA polymerase III termination signal, therefore, once Alu sequences get transcribed, the transcription will carry through until a termination signal was reached [59]. As a result, most observed Alu transcriptions via the whole transcriptome approach would be read-through transcripts.

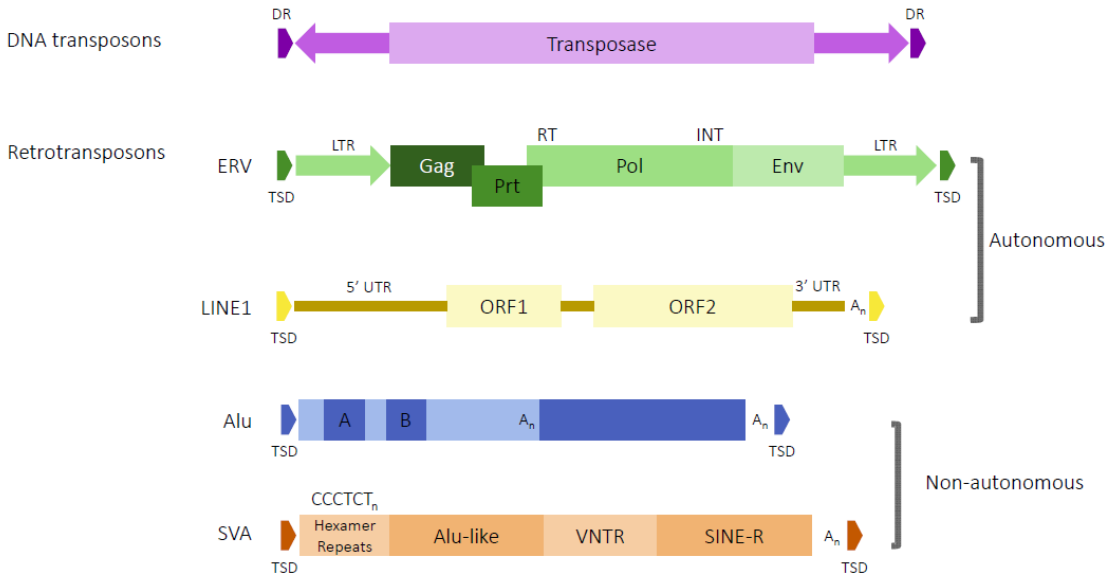


Figure 1 Schematic showing the structure of different TEs in the human genome

The SINE-VNTR-Alus (SVAs) are hominoid specific retrotransposons that are currently active in the human genome. SVA, as its name SINE-VNTR-Alu indicates, is a chimeric retrotransposon that is comprised of SINE, VNTR and Alu sequences. SVAs usually have a full length of ~2 kb long [7, 8]. Like Alu elements, they are also non-autonomous retrotransposons and therefore, its transposition is also likely to rely on the L1 machinery. Unlike Alu sequences that carries an internal RNA polymerase III promoter, SVAs do not encode any internal promoter sequence. It has been shown that SVAs can be by RNA polymerase II and it is likely to rely on the promoter sequences from its flanking regions [7, 8]. In other words, the observed transcription of SVA sequences is likely to be initiated in its flanking regions.

1.5 Human TE Polymorphisms and Genome Regulation

The retrotransposition rate (RR) of active human TEs have been estimated be about one insertion every 10 to 100 births [60]. When members of the active TE families jump in the somatic cells, they generate cellular heterogeneity. Similarly, when they jump in germline cells, they create inter-individual differences at the TE insertional sites. These sites segregate human individuals within and between populations through TE insertional polymorphisms.

The initial studies on TE polymorphisms were started in the 1990s where a small number of sites in the human genome with very recent insertions were studied. The Alu elements, namely Ya5/8 and Yb8 Alu elements were found to be present in some human populations while absent in others [3, 61, 62]. Four Alu insertion polymorphisms were identified for the first time in 16 populations worldwide [62]. Subsequent study characterized 8 polymorphic Alu insertions in 1,500 individuals from 34 populations worldwide [63]. In the past decade, new experimental and computational approaches have been developed to systematically characterize structural variations, including TE polymorphisms in the entire human genome [64]. The 1000 genomes project (1KGP) has applied multiple approaches to characterize TE polymorphisms in 2,504 individuals from 26 populations worldwide [65]. Given the disruptive nature of TE insertions, along with the demonstrated impact of TE-derived sequences to on the regulation of the human genome, TE insertion polymorphisms opened a new opportunity for us to systematically study the impact of TEs on inter-individual differences in the regulation of gene expression. For example, some human TE polymorphisms may lead to differences in gene expression patterns between individuals. Therefore, inter-individual differences in regulatory

variations generated by these recent TE insertions may have important implications for health and disease.

CHAPTER 2. GENOME-WIDE SCREEN FOR MODIFIERS OF HUMAN LINE-1 EXPRESSION

2.1 Abstract

LINE-1 (L1) is an active family of transposable elements (TEs), which exerts a major impact on the structural integrity of the human genome. L1 transpositional activity is highly disruptive, and L1-mediated insertions have been linked to more than one hundred human diseases. Accordingly, human cells employ a wide variety of mechanisms to mitigate the impact of L1 activity by suppressing their expression. Here, we report the development of a novel genome-wide screen for genetic modifiers of human L1 expression. Our approach relies on expression quantitative trait loci (eQTL) analysis for the discovery of individual genetic variations, *i.e.* single nucleotide polymorphisms (SNPs), which are jointly associated with both L1 and gene expression. We applied this approach to known modifiers of L1 expression (n=34), as a positive control, along with sets of transcription factor (n=1,244) and chromatin associated (n=450) protein encoding genes. The joint L1-gene eQTL association analysis was able to recover 35% of 34 known L1 modifier genes, including the DNA methyltransferase gene *DNMT3A* and the RNA helicase gene *MOV10*. We also discovered 24% putative transcription regulator genes and 30% chromatin modifier genes, which that can be considered as putative L1 modifiers based on the results of our screen. Notable transcription regulator genes uncovered by our screen include the *TAF13* gene, which encodes part of the transcription factor IID complex, and the *TCF19* cell-type specific transcription factor gene. With respect to putative chromatin modifiers of L1, a number of histone deacetylase genes, including *HDAC1*, *HDAC3*, and *HDAC5*, as

well as the histone H3 lysine demethylase gene *KDM5A*, were detected via the joint L1-gene eQTL analysis. The list of putative L1 modifiers detected by our genome-wide screen can be considered as a set of working hypotheses regarding human regulation of L1 activity; definitive proof for the role of these genes in the regulation of L1 expression will require further experimental interrogation. Nevertheless, our results underscore the potential utility of genome-scale approaches to the analysis of host-TE interactions, particularly in terms of the ability to substantially narrow down the search space for candidate genes of interest.

2.2 Introduction

Transposable element (TE) derived sequences make up more than half of the human genome, and members of a single TE family alone, LINE-1 (L1), comprise 17% of the genome sequence [1]. L1s are retrotransposons, which transpose via the reverse transcription of an RNA intermediate. Full-length L1 elements are ~6 kilobases (kb) in length and encode an RNA polymerase II promoter along with the ORF1 and ORF2 proteins that catalyze reverse transcription [55, 56]. L1s are the only family of autonomous TEs that remain active in the human genome [5], and the L1 transcriptional machinery is also responsible to catalyzing the transposition of the non-autonomous Alu and SVA TE families [58, 66]. Accordingly, L1 activity has a major impact on the structure of the human genome.

Ongoing L1 activity poses a direct threat to human genome stability. In fact, L1 transpositional activity was discovered by virtue of a Hemophilia A causing insertion in

the Coagulation Factor VIII gene (*F8*) [6]. To date, 124 independent TE insertions resulting from L1 activity, including L1-mediated Alu and SVA insertions, have been linked to human disease [67]. For example, germline L1 insertions are causal mutations for chronic granulomatous disease, Duchenne muscular dystrophy, and Hemophilia B [68-70]. Somatic L1 insertions have been implicated in a variety of cancer types [71], including colorectal cancer [72], head and neck cancers [73, 74], and lung carcinomas [75]. L1s can also disrupt genome stability by facilitating non-homologous recombination and chromosome breakage [66, 76].

Host genomes' TE regulatory machinery are a crucial component of genome stability. Human cells employ a wide variety of mechanisms to mitigate the impact of L1 insertions by tightly regulating their activity. As transcription is rate-limiting step for the activity of retrotransposons, such as L1, host genomes keep these elements in check by regulating their expression at the pre- and post-transcriptional levels [76-78]. Different cell types can employ distinct cellular machineries to suppress L1 activities.

In germline cells, the PIWI-interacting RNA (piRNA) pathway inhibits L1 activity via DNA methylation genomic L1 genomic sequences and RNA degradation of L1 transcripts. This pathway involves the methylation regulator *DMNT3L* and the piRNA bidnding protein *PIWIL4* [79, 80]. Recent evidence has shown that piRNA pathway may be active in tissues other than the germline [81, 82]. In embryonic stem cells (ESCs), three DNA methyltransferases genes, *DNMT1*, *DNMT3A* and *DNMT3B* coordinately maintain the repressive L1 methylation [83]. The Sirtuin 6 (*SIRT6*) protein has been shown to repress L1s in mice embryonic fibroblast cells by packaging the L1 sequences into heterochromatin in collaboration with the tripartite motif containing 28 (*TRIM28*) gene

(also known as *KAP1*) [84]. In neural stem cells, the transcription factor Sox2 (*SOX2*) and the histone deacetylase 1 protein (HDAC1) form a complex on the L1 5' promoter region to represses element transcription [85, 86].

Post-transcriptional suppression mechanisms usually target L1 mRNA or the formation and transportation of the L1 ribonucleoprotein (RNP) complexes. For example, dicer proteins of the RNA interference (RNAi) pathway cleave dsRNAs from L1 transcripts to yield small interfering RNAs (siRNAs). The siRNA-guided RNA-induced silencing complex (RISC) then identifies and the degrades L1 transcripts complementary to guide siRNA sequences [48]. The ribonucleoprotein hnRNPL has been shown to interact with L1 mRNA directly and inhibits retrotransposition by decreasing steady-state levels of L1 transcripts [50]. The Mov10 RNA helicase (MOV10), which is a component of the RNA-induced silencing complex (RISC), physically associates with the L1 RNP and represses L1 retrotransposition by promoting stress granule formation [52, 87, 88]. The endonucleases TREX1 and ERCC1 inhibit L1 retrotransposition post-transcriptionally by cleaving the reverse-transcribed L1 cDNA molecules [89, 90].

The experimental approaches used to characterize the modifiers of L1 expression described above traditionally rely on candidate gene approaches, which require *a priori* knowledge regarding host L1 activity related pathways and their constituent genes (proteins). Candidate L1 modifier genes are experimentally manipulated, *e.g.* via knock-out or over-expression techniques, and the impact on L1 activity is then observed. For example, the role of the piRNA pathway in L1 repression was investigated by insertional mutation of the *PIWIL4* gene followed by the use of *in situ* hybridization to compare L1 transcript levels in wild-type versus mutant cells [80]. Similarly, for studies of the *SIRT6*

L1 suppressor gene, an L1 green fluorescent protein (L1-EGFP) [91] reporter system was used to measure *de novo* retrotransposition events in wild type versus SIRT6 knockout cells [84]. While approaches of this kind have been extremely valuable in characterizing the host L1 regulatory machinery, they are typically limited to a relatively small set of candidate genes (proteins).

For this study, we developed and applied a genome-wide screen for putative genetic modifiers of L1 activity. Our genome-wide screen uses the expression quantitative trait loci (eQTL) approach to look for genetic variants, *i.e.* single nucleotide polymorphisms (SNPs), which are associated with both L1 and gene expression levels (Figure 1). We considered SNPs that are jointly associated with L1 and gene expression to be putative genetic modifiers of L1 expression, with the paired genes implicated as members of L1 regulatory pathways. This approach can be considered to be largely hypothesis free, and therefore less biased, compared to more traditional candidate gene approaches, and it is also distinguished by its genome-wide scale, with respect to the ability to screen thousands of genes at a time.

We applied our eQTL approach to known L1 modifier genes, as a positive control, along with potential L1 modifier genes encoding transcription factors and chromatin associated proteins. In addition to recovering known L1 modifier genes, such as *DNMT3A* and *MOV10*, our screen also identified a number of novel L1 transcriptional regulator genes (*TAF13* and *TCF19*) as well as chromatin associated proteins implicated in L1 regulation (*HDAC5* and *KDM5A*).

2.3 Materials and Methods

2.3.1 Genome-wide SNP genotypes

Genome-wide SNP genotype calls for 358 individuals from four European populations were accessed from the phase 3 variant release of the 1000 Genomes Project (1KGP) [92], corresponding to the human genome reference sequence build GRCh37/hg19. The four populations are CEU: Utah Residents (CEPH) with Northern and Western Ancestry, FIN: Finnish in Finland, GBR: British in England and Scotland and TSI: Toscani in Italy. As previously described for the 1KGP [92], whole genome sequencing (DNA-seq) was performed for lymphoblastoid cell lines, *i.e.* Epstein–Barr virus (EBV) transformed B-lymphocytes (B cells), from these individuals and used to call genetic variants. The SNP genotype data were accessed from the 1000 Genomes Project ftp server maintained at the NCBI:

http://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/release/20130502/.

2.3.2 Gene expression quantification and normalization

Matched gene expression data for the same 358 individuals were taken from Genetic European Variation in Health and Disease (GEUVADIS) project [93]. Gene expression levels were measured for the same lymphoblastoid cell lines that were used for DNA-seq analysis in the 1KGP. PEER normalization was applied on the gene expression levels, with parameters optimized for cis eQTL discovery, as previously described [94]. The PEER normalization controls for multiple, potentially confounding sample covariates

in a similar way as described for the L1 expression analysis below. Normalized RNA-seq gene expression levels were accessed from the GUEVADIS project ftp server maintained at EBI:

ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/analysis_results/.

2.3.3 L1 expression quantification and normalization

Matched RNA-seq data from the GEUVEDIS project were also used to quantify the expression levels of full-length, intact and potentially active L1 elements genome-wide. The locations of 145 full-length, intact human L1s were obtained from L1Base version 2 (http://l1base.charite.de/) [95]. RNA-seq reads mapped to the human genome reference sequence build GRCh37/hg19 were access from the GEUVADIS project ftp server maintained at EBI:

ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-6/processed/

The program TEtranscripts was used to quantify the expression levels for each of the 145 individual, full-length L1 elements. The TEtranscripts program works by re-mapping RNA-seq reads genome-wide, allowing for ambiguous or multi-mapped reads, and generating read counts for both genes and TEs [96]. The program subsequently uses an EM algorithm to optimally map ambiguous reads to a unique genomic location, taking into account the entire space of reads mapped genome-wide. The resulting uniquely

mapped reads were used to generation L1-specific read counts, which were log transformed and quantile normalized to a standard normal distribution using the edgeR package [97]. A series of additional normalization steps were used to control for sample covariates, as described below, in order to allow for accurate downstream eQTL analysis.

2.3.4 Controlling sample covariates for L1 expression levels

We controlled for the effects of potential confounding variables in the eQTL analysis by regressing sample covariates out of the expression dataset, following previously suggested analysis standards [93, 98]. The specific sample covariates controlled for are (1) gender, (2) sequencing batch, (3) population group, (4) gene expression heterogeneity, and (5) population structure. All covariates were combined in a single covariates matrix C with sub-matrices representing each of the five covariates. (1) Gender: a binary vector of gender labels for each individual sample. (2) Sequencing batch: a vector with sample labels for the sequencing labs where the RNA-seq experiments were conducted. (3) Population group: a matrix of sample indicator variables for each of the four populations – CEU, FIN, GBR, and TSI. (4) Gene expression heterogeneity: to control for variance due to unknown experimental factors or confounding factors in the expression dataset, we used the top 10 principal components (PCs), corresponding to the top 10 eigenvectors, of the covariance matrix of quantile normalized expression levels as sample covariates. (5) Population structure: in addition to the population group sample labels, we also used the top 2 PCs from the SNP genotype matrix of the individuals to control for any additional population structure. The genotype matrix was built using biallelic SNPs with

minor allele frequency (MAF) > 10%. The program PLINK was then used to perform LD-pruning with a window size of 50 SNPs, a step size of 5 SNPs and a pairwise correlation cutoff of 0.5 [99]. The final genotype matrix contains genotypes for pruned SNPs across all individuals, where for each SNP location, 0, 1, and 2 correspond to reference homozygous, heterozygous, and alternative homozygous.

Using the combined covariates matrix C, made up of the five submatrices described above, we computed the hat matrix H as:

$$H = C(C^T C)^{-1} C^T$$

We then applied the hat H matrix to residualize the gene expression levels. The residualized gene expression matrix was computed as:

$$Y = (1 - H)Y'$$

where Y' is the matrix of quantile normalized expression levels and Y is the matrix of residuals. Finally, L1 expression levels were extracted from the residualized expression matrix for downstream analysis.

2.3.5 eQTL association analysis

We performed two kinds of eQTL association analyses using the program Matrix eQTL [100]: (1) association of SNP genotypes with L1 expression levels, and (2) association of SNP genotypes with gene expression levels. For the L1 eQTL associations, SNP genotypes with MAF > 5% were regressed against residualized L1 expression levels

using the additive linear analysis option of Matrix eQTL with no covariates. Sample covariates were not included since they were explicitly controlled for as described above. The false discovery rate (FDR) was used to control for multiple statistical tests, with an FDR cutoff of 0.05, corresponding to $P<1.37 \times 10^{-7}$, used to define statistically significant associations. As an eQTL association control, we permuted the unlinked SNP genotypes, by randomly assigning them to individual genomes, and then ran the same eQTL association with permuted SNP genotypes analysis against residualized L1 expression levels. The resulting random eQTL association *P*-values were used as a null distribution against which the observed eQTL association *P*-values were compared.

For the gene eQTL associations, SNP genotypes with MAF > 5% were regressed against PEER normalized gene expression levels using the additive linear analysis option of Matrix eQTL with gender and population sample covariates included. The gene eQTL associations were limited to cis SNPs, which were defined as falling within 1Mb upstream or downstream of annotated gene model boundaries (transcription start and stop sites). An FDR cutoff of 0.05, corresponding to $P < 9.54 \times 10^{-4}$, was used to define statistically significant associations. The same random SNP permutation procedure as described above was used to generate a null *P*-value distribution for comparison with the observed eQTL association *P*-values.

2.3.6 Genetic modifiers of L1 expression

To search for potential genetic modifiers of L1 expression, individual SNPs that are jointly associated with both L1 and gene expression were identified. Fisher's combined

probability test was used to combine *P*-values for matched SNP-L1 and SNP-gene eQTL associations. The χ^2 distribution, with 4 degrees of freedom (2 x 2 *P*-values combined), was used to measure the significance of the combined *P*-values. An FDR cutoff of 10^{-4}, corresponding to $P < 3.21 \times 10^{-5}$, was applied to filter out jointly significant gene-L1 pairs. We focused on eQTL associated with three classes of genes, encoding for: known L1 suppressors, transcription factors, and chromatin associated proteins. A collection of 34 previously characterized L1 suppressor genes were curated from the literature. Human gene ontology (GO) annotations provided NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz) were parsed in order to identify 1,244 transcription factor genes and 450 chromatin association protein encoding genes.

2.4 Results and Discussion

2.4.1 Genome-wide screening with shared eQTL associations

We used joint eQTL association analysis of L1 and gene expression levels in order to conduct a genome-wide screen for genetic modifiers of L1 expression (Figure 1). Genome-wide SNP variant calls for 358 individuals from 4 European populations – CEU, FIN, GBR, and TSI – were taken from the phase 3 release of the 1KGP, and matched gene expression levels for the same individuals were taken from the GEUVEDIS RNA-seq project. Gene expression levels were characterized for the same lymphoblastoid cell lines that were used for DNA-seq analysis in the 1KGP project. Gene expression levels were normalized using the PEER method in order to optimize downstream eQTL association analyses. We re-analyzed the GEUVEDIS RNA-seq read data in order to measure gene

expression levels at 145 full-length, potentially active L1 loci. L1 expression analysis was performed in such a way as to find the best locations for multi-mapping sequence tags and to control for potentially confounding sample covariates, such as population structure and sequencing batch (Supplementary Figure S1). The details of our L1 and gene expression analyses are provided in the Materials and Methods.

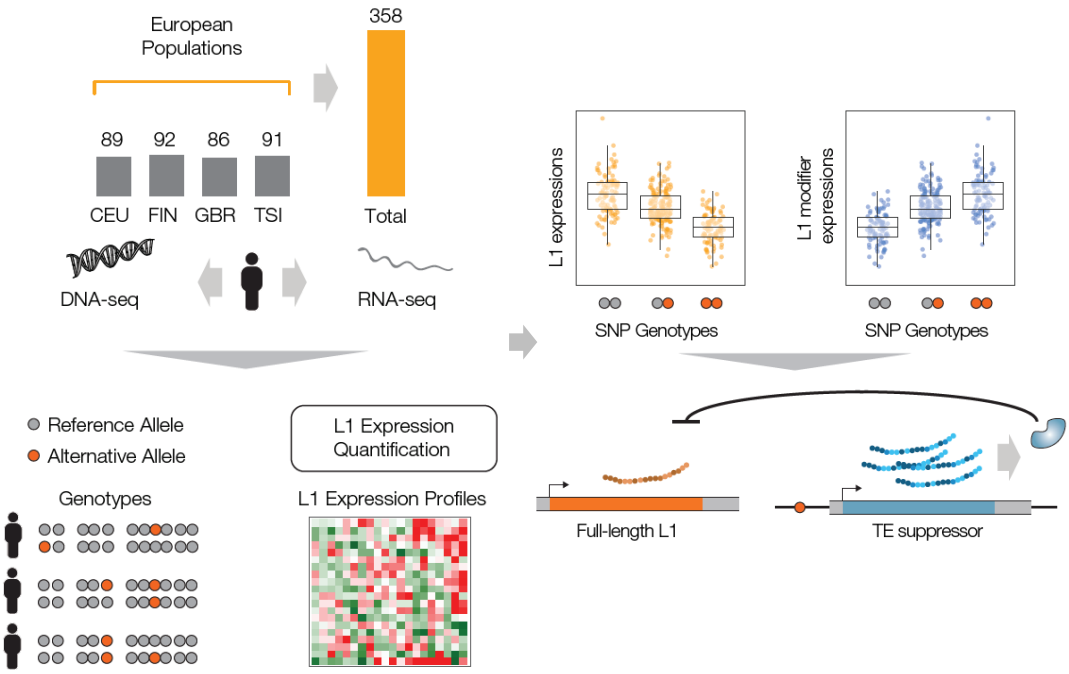


Figure 2 Genome-wide approach to screen for genetic modifiers of L1 expression

SNP genotype and transcriptome profiling of 358 European individuals were combined for expression quantitative trait loci (eQTL) analysis. SNP genotypes were regressed against L1 and gene expression levels to search for shared eQTL associations. An example of a shared eQTL association that represents a L1 suppressor is shown, where the SNP genotypes associated with decreased L1 expression are simultaneously associated with increased expression of the putative L1 modifier gene.

Individual SNP genotypes were regressed against both L1 and gene expression levels to search for shared associations (Figure 1). SNPs that are jointly associated with

both L1 and gene expression are considered to be putative genetic modifiers of L1 expression. We did this for known L1 modifier genes, as a proof of principle (*i.e.* a positive control), along with candidate L1 modifier genes encoding transcription factors and chromatin associated proteins. Results of our genome-wide eQTL analysis are shown separately for L1 expression (Figure 3A) and for gene expression (Figure 3B). The observed eQTL associations for both L1 and gene expression are compared to a null distribution of eQTL associations generated via random permutation of SNP genotypes. The Q-Q plots for L1 and gene expression show substantially more significant associations than expected by chance, as indicated by the deviation of the observed values from both the diagonal line and the randomly permuted SNP genotype association values. Manhattan plots also show a striking difference between the genome-wide distributions of the observed versus permuted eQTL associations. The L1 eQTL association Manhattan plot shows distinct peaks on chromosome 6, 12, and 16. There are far more significant association peaks observed for gene expression, likely owing to the fact that we analyzed far more gene (22,128) than L1 (145) loci.

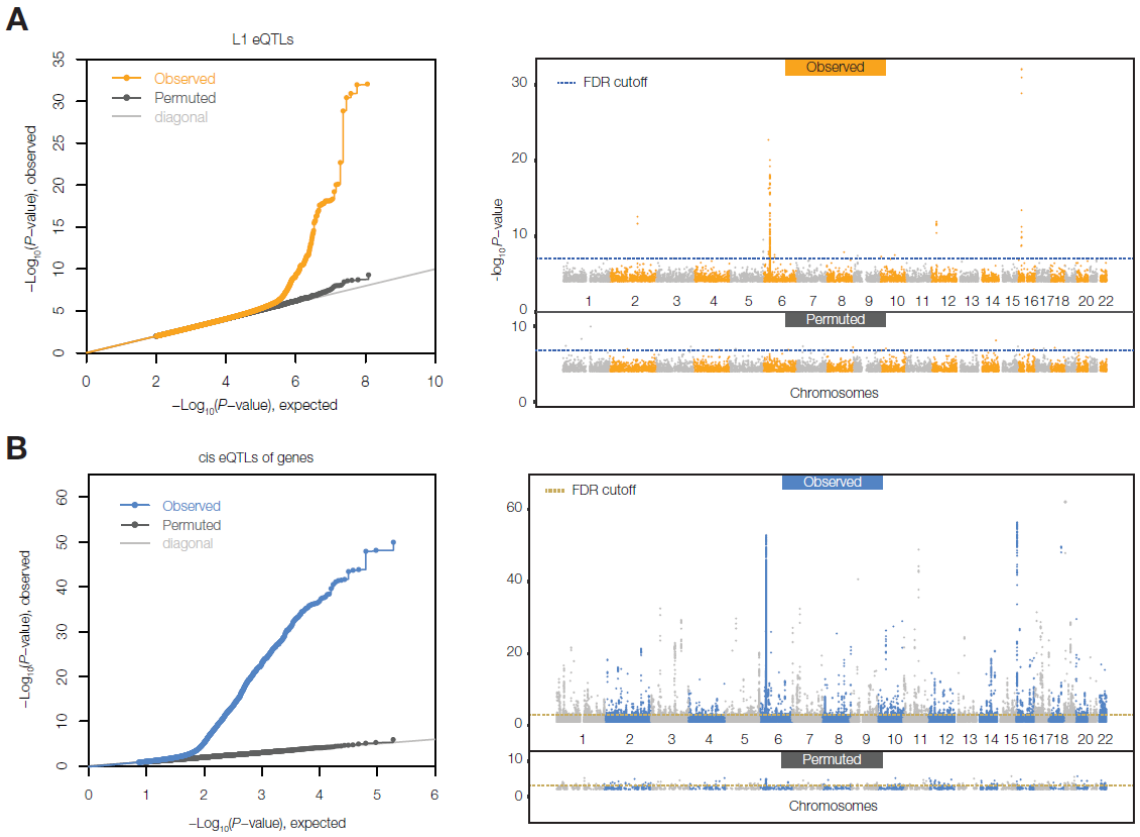


Figure 3 Results of the eQTL association analyses for L1 and gene expression

Quantile-quantile (Q-Q) plots (left) and Manhattan plots (right) are shown for the L1 (A - orange) and gene (B - blue) results. The Q-Q plots show the observed eQTL association log transformed P-values (L1-orange and gene-blue dots) compared to randomly permuted association P-values (gray dots) and the expected P-value distribution under the null hypothesis of no association (gray lines). The upper panels of the Manhattan plots show the observed eQTL association log transformed P-values and the lower panels show randomly permuted association P-values.

L1 and gene eQTL association *P*-values were combined to screen for shared SNP eQTL, *i.e.* putative L1 genetic modifiers. Fisher's combined probability test was used to combine L1 and gene association *P*-values, yielding χ^2 values for shared associations (Figure 4). The combined *P*-value χ^2 distributions show 2,346 statistically significant

shared L1-gene eQTL associations (χ^2>25.9, df=4, FDR q<10^{-4}) across all three classes of genes that we evaluated. This includes numerous shared associations with known L1 modifier genes as should be expected if our genome-wide screen is adequately powered to detect L1-gene regulatory interactions. We also found many shared associations with genes that encode transcription factor and chromatin associated proteins; these shared associations point to potentially novel genetic modifiers of L1 expression. Comparison of the three gene categories does not show any substantial difference in the strength of shared associations. The top shared associations observed for each of these three categories of genes are shown in Table 1, and examples for each of the three categories are described in the following sections of the manuscript.

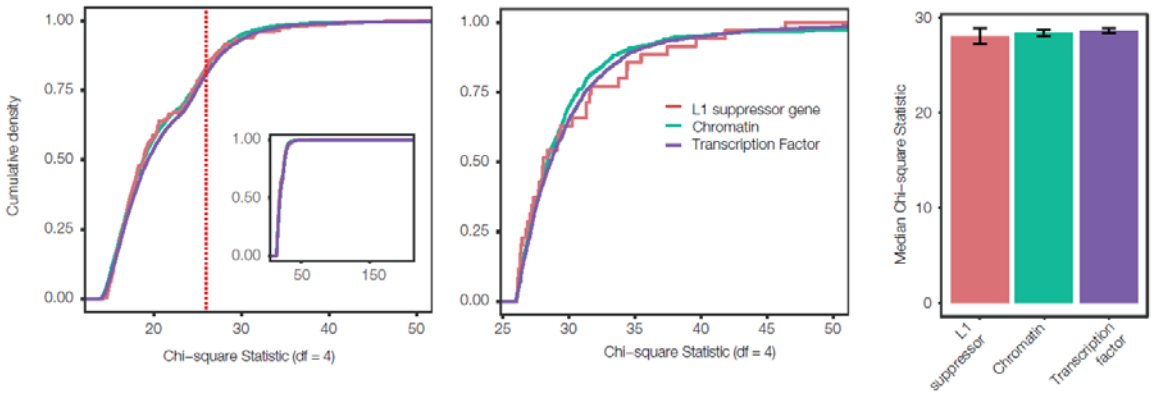


Figure 4 Results of the joint L1-gene eQTL association analysis

SNP rs11126321 genotypes Joint L1-gene eQTL results are shown for SNPs associated with known L1 suppressors (red), transcription factor (purple), and chromatin associated (green) genes. Cumulative distributions are shown for Fisher's combined probability test χ^2_4 values for all three gene sets. The left panel shows the entire distribution, and the middle panel shows only statistically significant values (FDR < 10^{-4}). The right panel shows the median and standard error of the χ^2_4 values for the three gene sets.

2.4.2 Known modifiers of L1 expression

The potential utility of our eQTL genome-wide screen for L1 genetic modifiers can be illustrated by the results seen for known modifiers of L1 expression, including both transcriptional and post-transcriptional modifiers. For example, the alternate allele (A) of the SNP rs11126321 was found to be jointly associated with increased L1 expression and decreased expression of the DNA Methyltransferase 3 Alpha gene (*DNMT3A*) (Figure 5A). *DNMT3A* encodes a *de novo* DNA methyltransferase that is known to repress L1 expression via the methylation of CpG islands proximal to element promoters [83]. Accordingly, the expression levels of *DNMT3A* and L1 are expected to be inversely correlated as can be seen here.

In a similar example, we found shared, and inverse, eQTL L1-gene associations for the SNP rs6537785. The alternate allele of this SNP (T) is associated with increased L1 expression and decreased expression of the Mov10 RISC Complex RNA Helicase (*MOV10*) (Figure 5B). *MOV10* has been implicated in post-transcriptional regulation of L1 elements by virtue of its role as an inhibitor of L1 mRNA transport between the nucleus and the cytoplasm, a critical, rate-limiting step it the retrotransposition cycle [52]. *MOV10* achieves this by promoting the degradation of L1 mRNA in the cytoplasm via the formation of cytoplasmic stress granules, thereby limiting their ability to return to the nucleus where reverse transcription takes place.

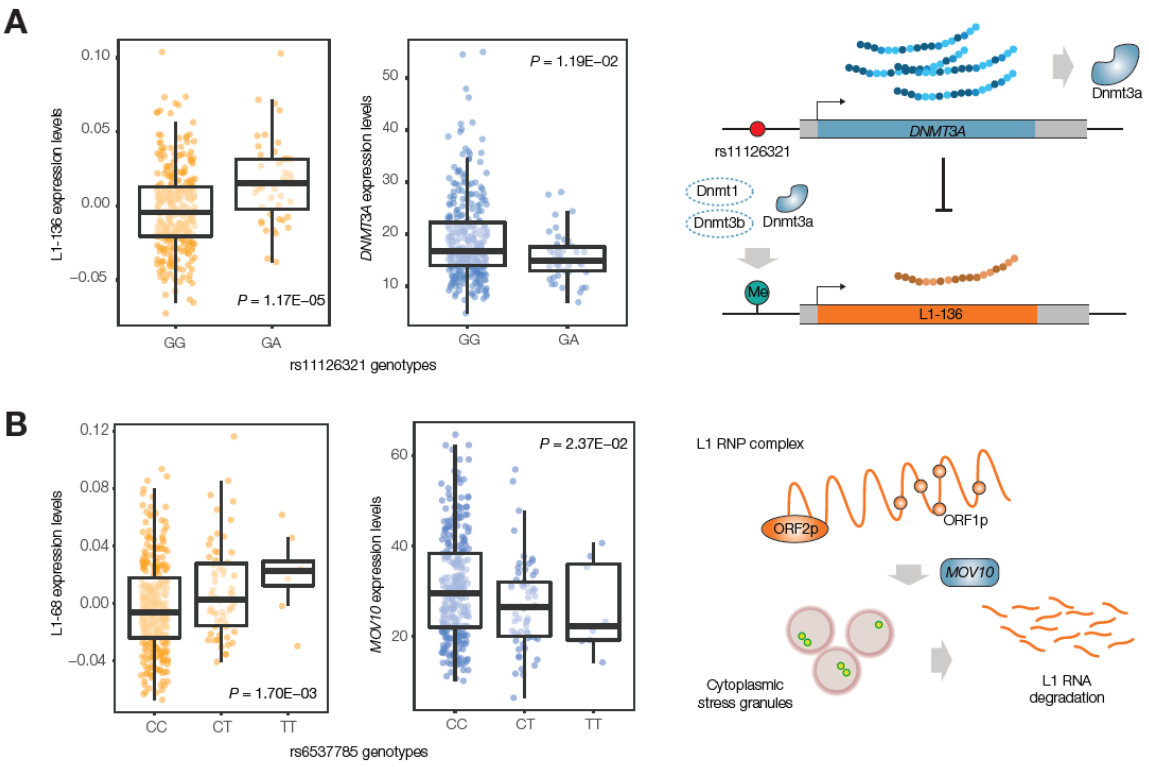


Figure 5 Known L1 suppressor genes implicated by the joint eQTL association analysis

(A) L1 and gene eQTL associations for a known L1 transcriptional suppressor. SNP rs11126321 genotypes are jointly associated with increased L1 expression and decreased expression of DNMT3A (left). A schematic for the role of DNMT3A in L1 suppression (right). (B) L1 and gene eQTL associations for a known post-transcriptional suppressor of L1. SNP rs6537785 genotypes are jointly associated with increased L1 expression and decreased expression of MOV10 (left). A schematic for the role of MOV10 in L1 post-transcriptional suppression (right).

An interesting counter example of shared L1-gene associations for known L1 modifiers was seen for the Nuclear RNA Export Factor 1 gene (*NXF1*). In this case, the SNP rs111438108 shows the same pattern of association for L1 and gene expression, whereby the alternate allele (T) is associated with decreased expression of both (Figure 6). *NXF1* also plays an important role in the transport of L1 mRNAs from the nucleus to the

cytoplasm by forming part of the nuclear pore through which the mRNAs pass to the cytoplasm [101]. Presumably, increased expression of the gene would facilitate greater transport of L1 mRNAs and accordingly higher apparent L1 expression.

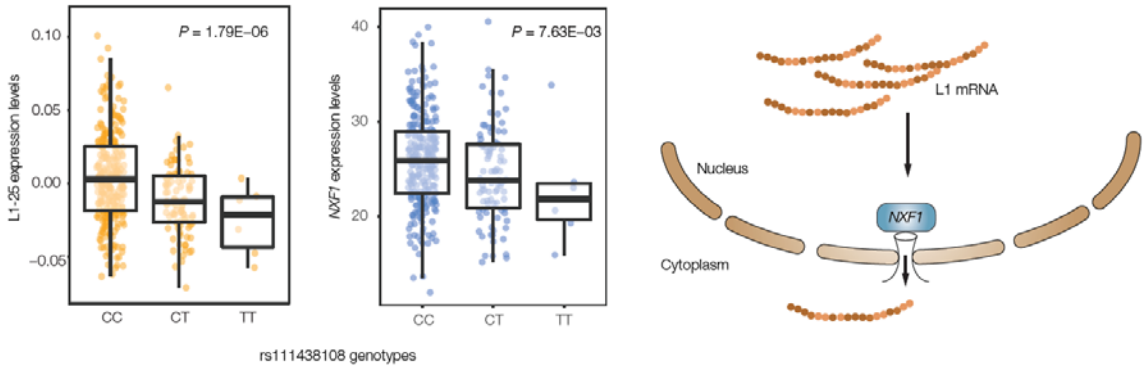


Figure 6 Known L1 retrotransposition promoting genes implicated by the joint eQTL association analysis.

L1 and gene eQTL associations for a known post-transcriptional modifier of L1. SNP rs111438108 genotypes are jointly associated with decreased L1 expression and decreased expression of NXF1 (left). A schematic for the role of NXF1 is enhancing L1 expression (right).

2.4.3 Novel transcription factor L1 modifiers

We were most interested in the discovery of novel potential modifiers of L1 expression via our shared eQTL approach. We found a number of transcription factors, which to our knowledge were not previously implicated in L1 regulation, that are associated with SNPs which are also eQTL for L1 expression (Table 1). For transcription factors, we focused on shared eQTL that showed the same direction of association, reasoning that their increased expression may lead in turn to upregulation of L1 elements. An example of this pattern can be seen for the SNP rs28515780, for which the alternate

allele (A) is associated with increased gene expression of the TATA-Box Binding Protein Associated Factor 13 gene (*TAF13*) along with increased expression of four L1 loci (Figure 7A). *TAF13* encodes the 18 KDa subunit of the basal transcription initiation factor protein TFIID, which is expected to be crucial for the expression of L1 elements by the RNA polymerase II machinery.

Table 1 Top modifiers genes identified via joint eQTL analysis

Gene set	SNP	Gene	L1 ID	χ^2 statistic	P-value	FDR q-value
TE suppressor	rs2887269	*HNRNPL*	L1-136	34.41	6.14E-07	4.60E-06
	rs11126321	*DNMT3A*	L1-136	31.57	2.34E-06	1.42E-05
	ss1388073530	*AICDA*	L1-49	29.35	6.64E-06	3.22E-05
	rs7529736	*AGO1*	L1-56	29.19	7.15E-06	3.41E-05
	rs467053	*SQSTM1*	L1-52	28.38	1.04E-05	4.53E-05
Chromatin	rs4668938	*DDX1*	L1-78	103.62	1.66E-21	9.27E-20
	rs28366287	*CSNK2B*	L1-28	98.19	2.39E-20	1.19E-18
	rs540748076	*BRD2*	L1-28	54.88	3.45E-11	5.88E-10
	rs10902548	*MEF2A*	L1-106	53.18	7.82E-11	1.27E-09
	rs13202640	*GMNN*	L1-120	48.61	7.05E-10	9.78E-09

Gene set	SNP	Gene	L1 ID	χ^2 statistic	P-value	FDR q-value
Transcription Factor	rs386525443	*TCF19*	L1-103	72.19	7.84E-15	2.04E-13
	rs113823631	*RXRB*	L1-57	64.07	4.05E-13	8.76E-12
	rs386525443	*POU5F1*	L1-103	61.28	1.56E-12	3.18E-11
	rs841657	*ARNTL2*	L1-94	56.91	1.29E-11	2.34E-10
	rs7258563	*ZNF790*	L1-129	51.12	2.11E-10	3.19E-09

A similar example of this kind can be seen for genes that encode a pair of transcription factors – Transcription Factor 19 (*TCF19*) and POU Class 5 Homeobox 1 (*POU5F1*) – both of which show increased expression associated with the alternate allele (T) of the SNP rs386525443 (Figure 7B). The same allele of rs386525443 is also associated with increased expression of four L1 elements. These two transcription genes are located in close proximity on the short arm of chromosome 6, within the major histocompatibility locus (MHC). There is evidence suggesting that these two genes function together, and they are both important in embryonic development and stem cell pluripotency. It may be the case that they are connected to developmentally regulated expression of L1 elements.

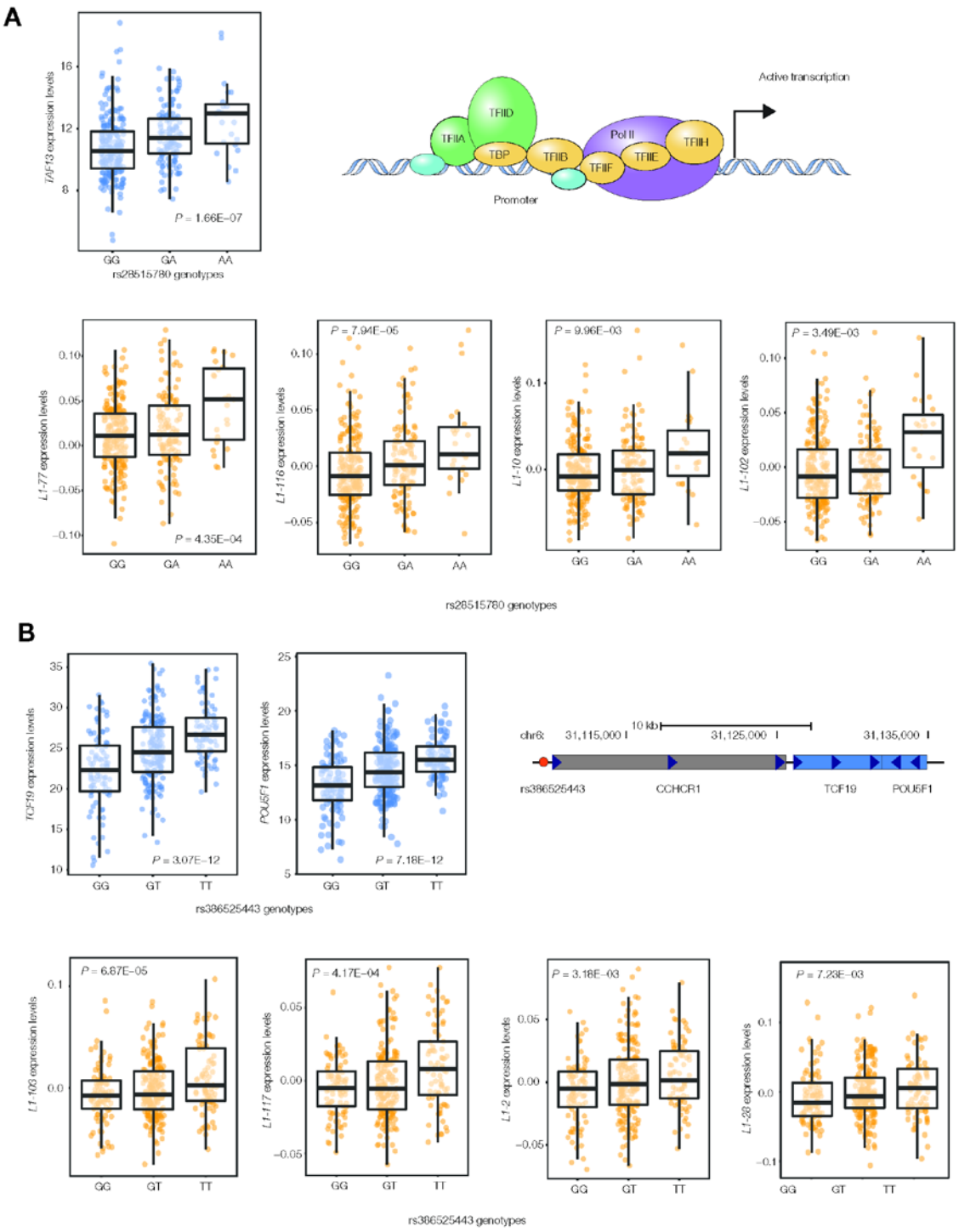


Figure 7 Putative L1 modifier transcription factors uncovered by the joint eQTL association analysis

(A) Shared gene and L1 eQTL associations for the putative L1 transcription factor TAF13. SNP rs28515780 genotypes are jointly associated with increased expression of TAF13

(upper – blue) and four L1 loci (lower – orange). A schematic for the role of TAF13 in regulating L1 transcription is shown. (B) Shared gene and L1 eQTL associations for the putative transcription factors TCF19 and POU5F1. SNP rs386525443 genotypes are associated with increased expression of both TCF19 and POU5F1 (upper – blue), along with increased expression of four L1 loci (lower – orange). A schematic showing the shared genomic location of TCF19 and POU5F1.

2.4.4 Novel chromatin L1 modifiers

We searched for novel chromatin modifiers of L1 expression, with an emphasis on shared L1-gene eQTL that showed the opposite direction of association. The rationale behind this approach is based on the fact that chromatin compaction plays an important role in the suppression of L1 activity [102] and can be considered to serve as the 'ground state' through which L1 elements are held silent across the genome. Accordingly, decreased expression of chromatin modifiers is expected to lead to increased expression of L1 elements. We found three histone deacetylase encoding genes – *HDAC1*, *HDAC3*, and *HDAC5* – each of which has a shared and inverse eQTL association with L1 expression (Figure 8 A-C). Activity of these histone deacetylases leads to decreased acetylation, decrease chromatin compaction and increased expression of L1s, consistent with the inverse directions of the shared associations observed here. We found a similar inverse shared eQTL association for the Lysine Demethylase 5A encoding gene (*KDM5A*), which demethylates lysine 4 of the histone H3 (Figure 8D). Histone methylation is also a repressive chromatin mark, so decreased expression of this gene is expected to be associated with increased L1 expression as we observe here (Figure 8E).

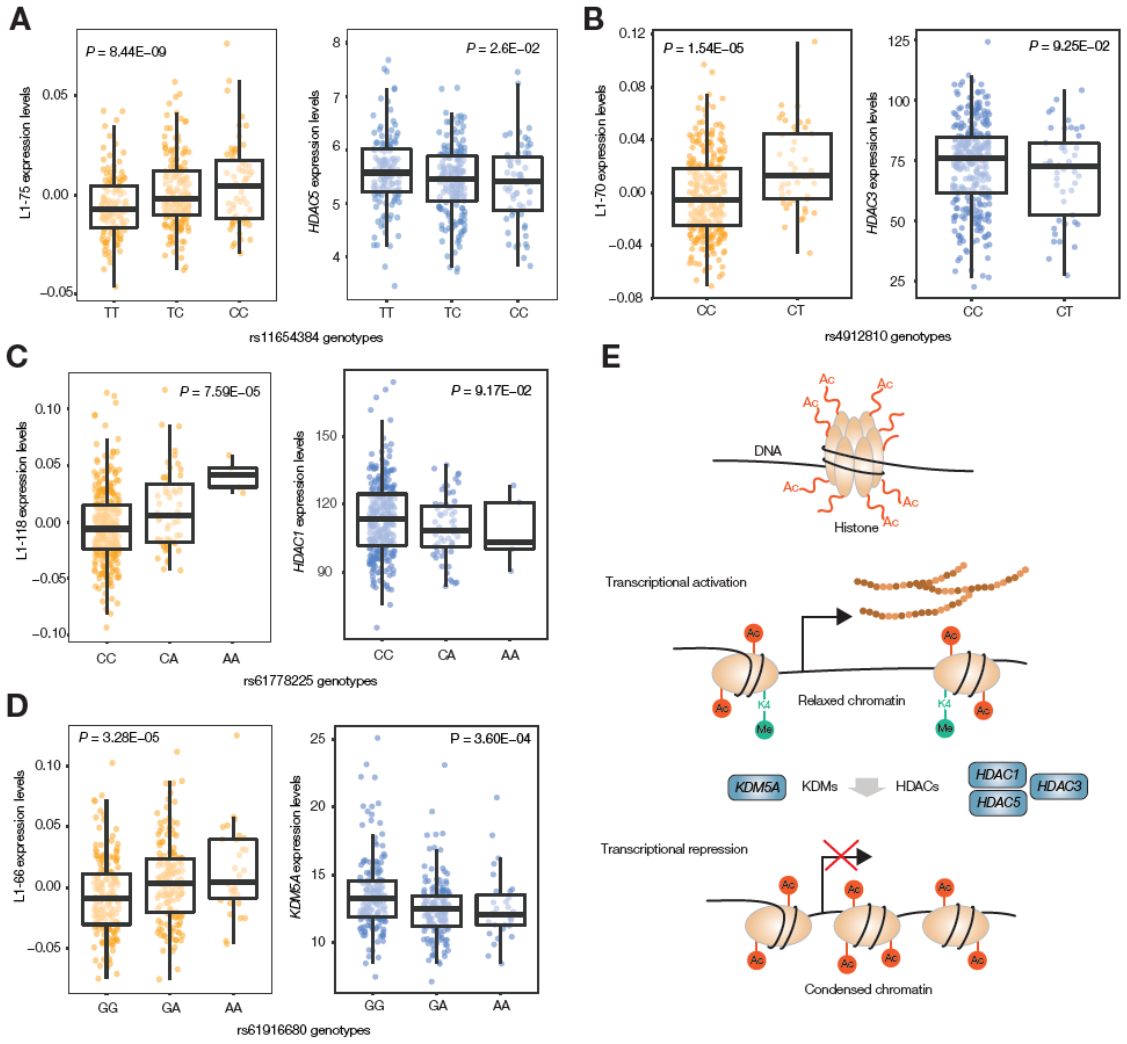


Figure 8 Putative chromatin L1 modifiers uncovered by the joint eQTL association analysis

Shared gene and L1 eQTL associations are shown for three histone deacetylase genes: HDAC5 (A), HDAC3 (B), and HDAC1 (C) along with the histone demethylation gene KDM5A (D). In each case, there is a single SNP with genotypes that are simultaneously associated with increased L1 expression and decreased expression. (E) A schematic illustrating the L1 suppression roles of the histone deacetylases and the histone demethylation genes.

2.5 Conclusions

Our genome-wide approach to the discovery of genetic modifiers of L1 expression yielded a number of promising results, including confirmation of previously characterized L1 modifiers (*i.e.* positive controls) as well as potentially novel L1 modifiers that act at the level of transcription initiation or chromatin modification. These results attest to the power of genome-scale eQTL studies to decipher the regulatory architecture governing the expression of human TEs. Given TEs' known roles in mutation and genome dynamics, a deeper understanding of genetic modifiers of L1 expression has important implications for studies of mutagenesis and genome stability.

The current study is limited by its focus on a single tissue type – the lymphoblastoid cell lines that were used for both the 1KGP DNA-seq and GEUVEDIS RNA-seq studies. This limitation is based on the fact that eQTL association studies of the kind employed here require the characterization of genome-wide sets of genotype calls and gene expression levels from hundreds of matched human samples. Until recently, this has only been available for the single tissue type analyzed here. However, the recently completed GTEx project provides matched genotype and expression data for scores of human samples. Exploration of this rich data set, together with consideration of the known tissue-specific expression profiles of L1 elements, should yield additional resolution for the discovery of novel L1 modifiers. In particular, paired discovery and validation analyses could be used to provide an additional line of support for putative L1 modifiers.

Another note of caution is that the identities of the novel L1 genetic modifiers discovered here should be treated with caution given the bioinformatics techniques

employed for their discovery. Indeed, these putative L1 modifiers can be best considered as predictions, or hypotheses, that will need to be validated by careful experimental efforts. Nevertheless, the genome-wide screen approach that we took here should prove to be useful in guiding future experiments. In particular, we hope that evaluation of our lists of putative L1 modifiers by L1 experts, and experimentalists with deep knowledge of L1 regulation, will prove to be useful in substantially narrowing the space of possible studies that need to be done to discover novel dimensions of human genome regulation and dynamics.

CHAPTER 3. HUMAN POPULATION-SPECIFIC GENE EXPRESSION AND TRANSCRIPTIONAL NETWORK MODIFICATION WITH POLYMORPHIC TRANSPOSABLE ELEMENTS

3.1 Abstract

Transposable element (TE) derived sequences are known to contribute to the regulation of the human genome. The majority of known TE-derived regulatory sequences correspond to relatively ancient insertions, which are fixed across human populations. The extent to which human genetic variation caused by recent TE activity leads to regulatory polymorphisms among populations has yet to be thoroughly explored. In this study, we searched for associations between polymorphic TE (polyTE) loci and human gene expression levels using an expression quantitative trait loci (eQTL) approach. We compared locus-specific polyTE insertion genotypes to B cell gene expression levels among 445 individuals from 5 human populations. Numerous human polyTE loci correspond to both cis and trans eQTL, and their regulatory effects are directly related to cell type-specific function in the immune system. PolyTE loci are associated with differences in expression between European and African population groups, and a single polyTE loci is indirectly associated with the expression of numerous genes via the regulation of the B cell-specific transcription factor PAX5. The polyTE-gene expression associations we found indicate that human TE genetic variation can have important phenotypic consequences. Our results reveal that TE-eQTL are involved in population-specific gene regulation as well as transcriptional network modification.

3.2 Introduction

Transposable elements (TEs) are mobile DNA sequences that create copies of themselves when they move among chromosomal locations. TE activity has had a major impact on the evolution and structure of the human genome; millions of TE sequence copies have accumulated over the last ~100my. The initial sequencing and subsequent analysis of the human genome revealed that >50% of the genome sequence is derived from past TE sequence insertions [1, 2].

TEs can also shape the function of the human genome, particularly with respect to the regulation of gene expression[17]. Human TE-derived sequences have been shown to provide a wide variety of gene regulatory sequences including promoters [19-21], enhancers [22-26], transcription terminators [28] and several classes of small RNAs [29-31]. Human TEs also influence various aspects of chromatin structure throughout the genome [1, 32, 103-106].

The vast majority of human TE sequences are remnants of ancient transposition events that occurred many millions of years ago [1]. Accordingly, studies that have uncovered the regulatory properties of TE-derived sequences have dealt with fixed TE insertions that are present at the same locations in the genome sequences of all human individuals. Such fixed TE-derived regulatory sequences are not expected to provide for gene regulatory variation based on insertional polymorphisms between individuals.

It has only recently become possible to systematically evaluate the effects of TE genetic variation within and between human populations, *i.e.* TE polymorphisms. Human TE polymorphisms are primarily generated via the activity of three families of

retrotransposons: Alu [3, 4], L1 [5, 6] and SVAs [7, 8]. Transposition events by members of these polymorphic TE (polyTE) families yield numerous differences in the presence/absence of insertions at specific loci among individual human genome sequences. The recent phase 3 variant release of the 1000 Genomes Project included a catalog of presence/absence genotypes for >16,000 polyTE loci among 2,504 individuals from 26 human populations [65, 99]. This genome-wide collection of polyTE genotypes provides an opportunity to explore the phenotypic consequences of TE activity at the level of human populations.

Considering the known regulatory properties of human TEs, together with the fact that TE insertional activity is known to be highly disruptive [34, 35], we hypothesized that polyTE activity can lead to gene expression differences between human individuals. We used an integrated analysis of polyTE genotypes and genome-wide expression profiles, for the same set of 1000 Genome Project samples, in order to test this hypothesis (Figure 9). Gene expression levels were regressed against polyTE genotypes to search for polyTE-gene expression associations. This approach revealed numerous human polyTE loci that correspond to expression quantitative trait loci (eQTL). The TE-eQTL uncovered here are involved in the establishment of population-specific expression profiles as well as transcriptional regulatory network modification.

3.3 Materials and Methods

3.3.1 Polymoprhic transposable element (polyTE) analysis

Genotype calls for three families of human polyTEs – Alu, L1 and SVA – in 445 individuals from 5 populations were taken from the phase 3 variant release of the 1000 Genomes Project[99]. The phase 3 variant release corresponds to the human genome reference sequence build GRCh37/hg19. The 5 human populations are CEU: Utah Residents (CEPH) with Northern and Western Ancestry, FIN: Finnish in Finland, GBR: British in England and Scotland and TSI: Toscani in Italy from Europe along with YRI: Yoruba in Ibadan, Nigeria from Africa (Figure 9). These populations were chosen because they have matching RNA-seq data for the same individuals (see RNA-seq analysis section). PolyTE genotypes were characterized by the 1000 Genomes Project Structural Variation Group using the program MELT as previously described [65]. The polyTE genotype data were accessed from the 1000 Genomes Project ftp server maintained at the NCBI: http://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/release/20130502/.

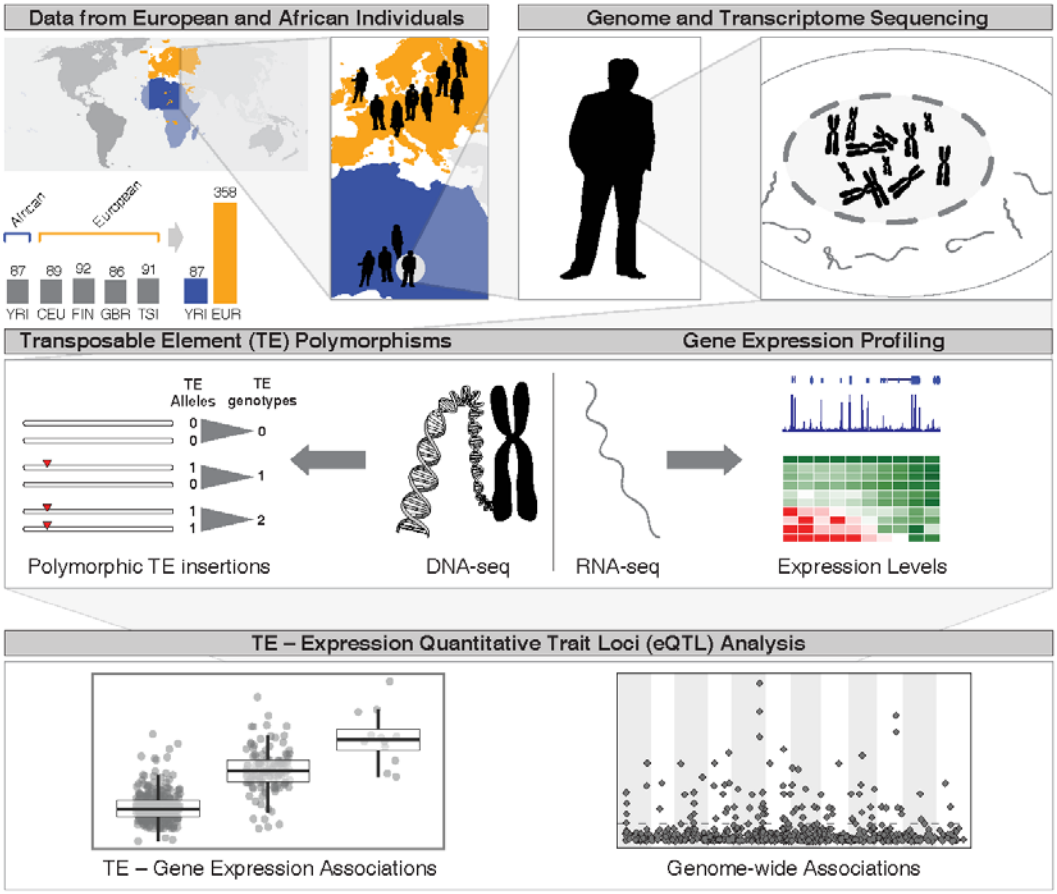


Figure 9 Scheme for the polymorphic transposable element (polyTE) expression quantitative trait loci (eQTL) analysis conducted

Data were taken from 87 African and 358 European individuals from the 1000 Genomes Project. Genome (DNA-seq) and transcriptome (RNA-seq) data were used to characterize polyTE genotypes and gene expression levels for all individuals in the study. Individual gene expression levels were regressed against polyTE insertion genotypes in an effort to reveal associations between polyTE loci and human gene expression, i.e. TE-eQTL.

For any given polyTE insertion site, there are three possible presence/absence genotype values for an individual genome: 0-no polyTE insertion (homozygote absent), 1-a single polyTE insertion (heterozygote), and 2-two polyTE insertions (homozygote

present). PolyTE genotypes were used for eQTL analysis as described below. PolyTE genotypes were also used to compute pairwise genetic distances between individuals as: $d_{xy}^{g} = \frac{1}{n}\sum_{i=1}^{n} |g_{xi} - g_{yi}|$, where g_{xi} and g_{yi} are the polyTE genotype value for individual x and individual y at insertion site i, for a total of n sites. The resulting pairwise polyTE genotype distance matrix was subject to dimension reduction using multidimensional scaling (MDS)[107], using the cmdscale function from the R statistical package version 3.2.2[108], in order to visualize the genetic relationships between individuals based on their polyTEs (Figure 10C).

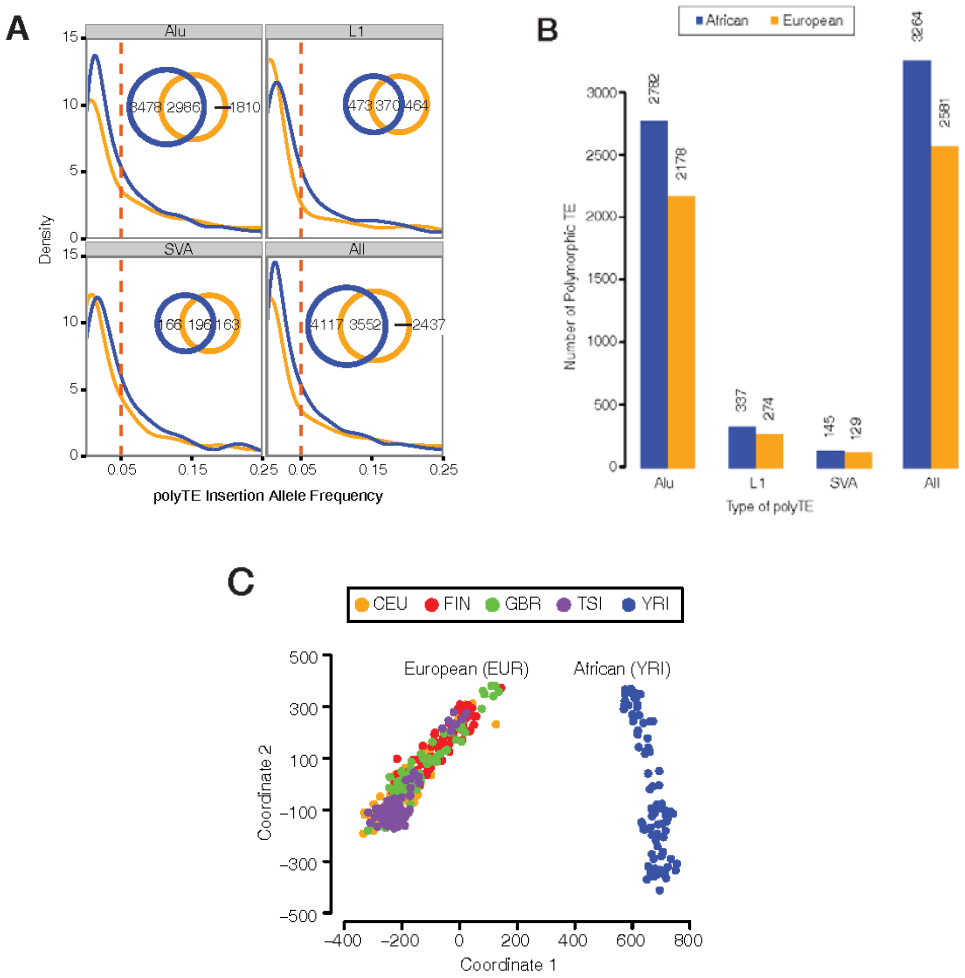


Figure 10 Distribution of polyTEs among the African and European population groups analyzed

Data are broken down into Alu, L1 and SVA polyTE families. (A) PolyTE insertion allele frequency distributions for African and European populations are shown along with the numbers of shared and population-specific polyTE loci. (B) The numbers of African and Europen polyTE insertions with allele frequencies >5%. (C) Genetic relationships among the individuals analyzed here based on their polyTE genotypes. Individual's population origins are color coded as shown in the key.

3.3.2 RNA sequencing (RNA-seq) analysis

RNA-seq expression data, for the same 445 individuals from 5 human populations with polyTE genotypes characterized as described in the previous section, were taken from the GUEVADIS RNA sequencing project for 1000 Genomes samples[109]. These RNA-seq data characterize genome-wide expression levels from the same lymphoblastoid cell lines, *i.e.* Epstein–Barr virus (EBV) transformed B-lymphocytes, used for DNA-seq analysis in the 1000 Genomes project. RNA isolation, library preparation, sequencing and read-to-genome mapping was performed as previously described [109]. As with the polyTE data, the RNA-seq read mapping corresponds to the human genome build GRCh37/hg19. Mapped reads were used to quantify gene expression levels for ENSEMBL gene models[110] and normalization of gene expression levels was done using a combination of a modified RPKM approach followed by the probabilistic estimation of expression residuals (PEER) method[111] as previously described[112]. The PEER normalized RNA-seq gene expression levels were accessed from the GUEVADIS project ftp server maintained at EBI:

ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/analysis_results/.

Genome-wide expression profiles were used to compute pairwise phenotypic distances between individuals as: $d_{xy}^{e} = \sqrt{\sum_{i=1}^{n}(e_{xi} - e_{yi})^2}$ where e_{xi} and e_{yi} are the normalized gene expression level for individual x and individual y at gene i, for a total of n genes. The resulting pairwise expression distance matrix was subject to dimension reduction using multidimensional scaling (MDS)[107], using the cmdscale function from

the R statistical package version 3.2.2[108], in order to visualize the relationships between individuals based on their genome-wide expression profiles (Figure 11A). Differential gene expression between African and European populations was evaluated using a paired ttest implemented with the *genefilter* package from Bioconductor[113] (Figure 11B).

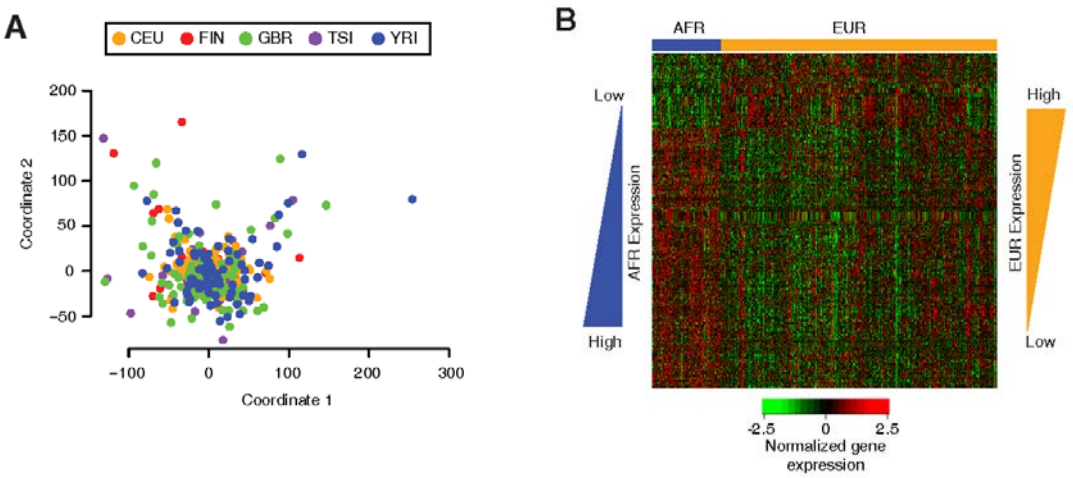


Figure 11 Gene expression profiles within and between populations analyzed

(A) Individuals from different populations are related based on their genome-wide expression profiles. Individual's population origins are color coded as shown in the key. (B) Heatmap showing genes that have expression profiles that are significantly different between the African and European population groups. Gene expression levels are color coded as shown in the key.

3.3.3 Expression quantitative trait loci (eQTL) analysis

PolyTE genotypes from the individuals analyzed here were related to their gene expression levels to identify eQTLs that correspond to polyTE insertion sites (TE-eQTLs) using the program Matrix eQTL[114] (Figure 9). Only polyTE insertion sites with >5% TE-present allele frequency were used for this purpose (Figure 10A and B). Matrix eQTL was run using the additive linear (least squares model) option with covariates for gender

and population. This was done for all possible pairs of polyTE insertion sites and genes. *Cis* versus *trans* eQTLs were defined later as polyTE insertion sites that fall inside (*cis*) or outside (*trans*) 1Mb from gene boundaries. *P*-values were calculated for all pairs of polyTE-gene expression comparisons, and FDR *q*-values were then calculated to correct for multiple statistical tests. The genome-wide significant polyTE-gene expression eQTL association threshold was set at FDR $q<0.05$ ($P<4.7e-7$).

A series of three additional control analyses were implemented in an effort to control for potentially confounding effects regulatory SNPs in particular, on the TE-eQTL associations that passed the genome-wide significance threshold (Figure 27). [Control 1] TE-eQTL versus SNP-eQTL comparisons: For all of the genes found to be associated with TE-eQTLs, we searched the results of the GEUVADIS RNA-seq project [109] to identify the number of SNPs that were previously implicated as eQTLs for the same genes (Figure 27A). [Control 2] Conditional association analysis: For the genes that were found to be associated with both TE-eQTLs and SNP-eQTLs, we performed conditional association analysis whereby multiple regression of expression against genotype is done using both TE and SNP genotype information used as explanatory variables in the same multiple regression model (Figure 27B). The conditional association analysis was performed using the same multiple regression approach as implemented in the GCTA program[115]. [Control 3] Regional association scans: Regional eQTL association scans were done by defining linked 1Mb regions that are centered on individual polyTE loci, and then all SNP and polyTE genotypes from the linked regions were further evaluated for association with gene expression using the same approach used for TE-eQTLs (Figure 27A). Results of the

regional eQTL association scans were visualized using the regional association plot R script from the Broad Institute of MIT and Harvard [116].

3.3.4 Functional enrichment analysis

Genes that correspond to best TE-eQTLs were used for gene set enrichment analysis using the KEGG, BIOCARTA and REACTOME data sets from the Molecular Signatures Database webserver (version 5.1)[117] in order to identify functionally enriched gene categories. A FDR *q*-value threshold of 0.05 was used for this purpose.

3.3.5 Transcription factor (TF) target identification

TF (*PAX5*) target genes were taken from annotations of experimentally characterized TF binding sites from the 2015.1 version of GENOME TRAX™ (www.biobase-international.com/genome-trax) from BIOBASE corporation[118]. TF-target gene interactions were visualized using the program Circos (version 0.69)[119].

3.4 Results

3.4.1 The landscape of human TE polymorphisms

Computational analysis of next-generation (re)sequencing data can be used to identify the locations of polyTE insertions genome-wide[120]. Recent applications of this

approach to human genome sequences from the 1000 Genomes Project has resulted in a deep characterization of human genetic variation resulting from TE activity[65, 99]. We analyzed polyTE loci from the genome sequences of 445 individuals sampled from 5 human populations (4 European and 1 African) characterized as part of this project. There are a total of 10,106 polyTE insertions observed for these 445 individual genome sequences: 8,274 for Alu, 1,307 for L1 and 525 for SVA (Figure 10A). Most of the polyTE insertions that we observe (9,799) can be considered to be *cis* to human genes as they either fall within gene boundaries or within 1Mb upstream or downstream of genes. Furthermore, consistent with previous results, the majority of polyTE loci for these 5 populations show low frequencies of TE insertions (*i.e.*, low minor allele frequencies), suggesting that TE insertions are highly disruptive and subject to strong purifying selection[65, 121]. Nevertheless, there are 2,617 polyTE loci that show >5% TE insertion frequency for these populations (Figure 10B); these common polyTE loci were used for the subsequent eQTL analysis. The vast majority of these are Alu polyTE loci with an order of magnitude fewer L1 and fewest SVA loci.

Despite the similar shapes of the TE insertion allele frequency distributions, many of the loci are specific to individual populations or continental population groups. Indeed, genetic distances between individuals calculated based on their polyTE genotypes clearly separates European from African populations (Figure 10C). Population-specific polyTE loci with higher insertion frequencies can be considered to be more likely to exert broad regulatory effects across individuals and populations. Accordingly, we focused our subsequent analysis on these (relatively) high frequency polyTE loci, and searched for possible population-specific regulatory effects of such loci.

3.4.2 TE expression quantitative trait loci (TE-eQTL)

We analyzed genome-wide expression profiles for these same individuals in an effort to evaluate the relationship between TE genetic variation and human gene regulation. Genome-wide expression profiles were compared to compute a pairwise phenotypic (regulatory) distance matrix for the individuals analyzed here. Unlike what is seen for the polyTE genetic distances, genome-wide expression profiles do not separate individual humans among different population groups (Figure 11A). In other words, gene expression variation does not segregate globally in the same way that TE genetic variation does. Nevertheless, there are several hundred genes that do show statistically different levels of expression between the African and European populations analyzed here (Figure 11B).

We evaluated the relationship between TE genetic variation and human gene regulation by searching for expression quantitative trait loci (eQTL) that correspond to polyTE insertion sites. To do this, gene-specific expression levels were regressed against presence/absence genotypes – 0, 1 or 2 TE insertions – for individual polyTE loci (Figure 9). We used an additive linear model as described in the Methods section to search for statistically significant associations between the polyTE genotypes at any given locus and expression levels for individual genes. This was done separately for African and European population groups as well as for all individuals considered together. The total number of statistically significant (FDR q-value<0.05, P<4.7e-7) polyTE-gene expression associations (TE-eQTLs) for the different population cohorts, and different polyTE families, are shown in Figure 12A. Alu polyTE loci provide the greatest number of TE-eQTL by far, consistent with their substantially higher numbers in the genome. A quantile-quantile (Q-Q) plot for these data confirms a strong overall signal of statistically significant

associations (Figure 12B), which are shown along individual chromosomes, and broken down by polyTE family, in the Manhattan plot in Figure 12C. A complete list of the TE-eQTL discovered here is provided as Table 3.

A

		African (*n*=87)	European (*n*=358)	All (*n*=445)
Alu	all	357	155	251
	cis	3	37	41
	trans	354	118	210
L1	all	43	16	36
	cis	0	5	4
	trans	43	11	32
SVA	all	22	26	27
	cis	0	10	9
	trans	22	16	18
All	all	422	197	314
	cis	3	52	54
	trans	419	145	260
Gene		380	186	279

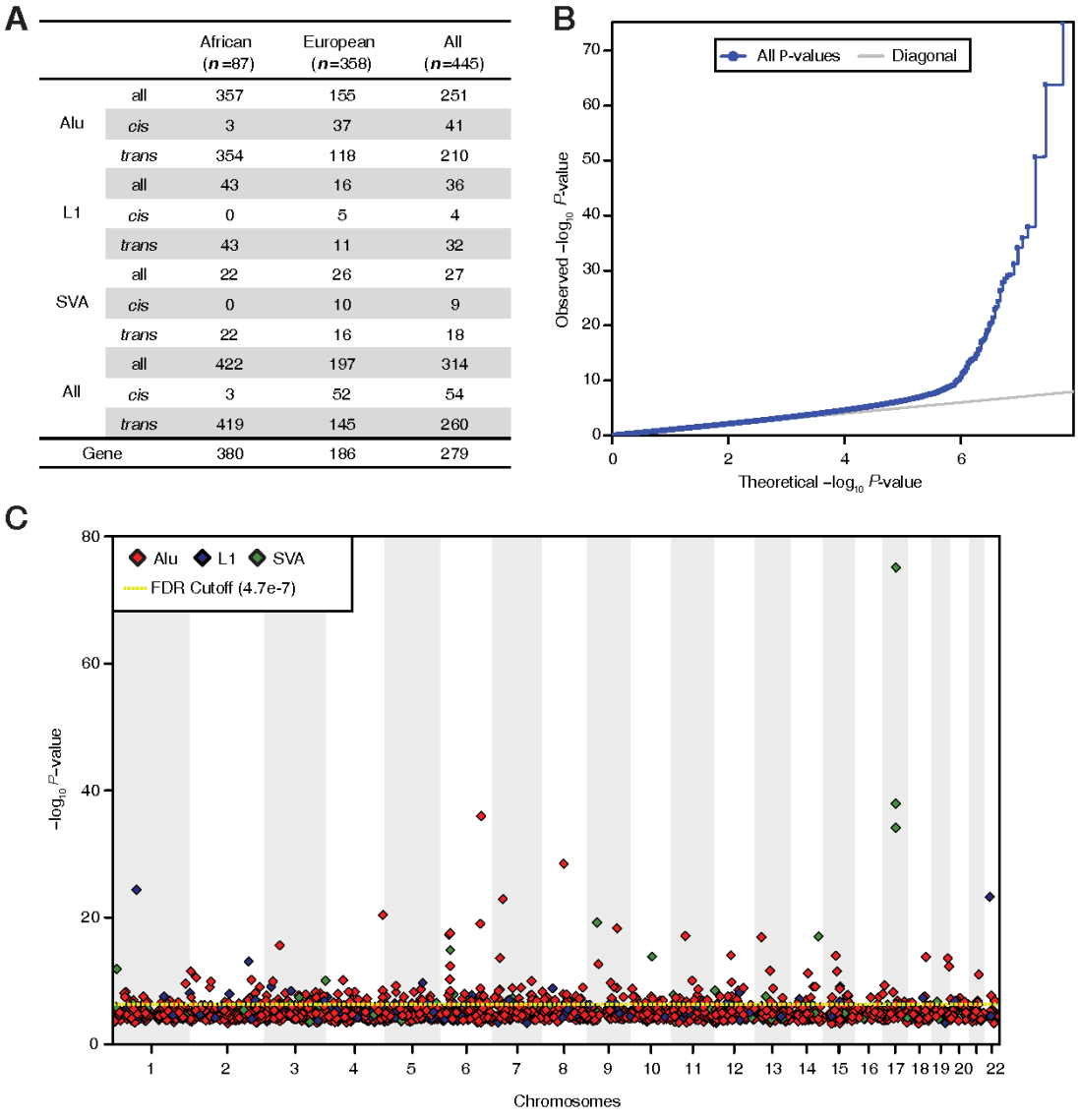


Figure 12 Polymorphic transposable element expression quantitative trait loci (TE-eQTL) detected

(A) Numbers of statistically significant TE-eQTL are shown for different population group cohorts and different polyTE families. The numbers of individuals (n) are shown for each

population cohort at the top of the table, and the total number of genes involved in TE-eQTL associations are shown at the bottom of the table. (B) Quantile-quantile (Q-Q) plot showing observed versus expected distributions of polyTE loci-gene expression association P-values (negative log transformed). (C) Manhattan plot showing the genomic distribution of polyTE-gene expression association values. The dashed yellow line indicates the FDR q-value cutoff of 0.05, which corresponds to a P-value of 4.7e-7. P-values are color coded according to polyTE families as shown in the key.

The set of genes that are associated with TE-eQTL is enriched for a number of immune-related functions including IgA production, antigen processing/presentation and several signalling pathways that lead to immune cell differentiation and activation (Figure 13 and Table 4). This result is consistent with the fact that the expression data were taken from lymphoblastoid cell lines (*i.e.,* transformed B-lymphocytes) and points to cell type-specific functional relevance of polyTE mediated gene regulation.

We performed a series of additional analyses in an effort to control for the potential effects of other genomic variations, regulatory SNPs in particular, on the TE-eQTL associations uncovered by our initial screen (see Materials and Methods; Figure 27). First, we assessed the extent of overlap between the genes that we observe to be associated with TE-eQTL here and genes previously found to be associated with SNPs using the same sequence and expression data. The overlap between the TE-eQTL genes identified here and the previously identified SNP-eQTL genes is extremely low (n=71 or ~1%), consistent with the fact that we are primarily identifying novel regulatory associations (Figure 28). Second, for those genes that were found to be associated with both TE-eQTL and SNP-eQTL, we performed conditional association analyses that combine both TE and SNP genotypes. The majority of the TE-eQTL from the initial screen remain significant after conditioning on the SNP genotypes (Table 5). Third, regional association scans were used

to evaluate the regulatory effects of all genomic variants linked to the TE-eQTL discovered here (Table 6). Examples of this analysis can be seen in the following section on population-specific TE-eQTLs.

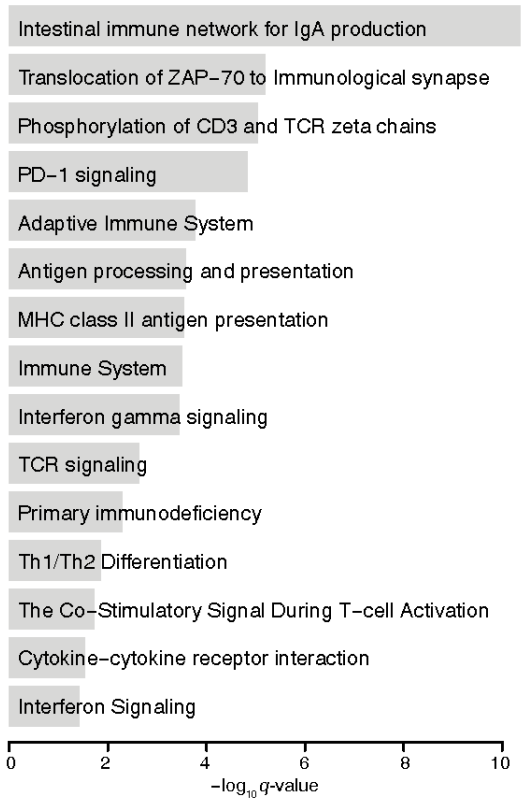


Figure 13 Functional enrichment of polyTE loci associated genes

Enriched, immune-related gene sets are shown along with the FDR q-values indicating the significance of the enrichments.

3.4.3 Population-specific TE-eQTL

A number of factors suggested the possibility that TE-eQTL may exert population-specific effects on human gene regulation. The polyTE landscapes of human populations

are very distinct, with polyTE genetic variation clearly delineating African from European populations. While gene expression does not show the same overall pattern of population divergence, there are hundreds of genes that do show population-specific expression. Finally, the numbers of TE-eQTL vary substantially for the African, European and merged population cohorts.

To evaluate the population-specific effects of TE-eQTL, we searched for gene-by-population interactions whereby specific polyTE loci are only associated with gene expression in the European or African populations (but not both). There are a total of 589 TE-eQTL that show such gene-by-population interactions: 407 for African and 182 for Europe (Figure 29). These apparent population-specific effects of TE-eQTL can be attributed to cases where the polyTE genotypes are differentially distributed across population groups (Figure 14A and B) or where polyTE genotypes are shared across populations but their effects on gene expression are limited to one population group (Figure 14C).

The polyTE locus Alu-5788 is strongly associated with *REL* expression levels when both population groups are considered together (Figure 14A). However, polyTE insertions at this locus are almost entirely African-specific and are associated with higher expression of the gene. Thus, consideration of the European population group alone would not turn up any association between this polyTE locus and the *REL* gene. *REL* encodes the c-Rel protein, which is part of the NF-κB family of transcription factors[122]. *REL* is considered to be a proto-oncogene that influences the survival and proliferation of B-lymphocytes. The gene's function has clinical significance with somatic mutations that are associated

with B-cell lymphomas[123] and SNPs that are associated with ulcerative colitis and rheumatoid arthritis[124, 125].

A similar kind of a population-specific TE-eQTL is seen for the Alu-10841 locus, which is associated with *PSD4* expression levels (Figure 14B). In this case, the presence of Alu insertions at the locus is associated with a reduction in gene expression levels. Alu insertions at this locus are far more common in European populations, and African individuals that lack the insertions tend to show higher expression levels for the gene. *PSD4* encodes a guanine nucleotide exchange factor that works with the ARF6, ARL14/ARF7 protein complex to control the movement of MHC class II containing vesicles along the actin cytoskeleton.

Gene-by-population interactions can also be seen for polyTE loci that are found in both the African and European population groups. While insertions at the Alu-1870 locus are commonly found in both population groups, polyTE insertion genotypes are only associated with decreased *PRDM2* expression in the African population (Figure 14C). The population-specific effects of Alu insertions at this locus could be attributable to the distinct genetic background of each population, via interactions with population-enriched variants for instance. On the other hand, insertions at the Alu-8559 locus are similarly found in both African and European populations, but both populations show polyTE insertion associations with decreased expression levels of the *HSD17B12* gene (Figure 14D). Interestingly, this particular example was detected with the FDR *q*-value cutoff employed here (0.05) for the both the European and merged population cohorts but not for the African population alone (P=8.7e-5 and FDR q-value=3.9e-1). This may be attributable to the relatively low number of human samples analyzed for the single African population and

suggests the possibility that some *bona fide* African-specific associations may have been overlooked in this study.

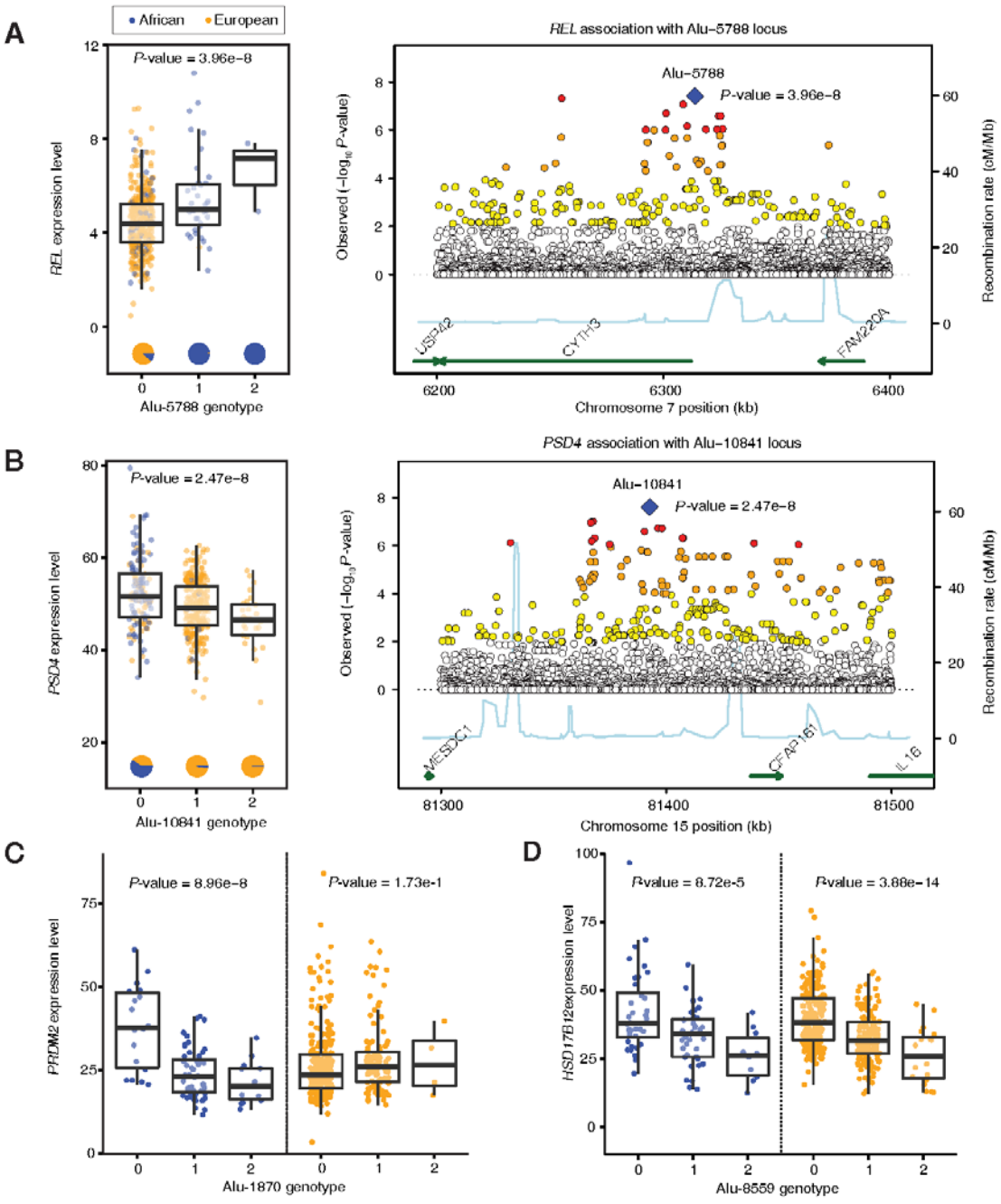


Figure 14 Examples of population-specific TE-eQTL detected

Population-specific TE-eQTL where polyTE insertions are found primarily in only one population group are shown for the (A) REL and (B) PSD4 genes. Box-plots show the distributions of individual gene expression levels for each of the three possible polyTE insertion genotypes. Regional association plots show all associations with the gene expression centered on the polyTE locus. Association P-values are shown as indicated on the left y-axis along with the local recombination rate shown on the right y-axis. (C) A population-specific TE-eQTL is shown for PRDM2 gene where the associated polyTE locus has insertions found in both population groups but an association with gene expression is only seen in the African population. (D) A counter example of a polyTE locus with insertions shared among both population groups and similar associations with HSD17B12 are seen for both groups.

3.4.4 Transcriptional network TE-eQTLs

We found a number of cases where polyTE loci corresponded to TE-eQTL for more than one human gene (Figure 12A and Table 3). This suggested the possibility that individual polyTE loci may participate in coordinated gene regulatory networks. One possible mechanism by which this may occur is through indirect polyTE-expression associations that are mediated by transcription factors (TFs), which regulate the expression of multiple genes. In other words, if a polyTE loci affects the expression of a TF, it may also appear to affect the regulation of one or more gene targets of that TF (Figure 15A). The Alu-7481 locus exemplifies this phenomenon. Alu insertions at this locus are associated with increased expression of *PAX5* (Figure 15B), which encodes a transcription factor crucial to the specific identity and function on B cells. In particular, *PAX5* expression is critical for differentiation of lymphoid progenitor cells into B cells (Figure 15C). It achieves this by simultaneously activating B lineage-specific genes and repressing genes active in distinct lineages[126]. There are 274 known Pax5 target genes that show the identical Alu-7481 insertion genotype expression pattern as seen for their cognate TF (Figure 15D and Table 7). While the majority of these do not reach the FDR *q*-value cutoff

used here, there are three immune related target genes – *PIK3AP1*, *REL* and *ZSCAN23* – that all remain statistically significant after controlling for multiple tests (Figure 15E). These data suggest that polyTE insertions are also involved in establishing cell type-specific regulatory networks with phenotypically important consequences.

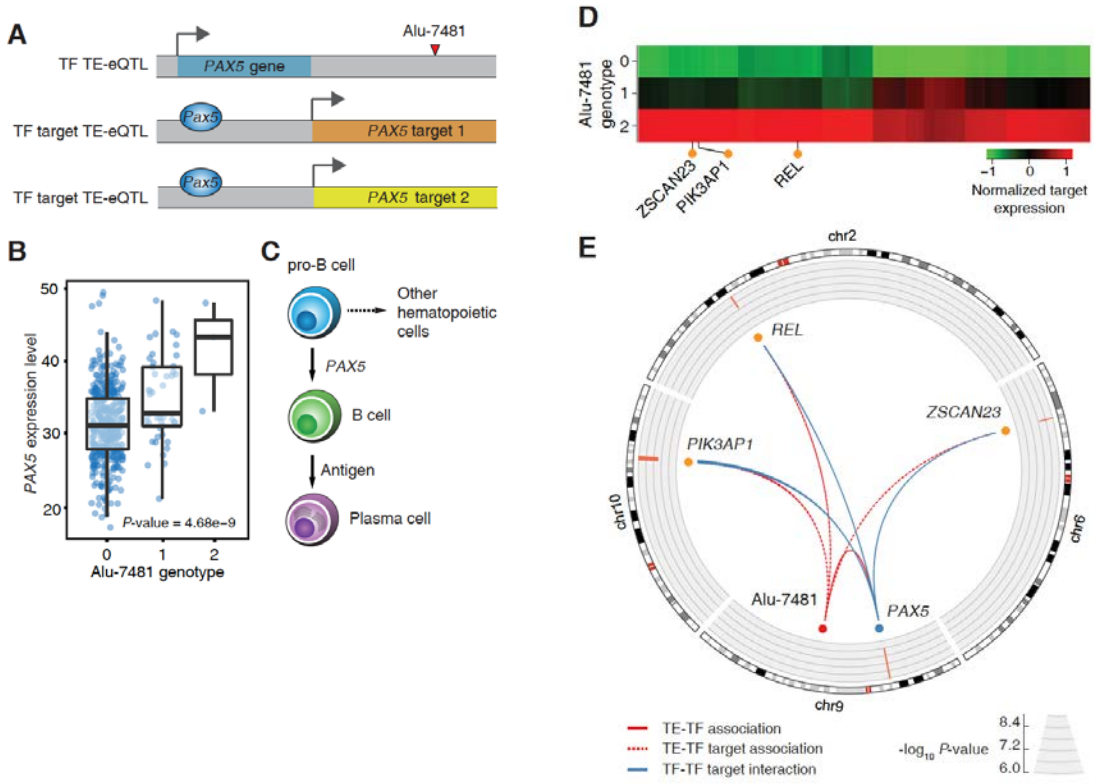


Figure 15 TE-eQTL and a *PAX5* transcriptional regulatory network

(A) Scheme for how a single polyTE loci can provide trans eQTL for multiple genes by modifying the expression of a transcription factor encoding gene (PAX5) and its downstream target genes. (B) The Alu-7481 PAX5 TE-eQTL. PAX5 expression levels are shown for individuals with different Alu-7481 insertion genotypes (0, 1 or 2 insertions); the association P-value is shown. (C) The role of PAX5 in B-cell development. (D) Average expression level of 274 PAX5 target genes for individuals with different Alu-7481 insertion genotypes. Normalized (z-score transformed) gene expression levels are color-coded as shown in the key. Target genes that correspond to the most significant Alu-7481 TE-eQTL (FDR q-value<0.05, P<4.7e-7) are indicated. (E) Circos plot showing the chromosomal locations of the Alu-7481 TE-eQTL, PAX5 and the downstream target genes. TE-gene associations are shown in red, and PAX5-target gene interactions are shown in blue. The association P-values are shown on the inner circle as indicated in the key.

3.5 Discussion

Numerous previous studies have uncovered gene regulatory contributions of human TE sequences[17, 19-26, 28-32, 104-106]. However, these studies have dealt with TE sequences derived from relatively ancient insertion events, which now are fixed in human populations. In other words, these TE-derived sequences exist at the same genomic locations in all human genomes, and thus may not contribute substantially to regulatory variation between individuals. Here, we present a systematic analysis of the regulatory contributions of polyTE loci that were generated by recent transpositional activity and thereby differ between individuals. The TE-eQTL that we discovered underscore the extent to which TE-generated human genetic variation can affect regulatory differences within and between populations.

Our results indicate that polyTE loci provide greater numbers of trans compared to cis eQTL (Figure 12A). This may be considered somewhat surprising given the fact that most human eQTL studies focus on cis eQTL[109, 127]. However, studies on eQTL are often limited to cis associations owing the large number of possible SNP-by-gene comparisons that whole genome (*i.e.*, both cis and trans) analyses entail. Thus, it is not entirely clear whether cis eQTL are actually expected to be more common that trans eQTL. The relatively low number of polyTE loci studied here (~16,000), combined with the introduction of a more computationally efficient eQTL detection algorithm[114], allowed us to evaluate all possible cis and trans TE-eQTL. There are several possible mechanisms by which polyTE loci could serve as trans eQTL. For example, they may exert trans eQTL effects indirectly by regulating transcription factors, which in turn regulate numerous target genes, as we have shown for *PAX5* (Figure 15). In addition, TEs have been shown to

influence three-dimensional genome architecture, via the formation of chromosome loops or association with the nuclear scaffold/matrix for instance[20]. It is tempting to speculate that TEs can exert trans eQTL effects via similar mechanisms that bring distal, homologous TE sequences into close proximity.

It is worth noting that human TE activity has often been associated with disease[34, 35]. Indeed, transpositional activity of human TEs was confirmed via the discovery of *de novo* insertions with obvious effects on health[6]. However, the samples analyzed here correspond to (presumably) healthy individuals from the 1000 Genomes Project and are thereby taken to represent the normal scope of human genetic variation. The fact that many of the polyTE loci analyzed here have accumulated to relatively high insertion allele frequencies (Figure 10) is consistent with the notion that they are not deleterious. Thus, the phenotypic impact of human TE activity is not limited to deleterious effects; it also includes the generation of regulatory differences that fall within the scope of naturally occuring human variation. These kinds of functionally relevant but subtle TE-genetic variations, which necessarily avoid elimination by purifying selection, may provide an important substrate for ongoing human evolution.

CHAPTER 4. HUMAN RETROTRANSPOSON INSERTION POLYMORPHISMS ARE ASSOCIATED WITH HEALTH AND DISEASE VIA GENE REGULATORY PHENOTYPES

4.1 Abstract

The human genome hosts several active families of transposable elements (TEs), including the Alu, LINE-1, and SVA retrotransposons that are mobilized via reverse transcription of RNA intermediates. We evaluated how insertion polymorphisms generated by human retrotransposon activity may be related to common health and disease phenotypes that have been previously interrogated through genome-wide association studies (GWAS). To address this question, we performed a genome-wide screen for retrotransposon polymorphism disease associations that are linked to TE induced gene regulatory changes. Our screen first identified polymorphic retrotransposon insertions found in linkage disequilibrium (LD) with single nucleotide polymorphisms (SNPs) that were previously implicated in common complex diseases by GWAS. We further narrowed this set of candidate disease associated retrotransposon polymorphisms by identifying insertions that are located within tissue-specific enhancer elements. We then performed expression quantitative trait loci (eQTL) analysis on the remaining set of candidates in order to identify polymorphic retrotransposon insertions that are linked to gene expression changes in B-cells of the human immune system. This progressive and stringent screen yielded a list of six retrotransposon insertions as the strongest candidates for TE polymorphisms that lead to disease via enhancer-mediated changes in gene regulation. For example, we found an SVA insertion within a cell-type specific enhancer located in the second intron of the *B4GALT1* gene. *B4GALT1* encodes a glycosyltransferase that functions in the glycosylation of the Immunoglobulin G (IgG) antibody in such a way as

to convert its activity from pro- to anti-inflammatory. The disruption of the *B4GALT1* enhancer by the SVA insertion is associated with down-regulation of the gene in B-cells, which would serve to keep the IgG molecule in a pro-inflammatory state. Consistent with this idea, the *B4GALT1* enhancer SVA insertion is linked to a genomic region implicated by GWAS in both inflammatory conditions and autoimmune diseases such as systemic lupus erythematosus and Crohn's disease. We explore this example and the other cases uncovered by our genome-wide screen in an effort to illuminate how retrotransposon insertion polymorphisms can impact human health and disease by causing changes in gene expression.

4.2 Introduction

At least one half of the human genome sequence is derived from the replication and insertion of retrotransposons – RNA agents that transpose among chromosomal locations via the reverse transcription of RNA intermediates [1, 2]. The vast majority of retrotransposon-related sequences in the human genome are derived from ancient insertion events and are no longer capable of transposition. Nevertheless, there are several families of human retrotransposons that remain active. The most abundant active families of human retrotransposons are the Alu [3, 4], LINE-1 (L1) [5, 6], and SVA [7, 8] retrotransposons; recent evidence indicates that a smaller number of HERV-K endogenous retroviruses also remain capable of transposition [54].

Sequences from active retrotransposon families generate insertional polymorphisms within and between human populations by means of germline transposition events. In this way, ongoing retrotranspositional activity of these RNA agents serves as an important source of human genetic variation. Retrotransposons are further distinguished by the fact that they are known to impact the regulation of human genes in a number of

different ways [17, 18, 128]. Nevertheless, the joint phenotypic implications of retrotransposon generated human genetic variation, coupled with their capacity for genome regulation, have yet to be fully explored. We previously studied the implications of somatic retrotransposition for the etiology of cancer vis a vis retrotransposon induced regulatory changes in tumor suppressor genes [129]. For the current study, we were curious to understand how insertion polymorphisms generated by human retrotransposon activity may be related to commonly expressed health and disease phenotypes.

In one sense, a link between retrotransposon activity and disease is already well established. Active human retrotransposons were originally discovered due to the deleterious effects of element insertions [6]. There are 124 genetic diseases that have been demonstrated to be caused by retrotransposon insertions, including cystic fibrosis (Alu), hemophilia A (L1) and X-linked dystonia-parkinsonism (SVA) [34, 67]. However, these cases represent so-called Mendelian diseases caused by very deleterious mutations that are expressed with high penetrance. Disease causing mutations of this kind are extremely rare and do not segregate among populations as common genetic polymorphisms. Complex multi-factorial diseases, on the other hand, are associated with more common genetic variants that exert their effects in a probabilistic as opposed to a deterministic manner. The contribution of common retrotransposon polymorphisms to complex health and disease related phenotypes has yet to be systematically explored.

Given the known connection between retrotransposon activity and genetic disease, we hypothesized that retrotransposon insertion polymorphisms may also contribute to inter-individual phenotypic differences that are associated with common diseases that have complex, multi-factorial genetic etiology. Since we previously showed that retrotransposon insertions contribute to inter-individual and population-specific differences in human gene regulation [130], we also hypothesized that the impact of

retrotransposon insertion polymorphisms on human health could be mediated by gene regulatory effects.

Previously, it has only been possible to investigate the impact of retrotransposon polymorphisms on disease phenotypes for a limited number of individuals owing to the number of genomes that were available [131]. For the current study, we leveraged the accumulation of whole genome sequence and expression datasets, along with data on single nucleotide polymorphism (SNP) disease-associations, in order to perform a population level genome-wide screen for retrotransposon polymorphisms that are linked to complex health- and disease-related phenotypes.

4.3 Materials and Methods

4.3.1 Polymorphic transposable element (polyTE) and SNP genotypes

Human polymorphic TE (polyTE) insertion presence/absence genotypes for whole genome sequences of 445 individuals from five human populations were accessed from the phase 3 variant release of the 1000 Genomes Project (1KGP) [99]. Whole genome SNP genotypes were taken for the same set of individuals. The phase 3 variant release corresponds to the human genome reference sequence build GRCh37/hg19, and the 5 human populations are YRI: Yoruba in Ibadan, Nigeria from Africa, CEU: Utah Residents (CEPH) with Northern and Western European Ancestry, FIN: Finnish in Finland, GBR: British in England and Scotland and TSI: Toscani in Italia from Europe. We chose these genome sequence datasets because they have matching RNA-seq data for the same individuals (see eQTL analysis section). The YRI population was taken to represent the African continental population group (AFR), and the 4 populations from Europe (CEU, FIN, GBR and TSI) were grouped together as the European (EUR) continental population

group for downstream analysis (Figure 16). PolyTE insertion genotypes were characterized by the 1KGP Structural Variation Group using the program MELT as previously described [65]. Previously, we performed an independent validation the performance of this program for human polyTE insertion variant calling from whole genome sequences [64]. The polyTE genotype data were downloaded via the 1000 Genomes Project ftp hosted by the NCBI: http://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/release/20130502/. For a given polyTE insertion site in the genome, there are three possible presence/absence genotype values for an individual genome: 0-no polyTE insertion (homozygote absent), 1-a single polyTE insertion (heterozygote), and 2-two polyTE insertions (homozygote present). PolyTE genotypes were used for eQTL analysis as described in section 2.5. For each of the two continental population groups, only polyTE insertions and SNPs with greater than 5% minor allele frequencies (MAF) were used for the downstream analysis to ensure both the confidence of genotype calls and the reliability of the association analyses. Minor alleles for TEs are assumed to be the insertion present allele, since the ancestral state for any polyTE insertion site corresponds to the absence of an insertion [121].

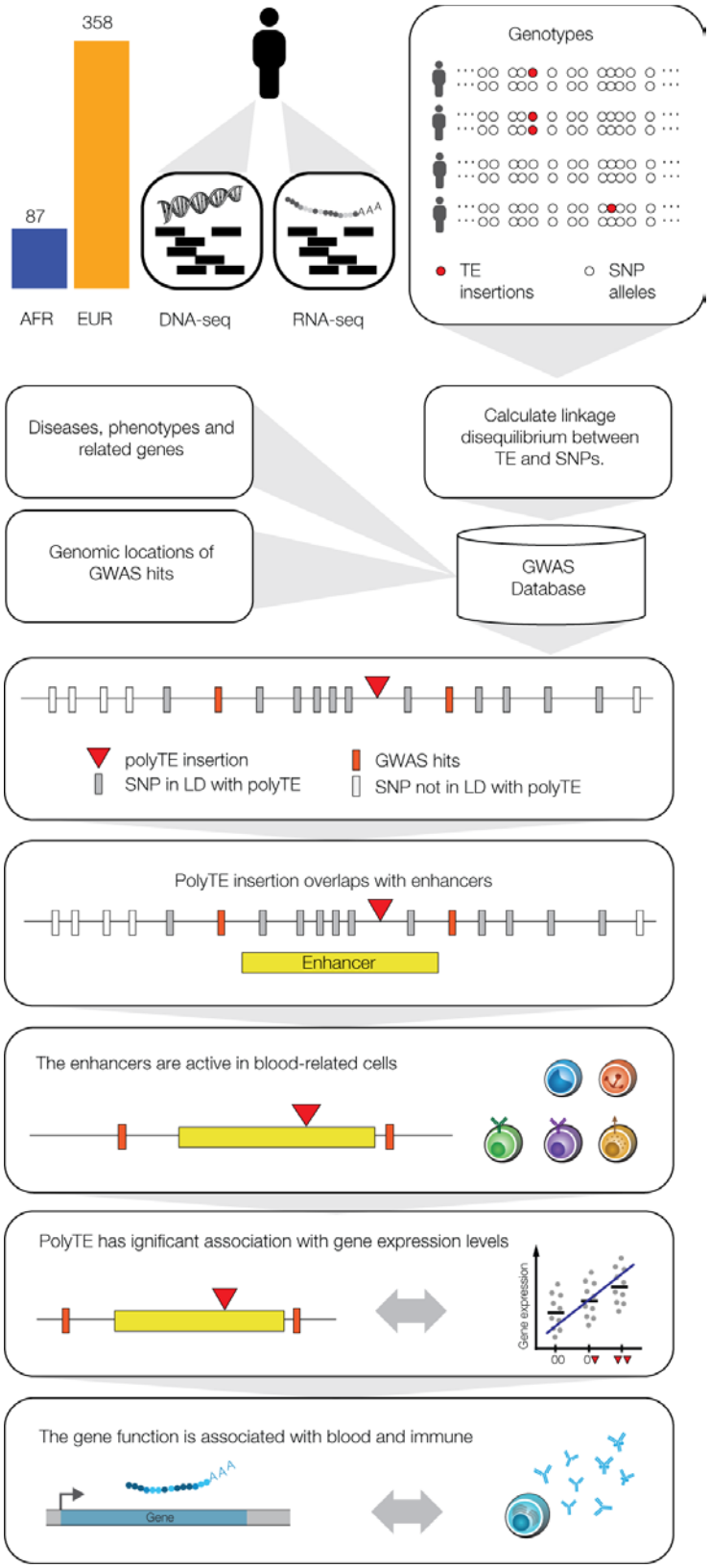


Figure 16 Integrative data analysis used to screen for polyTE disease-associations.

Matched DNA-seq and RNA-seq samples were analyzed for 445 individuals, 87 Africans (AFR - blue) and 358 Europeans (EUR - orange). Genome-wide genotypes were characterized for polyTE insertions (presence/absence) and SNPs, and the linkage disequilibrium (LD) structure for polyTE insertions and SNP alleles was characterized for all samples. The NHGRI-EBI GWAS database was mined to extract SNP disease-associations and related information on diseases, phenotypes, genes and SNP genomic locations. A series of filters was applied to screen for a set of high-confidence polyTE disease-associations. PolyTEs were evaluated for: (1) minor allele frequency (MAF), (2) linkage to disease-associated SNPS (i.e., GWAS hits), (3) overlap with tissue-specific enhancers, (4) associations with gene expression, and (5) functional relevance for blood- and immune system-related diseases.

4.3.2 PolyTE-SNP linkage analysis

The GCTA program (version 1.25.0) was used to estimate the linkage disequilibrium (LD) structure for polyTEs and SNPs in genomic regions centered at each polyTE insertion site. For each polyTE insertion site, pairwise correlations (r) between the target polyTE insertion alleles and all SNP alleles in the same LD block were computed across all individual genome samples. A correlation (r) significance P-value threshold of 0.05 was used to identify all SNPs considered to be in LD with each polyTE insertion. Pairwise distances between polyTE insertion sites and linked SNPs were calculated as the number of base pairs between each polyTE insertion site and all linked SNP locations.

4.3.3 Genome wide association studies (GWAS) for disease

Associations between human genetic variants (SNPs) and health- or disease-related phenotypes were explored using the NHGRI-EBI GWAS database [132]. GWAS database SNPs with genome-wide association values of $P<10^{-5}$ were taken for analysis, and the genomic location, specific health- or disease-related phenotype, identity of the risk allele and original reporting publications were recorded for each associated SNP. The GWAS

SNPs were screened for LD with polyTE insertions as described in section 4.3.2 to yield a set of candidate disease-linked polyTE insertions for further analysis.

4.3.4 Evaluating polyTE regulatory potential

The regulatory potential for polyTE insertions was evaluated by considering their co-location with known enhancer sequences. Active enhancers for 127 cell-types and tissues were characterized by the Roadmap Epigenomics Project using the ChromHMM program [133, 134]. ChromHMM integrates multiple genome-wide chromatin datasets (*i.e.*, epigenomes), such as ChIP-seq of various histone modifications, using a multivariate Hidden Markov Model to identify the locations of tissue-specific enhancers based on their characteristic chromatin states. The data files with genomic locations for enhancers across all 127 epigenomes were accessed through the project website at http://mitra.stanford.edu/kundaje/leepc12/web_portal/chr_state_learning.html. The genomic locations of polyTE insertions that are in LD with disease-associated SNPs were compared with the genomic locations for enhancers from the 127 epigenomes, and polyTE insertions found to be located within active enhancer elements were considered to have regulatory potential. A subset of 27 epigenomes characterized for cells and tissues related to blood and the immune system – such as T cells, B cells and hematopoietic stem cells – were selected for downstream eQTL analysis (see section 4.3.5).

The overall relative regulatory potential of polyTE insertions in a given epigenome i is quantified as:

$$r_i = \frac{t_i}{s_i}$$

where t_i is the proportion of polyTEs that are co-located with an enhancer element in a given epigenome i, and s_i is the proportion of SNPs from polyTE LD blocks that overlap with an enhancer element in the same epigenome i.

Physical associations between TE-enhancer insertions and nearby gene promoter regions were evaluated with chromatin-chromatin interaction map data, based on several different data sources, including 4C, 5C, ChIA-PET and Hi-C, using the Chromatin Chromatin Space Interaction (CCSI) database at http://songyanglab.sysu.edu.cn/ccsi/ [135].

4.3.5 Expression quantitative trait locus (eQTL) analysis

Associations between polyTE insertion genotypes and tissue-specific gene expression levels were characterized using eQTL analysis (Figure 16). PolyTE insertion presence/absence genotypes were characterized as described in section 2.1. RNA-seq gene expression data for the same 445 individual genome samples used for polyTE genotype characterization were taken from the GEUVADIS RNA sequencing project. Genome-wide expression levels were measured for the same lymphoblastoid cell lines, *i.e.* Epstein–Barr virus (EBV) transformed B-lymphocytes (B cells), as used for DNA-seq analysis in the 1KGP. RNA isolation, library preparation, sequencing and read-to-genome mapping were performed as previously described [109]. As with the polyTE genotype data, the RNA-seq reads were mapped to the human genome build GRCh37/hg19. The process of gene expression normalization and quantification based on these RNA-seq data has been extensively validated as part of the GEUVEDIS project [112]. The GEUVADIS RNA-seq data were used to compute gene expression levels for ENSEMBL gene models as previously described [110]. Normalization of gene expression levels was done using a combination of a modified reads per kilobase per million mapped reads (RPKM) approach

followed by the probabilistic estimation of expression residuals (PEER) method as previously described [111]. This procedure has been shown to eliminate batch effects among different RNA-seq samples and to reduce the overall variance across samples, thereby ensuring the most accurate and comparable gene expression level inferences among samples. The normalized gene expression levels were accessed from the GUEVADIS project ftp server hosted at the EBI:

ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/analysis_results/.

PolyTE insertions that are (1) linked to at least one disease-associated SNP, and (2) located within a blood- or immune system-related enhancer were taken as a candidate set for eQTL analysis with the lymphoblastoid cell line RNA-seq data. PolyTE insertion presence/absence genotypes were regressed against gene expression levels to identify eQTLs (TE-eQTLs) using the program Matrix eQTL [114]. Matrix eQTL was run using the additive linear (least squares model) option with gender and population used as covariates. This was done for all possible pairs of polyTE insertion sites from the candidate set and all genes. *Cis* versus *trans* TE-eQTLs were defined later as polyTE insertion sites that fall inside (*cis*) or outside (*trans*) 1 megabase from gene boundaries. *P*-values were calculated for all TE-eQTL associations, and FDR *q*-values were then calculated to correct for multiple statistical tests. The genome-wide significant TE-eQTL association threshold was set at FDR $q<0.05$, corresponding to $P=4.7 \times 10^{-7}$ (AFR) and $P=2.6 \times 10^{-7}$ (EUR).

4.3.6 Interrogation of disease-associated gene function and association consistency

The potential functional impacts of disease-associated TE-eQTL were evaluated via comparison of annotated gene functions and reported GWAS phenotypes for polyTE-

linked SNPs. Gene functions were taken from the NCBI Entrez gene summaries, and GWAS phenotypes were taken from the original literature where the associations were reported. Genes that were found to be functionally related to GWAS reported health- or disease-related phenotypes were further checked for the direction of association. If the GWAS SNP-gene pair shows the same direction of association as the polyTE-gene pair, then the pair was included in the final set of significant gene-polyTEs association pairs (Table 2). For each gene in the final set, its tissue-specific expression levels across 18 tissues, including 4 blood- and immune-related tissues, were taken from the Illumina BodyMap and GTEx projects [110, 136, 137].

4.4 Results

We used a genome-scale data analysis approach to explore the potential impact of human genetic variation generated by the activity of TEs on health and disease (Figure 16). This approach entailed an integrative analysis of (1) TE insertion polymorphisms, (2) single nucleotide polymorphisms (SNPs), (3) SNP-disease associations, (4) tissue-specific enhancers, (5) expression quantitative trait loci (eQTL), and (6) gene function/expression profiles. The rationale behind this approach was to employ a series of successive genome-wide filters, which would converge on a set of high-confidence TE insertion polymorphisms that are most likely to impact health- or disease-related phenotypes. Our analysis started with 5,845 polyTE insertions, with minor allele frequencies (MAF)>0.05 for two continental population groups (European and African), and converged on a final set of seven high-confidence TE disease-association candidates (Figure 17). The final set of seven health/disease-implicated TE insertion polymorphisms that we found are distinguished by their linkage to disease-associated SNPs as well as their regulatory and

functional properties. We describe the results and implications for each step in our TE disease-association screen in the sections below.

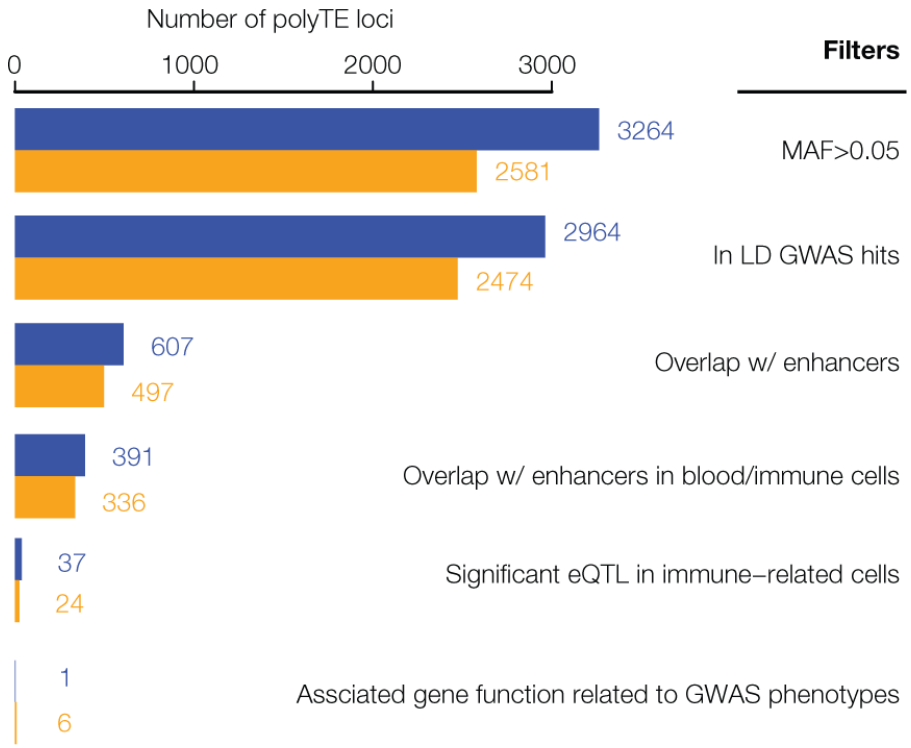


Figure 17 Results of the genome-wide screen for polyTE disease-associations

As illustrated in Figure 16, a series of filters was applied to screen for a final set of high-confidence polyTE disease-associations. The number of polyTE insertions that remain after the application of each successive filter is shown for the African (AFR – blue) and European (EUR – orange) population groups.

4.4.1 Linkage disequilibrium for polyTEs and disease-associated SNPs

The genomic locations of polyTE insertions were characterized for 445 individuals from one African (AFR) and four European (EUR) populations as described in section 2.1 of the Materials and Methods. This was done for the most common families of active human TEs: Alu, L1 and SVA. For each polyTE insertion location, individual genotypes were characterized as homozygous absent (0), heterozygous (1), or homozygous present

(2). The distributions of polyTE insertion genotypes among individuals from each population were used to screen for polyTEs that are found at relatively high MAF>0.05. The linkage disequilibrium (LD) structure of the resulting common polyTE insertions, with adjacent common SNPs (also at MAF>0.05), was then defined using correlation analysis across individual genome samples (Materials and Methods section 4.3.2). In addition, the genomic locations of common polyTE insertion variants and their linked SNPs were compared to the locations of disease-associated SNPs reported in the NHGRI-EBI GWAS database (Materials and Methods section 4.3.3). Linkage correlation coefficients between all polyTE insertions analysed here and GWAS SNPs are shown in Table 8 and Table 9.

Distributions of LD correlations between polyTEs and adjacent SNPs were compared separately for non-disease-associated versus disease-associated SNPs. For all three families of active human TEs, in both the AFR and EUR population groups, polyTEs are found in significantly higher LD with disease-associated SNPs compared to non-disease-associated SNPs (Figure 18A). In addition, polyTE variants are located closer to disease-associated SNPs than non-disease-associated SNPs for the EUR population group (Figure 18B). A similar enrichment was not seen for the AFR population group, which may be attributed to the lower number of samples available for analysis for this group (AFR=87 vs. EUR=358). Indeed, when a larger number of AFR samples, which do not have matched RNA-seq data, were used for the same linkage analysis, the results were qualitatively identical to those seen for the EUR samples analysed here. Taken together, these results indicate that polyTEs are more likely to be tightly linked to disease-associated SNPs compared to adjacent linked SNPs from the same LD blocks, suggesting a possible role in disease etiology for some TE variants. This is notable in light of the facts that (1) the TE genotypes were not considered in the initial association studies, and (2) TE insertions entail substantially larger-scale genetic variants than SNPs. Thus, polyTEs

found on the same haplotypes as disease-associated SNPs may be expected to have an even greater impact on health- and disease-outcomes in some cases.

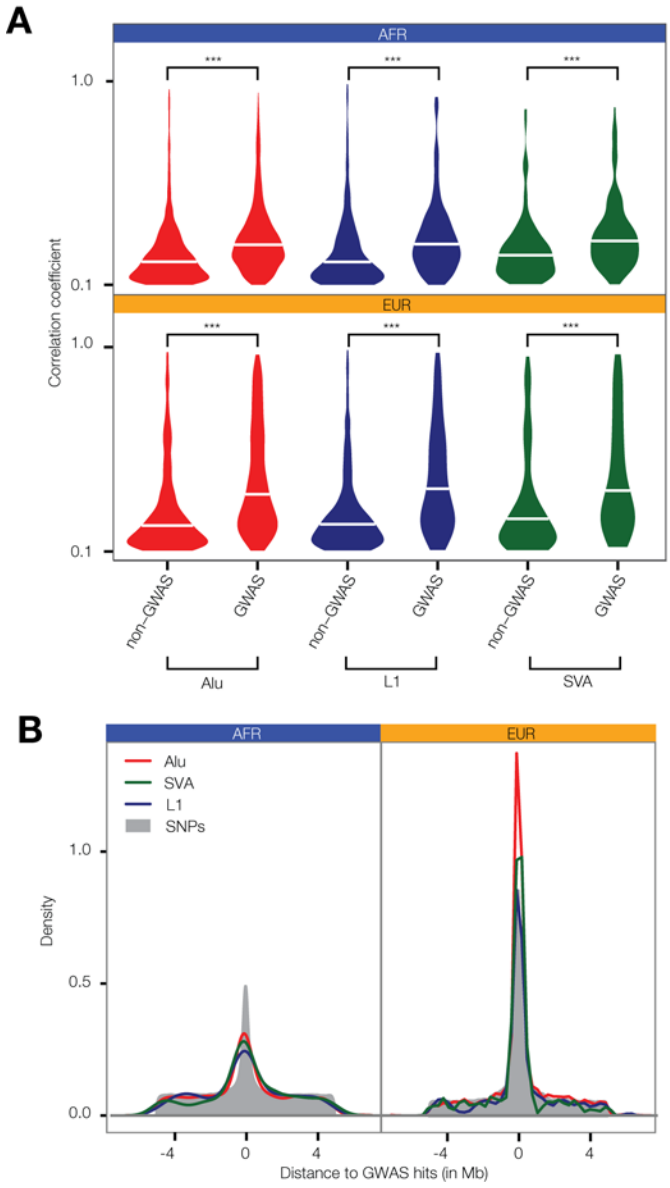


Figure 18 Linkage between polyTE insertions and SNP disease-associations from GWAS.

*(A) Distributions of the allele correlation coefficients (r) are shown for (1) polyTE insertions and non-GWAS SNPs, and (2) polyTE insertions and GWAS SNPs. Higher correlation coefficients indicate tighter linkage. The significance of the differences for the non-GWAS SNP versus the GWAS SNP correlation coefficient distributions are indicated (Kolmogorov-Smirnov test; *** = $P<2.4 \times 10^{-5}$). (B) Density distributions of the distance*

between polyTE insertions and SNPs to the nearest GWAS disease-associated SNP. Correlation coefficient (A) and density (B) distributions are shown separately for the Alu (red), L1 (blue) and SVA (green) TE families in the African (AFR – blue) and European (EUR – orange) population groups.

4.4.2 Co-location of disease-linked polyTEs with tissue-specific enhancers

Given the fact that TEs are known to participate in human gene regulation via a wide variety of mechanisms [17, 18, 128], we hypothesized that polyTEs may impact disease by virtue of gene regulatory effects. The regulatory potential of polyTEs linked to disease-associated SNPs was first evaluated by searching for insertions that are co-located with tissue-specific enhancers. The locations of enhancers were characterized for 127 cell- and tissue-types based on their chromatin signatures as described in section 2.4 of the Materials and Methods. An example of an enhancer co-located with a disease-linked polyTE is shown for an Alu element that is inserted 5' to the Immunoglobulin Heavy Variable 2-26 (*IGHV2-26*) encoding gene (Figure 19A). We found a total 607 disease-linked polyTEs co-located with enhancers in the AFR population group and 437 in EUR group; 391 (AFR) and 336 (EUR) of those enhancers correspond to blood- or immune-related tissues (Figure 17). Details on the co-localization of polyTE insertions and the enhancers characterized for each epigenome are shown in online Supplementary Data Table S2.

We estimated the overall regulatory potential for disease-linked polyTEs in all cell- and tissue-types by computing the relative ratio of enhancer co-located insertions as described in section 2.4 of the Materials and Methods. The results of this analysis are considered separately for the blood- and immune-related tissues (Figure 19B) and all other tissues from which enhancers were characterized. Enhancer co-located disease-linked polyTEs from blood/immune cell-types show higher overall regulatory potential than ones

that are co-located with enhancers characterized for the other tissue-types (Figure 12C and 4D). These results suggest that the set of disease-linked polyTEs studied here may have a disproportionate impact on immune-related diseases, and we focused our subsequent efforts on this subset of health conditions.

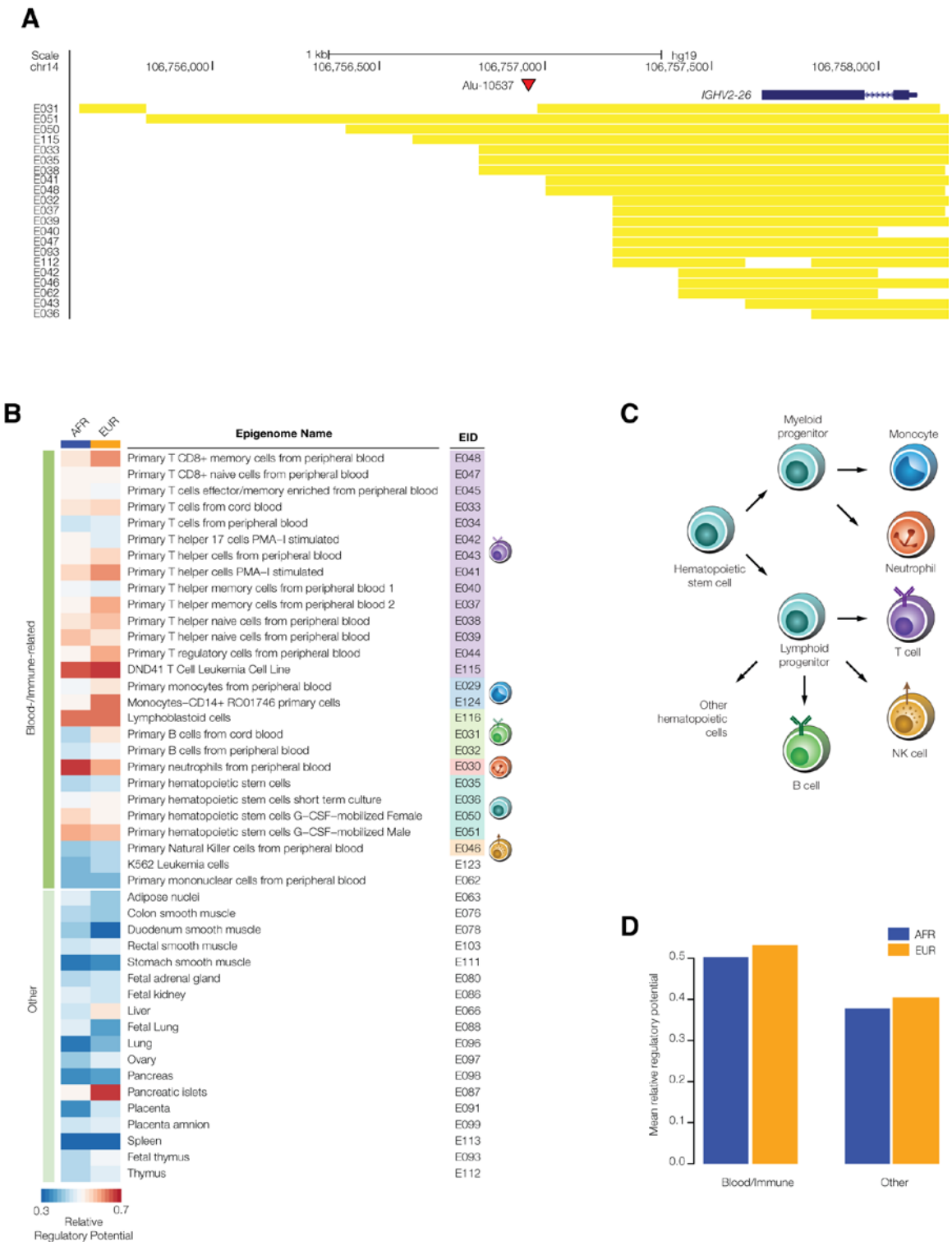


Figure 19 Regulatory potential of disease-linked polyTE insertions.

PolyTE insertions linked to disease-associated SNPs were evaluated for their co-location with tissue-specific enhancers. (A) UCSC Genome Browser screen capture showing an

example of a polyTE insertion – Alu-10537 – that overlaps with a number of tissue specific enhancers. The genomic location of the Alu insertion on chromosome 14, downstream of the IGHV2-26 gene (blue gene model), is indicated with a red arrow. The genomic locations of co-located enhancers, characterized based on chromatin signatures from a variety of tissue-specific epigenomes, are indicated with yellow bars. (B) Heatmap showing the relative regulatory potential (Materials and Methods section 2.4) of polyTE insertions for a variety of tissue-specific epigenomes. Blood- and immune-related tissues are shown separately from examples of the other tissue types analyzed here. (C) Developmental lineage of immune-related cells for which enhancer genomic locations were characterized. (D) The mean relative regulatory potential for disease-linked polyTE insertions is shown for blood- and immune-related tissues compared to all other tissue-types analyzed here. Values for the African (AFR – blue) and European (EUR – orange) population groups are shown separately.

4.4.3 Expression associations for disease-linked and enhancer co-located polyTEs

We further evaluated the regulatory potential of the polyTEs that were found to be both disease-linked and co-located with blood- and immune-related enhancers using an expression quantitative trait loci (eQTL) approach (see Materials and Methods section 2.5). Genotypes for this subset of polyTEs from the 445 genome samples analyzed here were regressed against gene expression levels characterized from lymphoblastoid cell lines for the same individuals. Quantile-quantile (Q-Q) plots comparing the observed versus expected *P*-values for the e-QTL analysis in the AFR and EUR population groups are shown in Figure 20A, revealing a number of statistically significant associations that are likely to be true-positives. There are 83 (AFR) and 42 (EUR) genome-wide significant TE-eQTL (Table 10 and Table 11), and they are enriched in genomic regions that encode immune-related genes (Figure 20B). We narrowed this list further by selecting the strongest TE-eQTL association for each individual polyTE, resulting in a final list of 37 (AFR) and 24 (EUR) immune-related TE-eQTL (Figure 17).

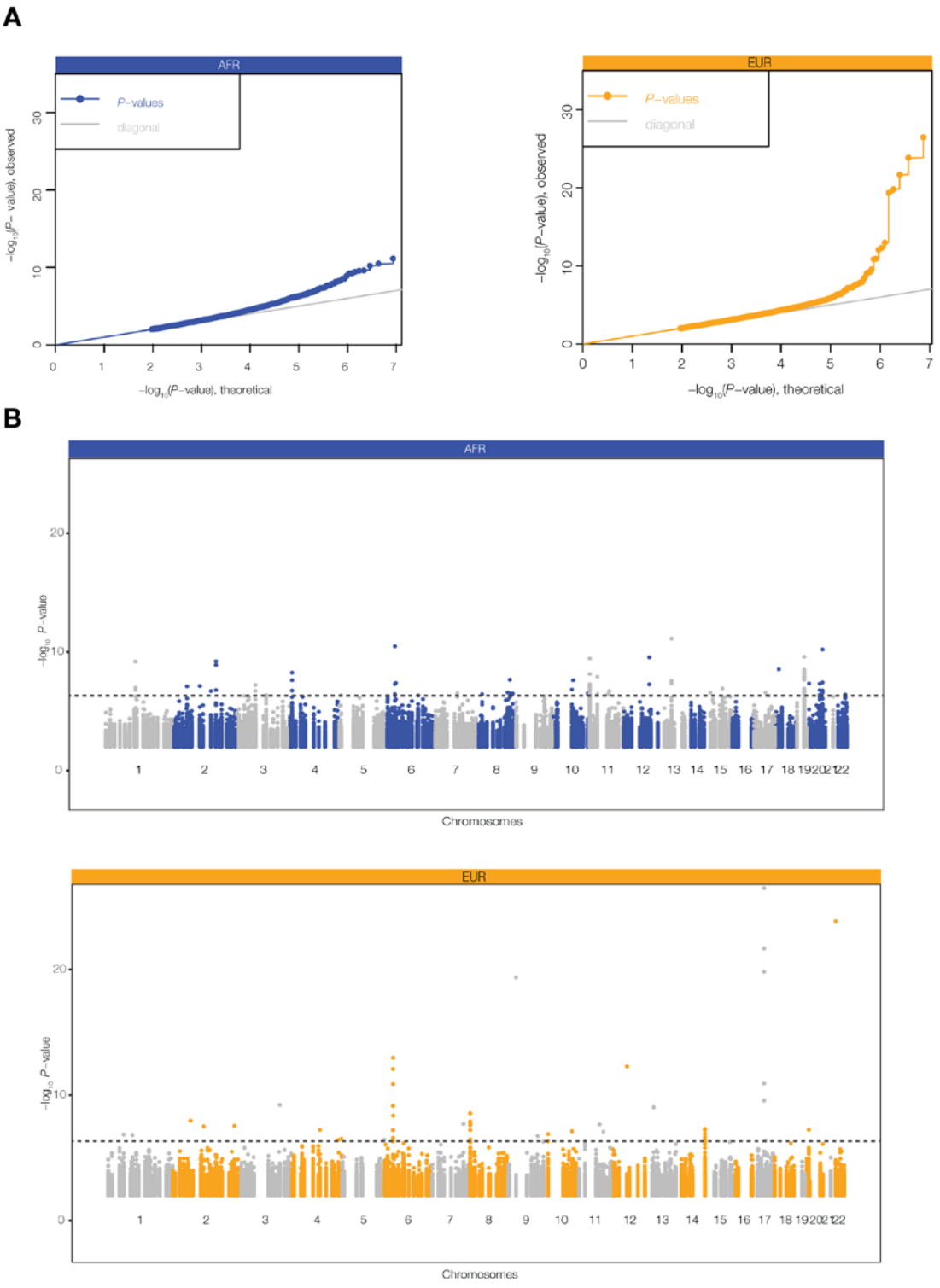


Figure 20 Expression quantitative trait (eQTL) analysis for disease-linked polyTEs.

eQTL analysis was performed by regressing lymphoblastoid gene expression levels against polyTE insertion genotypes for the for the African (AFR – blue) and European (EUR –

orange) individuals analyzed here. (A) Quantile-quantile (Q-Q) plots relating the observed (y-axis) to the expected (x-axis) TE-eQTL log transformed P-values. (B) Manhattan plots showing the genomic distributions of TE-eQTL log transformed P-values. The dashed line corresponds to a false discovery rate (FDR) threshold of q<0.05, corresponding to $P=4.7 \times 10^{-7}$ (AFR) and $P=2.6 \times 10^{-7}$ (EUR).

The results of the TE-eQTL analysis further underscore the regulatory potential of the disease-linked polyTEs characterized here and also allowed us to narrow down the list of candidate insertions. Starting with the list of TE-eQTL, we searched for 'consistent' examples where the disease-linked polyTE is associated with the expression of a gene that is functionally related to the annotated disease phenotype. This allowed us to converge on a final set of seven high-confidence disease-associated TE insertion polymorphisms (Figure 17 and Table 2). Four of these disease-associated polyTEs are illustrated in Figure 21, and we provide additional information on two examples in the following section (3.4). The four examples shown in Figure 21 all correspond to polyTEs that are linked to disease-associated SNPs and co-located with enhancers characterized from blood- or immune system-related tissues; in addition, the genes that these polyTE insertions regulate are all known to function in the immune system.

Table 2 Summary of TE-disease associations

polyTE	Chr	Pos	GWAS hits	GWAS Phenotype	GWAS gene	GWAS P-value	#Enhancer Overlaps	eGene	eQTL P-value	eQTL Type
Alu-2829	3	154966214	rs13064954	Diabetic retinopathy	LINC00881, CCNL1	7.00E-07	1	*LILRA1*	5.94E-10	*trans*
Alu-5072	6	32589834	rs4530903	Lymphoma	TRNAI25	2.00E-08	20	*HLA-DRB5*	8.49E-13	*cis*
Alu-5075	6	32657952	rs2858870	Nodular sclerosis Hodgkin lymphoma	TRNAI25	8.00E-18	15	*HLA-DQB1-AS1*	1.36E-11	*cis*
SVA-282	6	33030313	rs3077	Chronic hepatitis B infection	HLA-DPA1	5.00E-39	6	*HLA-DPB2*	1.05E-13	*cis*
SVA-401	9	33130564	rs10758189	IgG glycosylation	B4GALT1	2.00E-06	4	*B4GALT1*	4.47E-20	*cis*
SVA-438	10	17712792	rs6602203	Glucose homeostasis traits	ST8SIA6-AS1, PRPF38AP2	5.00E-06	7	*TMEM236*	1.30E-07	*cis*

Six of the seven disease-associated polyTE insertions are considered to be population-specific, based on significant eQTL results in only one population, whereas a single case is shared between both the AFR and EUR population groups (Figure 21C). However, two of the six cases considered to be population-specific using the eQTL criterion do show consistent trends across populations but failed to reach genome-wide significance when controls for multiple statistical tests were implemented (Figure 21D and Figure 22B).

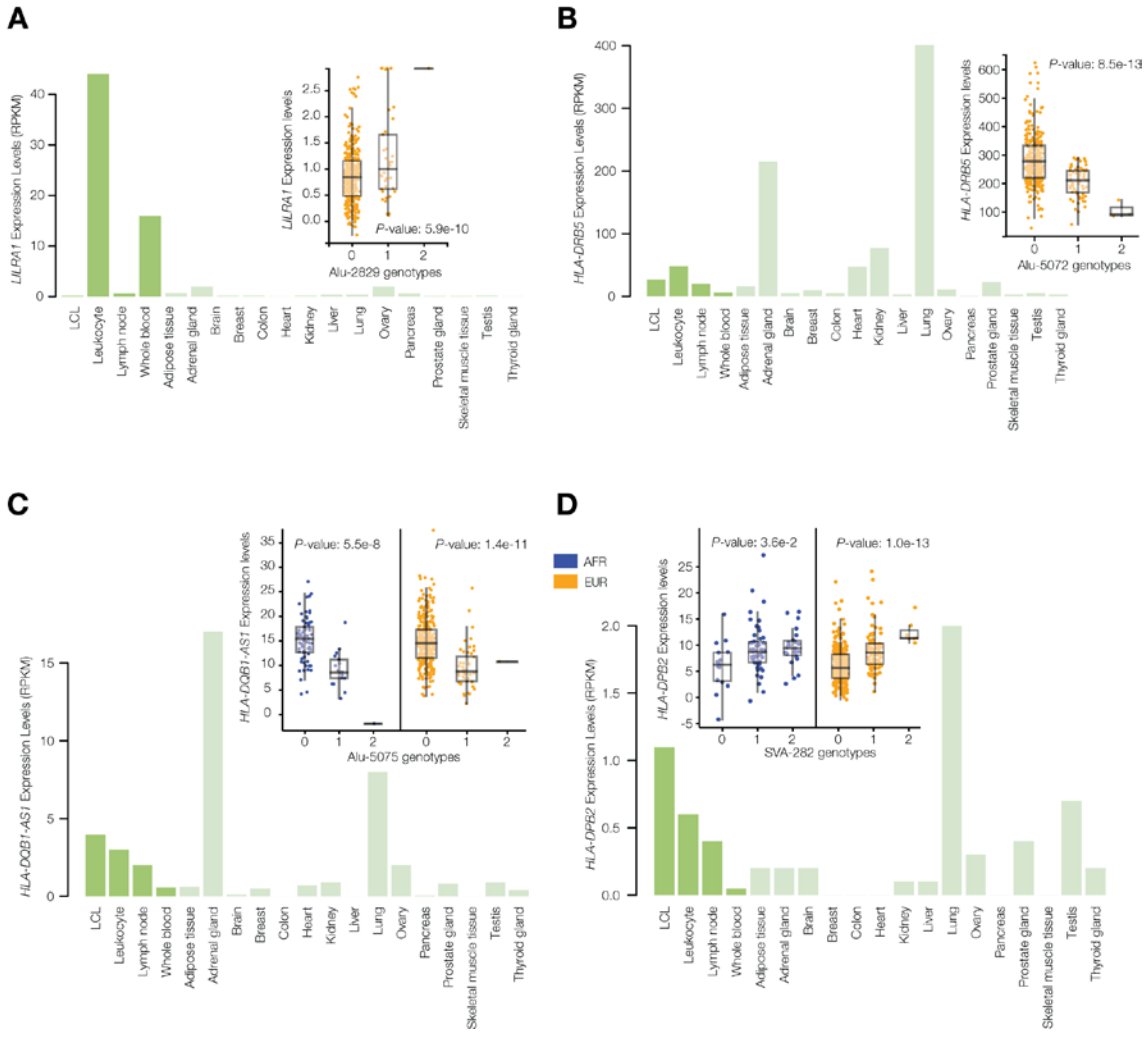


Figure 21 Gene expression profiles and eQTL results for disease-associated polyTE insertions.

Bar-charts of tissue-specific expression levels and box-pots of eQTL analyses are shown for four examples of TE-eQTLs corresponding to disease-linked and enhancer co-located polyTE insertions that regulate immune-related genes: (A) Alu-2829 and LILRA1, (B) Alu-5072 and HLA-DRB5, (C) Alu-5075 and HLA-DQB1-AS1, and (D) SVA-282 and HLA-DPB2. Bar-charts show tissue-specific expression levels as normalized RPKM values (green). The inset eQTL box-plots show individual sample gene expression levels (y-axis) regressed against polyTE insertion presence/absence genotypes (x-axis): 0-homozygous absent, 1-heterozygous, 2-homozygous present. Each dot represents a single individual from the African (AFR – blue) and/or European (EUR – orange) population groups.

4.4.4 Effects of polyTE insertions on immune- and blood-related conditions

Here, we described two specific examples of the effects that polyTE insertions can exert on immune- and blood-related disease phenotypes. Figure 22A shows the SVA-401 insertion that is co-located with a cell-type specific enhancer found in the second intron of the Beta-1,4-Galactosyltransferase 1 (*B4GALT1*) encoding gene, which is normally expressed at high levels in immune-related tissues. Chromatin interaction maps characterized for several different cell types – CD34, GM12878 and Mcf7 – show that this *B4GALT1* intronic enhancer physically associates with the gene's promoter region. The disruption of the *B4GALT1* enhancer by the SVA insertion is associated with down-regulation of the gene in B-cells, for both AFR and EUR population groups (Figure 22B). *B4GALT1* encodes a glycosyltransferase that functions in the glycosylation of the Immunoglobulin G (IgG) antibody in such a way as to convert its activity from pro- to anti-inflammatory (Figure 22C) [138-140]. Down-regulation of this gene in individuals with the enhancer SVA insertion should thereby serve to keep the IgG molecule in a pro-inflammatory state. Consistent with this idea, the *B4GALT1* enhancer SVA insertion is linked to a genomic region implicated by GWAS in both inflammatory conditions and autoimmune diseases such as systemic lupus erythematosus and Crohn's disease [140].

Another example of an SVA insertion into an enhancer element is shown for the adjacently located Signal Transducing Adaptor Molecule (*STAM*) and Transmembrane Protein 236 (*TMEM236*) encoding genes. The SVA-438 insertion is co-located with an enhancer in the first intron of the *STAM* gene (Figure 22D), but its presence is associated with changes in expression of the nearby *TMEM236* gene (Figure 22E). *TMEM236* is located ~100kbp downstream of the SVA-438 insertion and is most highly expressed in pancreatic islet α-cells (Figure 22F) [141, 142]. Islet α-cells function to secrete glucagon, a peptide hormone that elevates glucose levels in the blood [143]. The SVA-438 insertion is associated with increased expression of *TMEM236*, which would be expected to lead to increased blood glucose levels. This expectation is consistent with the fact that the SVA-438 insertion is also linked to the risk allele (T) of the SNP rs6602203, which is associated with a reduced metabolic clearance rate of insulin (MCRI), an endophenotype that is associated with the risk of type 2 diabetes [144]. In other words, up-regulation of *TMEM236* by the SVA-438 insertion may be mechanistically linked to insulin resistance by virtue of increasing blood sugar and decreasing insulin clearance.

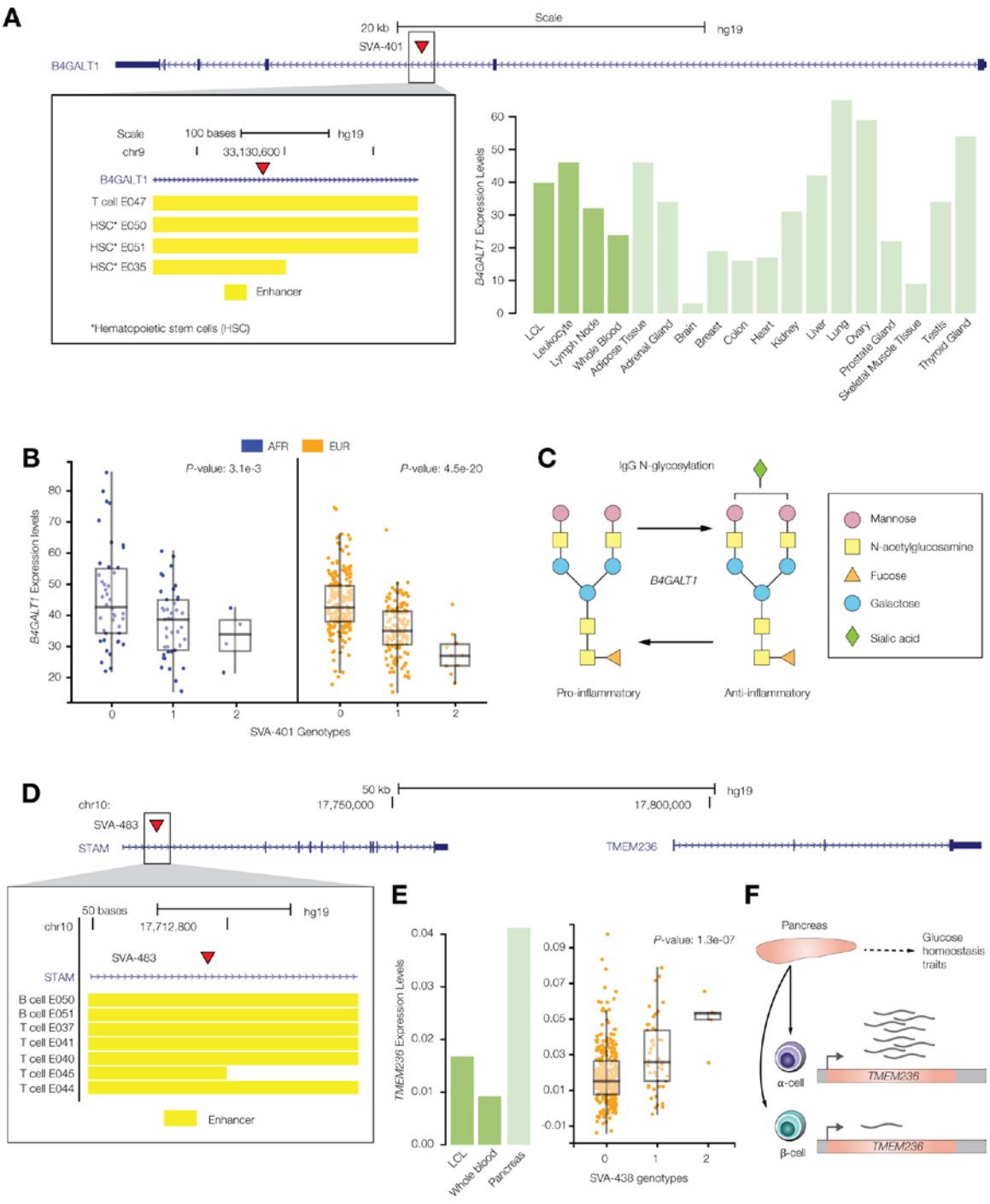


Figure 22 PolyTE insertions associated with immune- and blood-related conditions.

(A) UCSC Genome Browser screen capture showing the location of the SVA-401 insertion (red arrow) on chromosome 19 within the second exon of the B4GALT1 gene. The inset shows the genomic locations of co-located enhancers, characterized based on chromatin signatures from a variety of tissue-specific epigenomes locations, as yellow bars. The bar-chart shows B4GALT1 tissue-specific expression levels as normalized RPKM values

(green). (B) eQTL box-plots show individual sample gene expression levels (y-axis) regressed against SVA-401 insertion presence/absence genotypes (x-axis): 0-homozygous absent, 1-heterozygous, 2-homozygous present. Each dot represents a single individual from the African (AFR – blue) or European (EUR – orange) population groups. (C) B4GALT1 catalysed glycosylation of the Immunoglobulin G (IgG) antibody, resulting in conversion from pro- to anti-inflammatory activity. (D) UCSC Genome Browser screen capture showing the location of the SVA-438 insertion (red arrow) on chromosome 10 within the first exon of the STAM gene, upstream of the regulated TMEM236 gene. The inset shows the genomic locations of co-located enhancers (yellow bars). (E) Bar-chart of TMEM236 tissue-specific expression levels and box-pot of the SVA-438 TMEM236 eQTL analyses. (F) Functional role and cell-type specific expression profile for TMEM236.

4.5 Discussion

The results reported here underscore the influence that retrotransposon insertion polymorphisms can exert on human health- and disease-related phenotypes. The integrative data analysis approach that we took for this study also revealed how polyTE disease-associations are mediated by the gene regulatory properties of retrotransposon insertions. We adopted a conservative approach to screen for the potential regulatory effects of retrotransposon insertions by choosing candidate elements as those that were inserted into regions previously defined as tissue-specific enhancers in blood/immune cells. Retrotransposons that insert into enhancer sequences could entail loss-of-function mutants by virtue of disrupting enhancer sequences, or they could serve as gain-of-function mutants by altering enhancer activity. Our results can be considered to show instances of both loss- and gain-of-function enhancer mutations with respect to the decrease or increase, respectively, of gene expression levels that are associated with element insertion genotypes (Figure 21 and 22). Nevertheless, it is worth noting that our conservative approach could be prone to false negatives as it would not uncover novel enhancer activity provided by element insertions at new locations in the genome.

The TE regulatory findings that we report here are consistent with previous studies showing that TE-derived sequences have contributed a wide variety of gene regulatory elements to the human genome [17, 18, 128], including promoters [19-21], enhancers [22-26], transcription terminators [28] and several classes of small RNAs [29-31]. Human TEs can also influence gene regulation by modulating various aspects of chromatin structure throughout the genome [1, 32, 103-106].

It is important to note that the research efforts which have uncovered the regulatory properties of human TEs, including a number of our own studies, have dealt exclusively with sequences derived from relatively ancient insertion events. These ancient TE insertions are present at the same (fixed) locations in the genome sequences of all human individuals. In other words, previously described TE-derived regulatory sequences are uniformly present among individual human genomes and thereby do not represent a source of structural genetic variation. Such fixed TE-derived regulatory sequences may not be expected to provide for gene regulatory variation among individuals or for that matter to contribute to inter-individual differences related to health and disease.

Nevertheless, we recently showed that TE insertion polymorphisms also exert regulatory effects on the human genome [130]. Specifically, polyTE insertions were shown to contribute to both inter-individual and population-specific differences in gene expression and to facilitate the re-wiring of transcriptional networks. The results reported here extend those findings up the hierarchy of human biological organization by revealing potential mechanistic links between polyTE-induced gene regulatory changes and the endophenotypes that underlie human health and disease.

CHAPTER 5. TRANSPOSABLE ELEMENT ACTIVITY, GENOME REGULATION AND HUMAN HEALTH

5.1 Abstract

A recent convergence of novel genome analysis technologies is enabling population genomic studies of human transposable elements (TEs). Population surveys of human genome sequences have uncovered thousands of individual TE insertions that segregate as common genetic variants, *i.e.* TE polymorphisms. These recent TE insertions provide an important source of naturally occurring human genetic variation. Investigators are beginning to leverage population genomic data sets to execute genome-scale association studies for assessing the phenotypic impact of human TE polymorphisms. For example, the expression quantitative trait loci (eQTL) analytical paradigm has recently been used to uncover hundreds of associations between human TE insertion variants and gene expression levels. These include population-specific gene regulatory effects as well as coordinated changes to gene regulatory networks. In addition, analyses of linkage disequilibrium patterns with previously characterized genome-wide association study (GWAS) trait variants have uncovered TE insertion polymorphisms that are likely causal variants for a variety of common complex diseases. Gene regulatory mechanisms that underlie specific disease phenotypes have been proposed for a number of these trait associated TE polymorphisms. These new population genomic approaches hold great promise for understanding how ongoing TE activity contributes to functionally relevant genetic variation within and between human populations.

5.2 Introduction

Transposable elements (TEs) are distinguished by their ability to move, *i.e.* transpose, among genomic locations, often making copies of themselves as they go. TEs can replicate to extremely high copy numbers over time; at least 50% of the human genome sequence is thought to be derived from TE insertions [1, 2]. The abundance of TE sequences, along with their ability to colonize a seemingly endless variety of host genomes, begs an explanation for their evolutionary success. The selfish DNA theory holds that TEs are genomic parasites, which play no functional role for their hosts and exist simply by virtue of their ability to out-replicate the genomes in which they reside [145, 146]. The selfish DNA theory is still widely considered to represent the null hypothesis that best explains the presence of TEs from an evolutionary standpoint. Nevertheless, numerous studies have revealed instances of exaptation [147], also referred to as molecular domestication [148], whereby formerly selfish TE sequences have been co-opted to provide some functional utility for their host genomes. The most widely observed route of molecular domestication entails the conversion of TE sequences into host genome regulatory elements [17, 18, 128].

TE-derived sequences provide a wide variety of regulatory elements to the human genome, including promoters [19-21], enhancers [22-26], transcription terminators [28] and several classes of small RNAs [29-31]. Human TE-derived sequences can also exert higher order influences on gene regulation by shaping chromatin structure across the genome [32, 103-106]. It is important to note that, until this time, nearly all studies on human TE regulatory elements have focused on TE-derived sequences that are remnants of relatively ancient insertion events and no longer capable of transposition. In other

words, known human TE regulatory sequences largely correspond to so-called ‘fixed’ TE insertions, which are found at the same genomic insertion site locations within the genomes of all human individuals. This distinction is critical, since fixed TE insertions are not expected to contribute to regulatory variation among individual humans. In other words, fixed TE regulatory elements, while functionally important, do not provide a source of human population genetic variation.

Over the last several years, a convergence of genome-enabled technologies has begun to power studies that are focused squarely on structural variations generated by the ongoing activity of human TEs. There are several families of human TEs that retain the ability to transpose, primarily Alu [3, 4], L1 [5, 6], and SVA [7, 8]. Smaller numbers of HERV-K endogenous retroviruses also remain active in the human genome [54]. When members of these TE families transpose within the human genome, they generate inter-individual variations that segregate within and between populations in the form of TE insertion site polymorphisms. Given the known regulatory properties of human TEs, it is not unreasonable to expect that segregating TE polymorphisms could have significant regulatory consequences. In particular, some human TE polymorphisms may lead to differences in gene expression patterns between individuals. Furthermore, human regulatory variation generated by recent TE activity may have important implications for health and disease. This mini-review is focused on recent studies that are beginning to shed light on the ways in which ongoing TE activity can impact human health via changes in genome regulation. These studies are distinguished by their population level approach to the study of TE generated human variation (Figure 23).

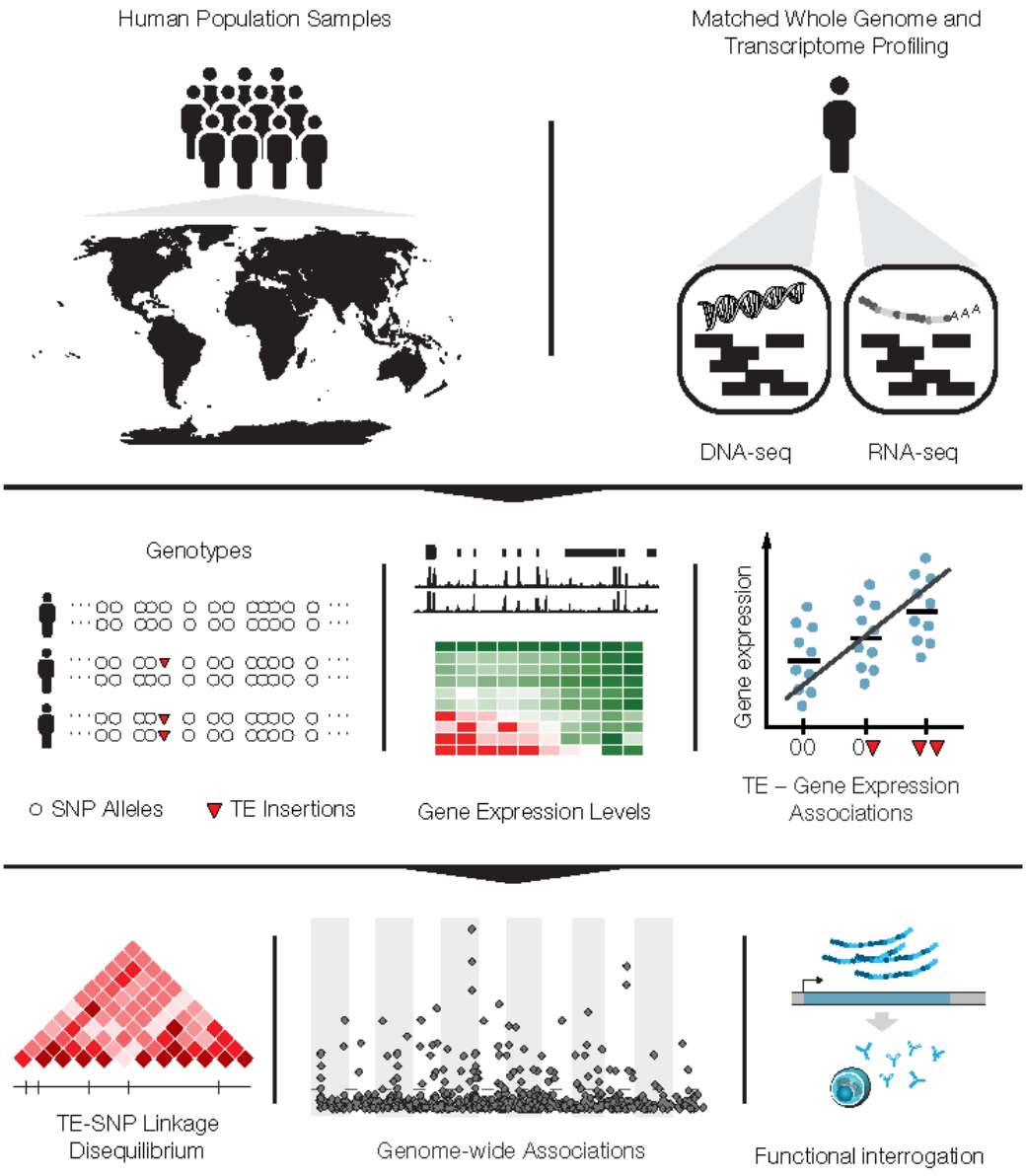


Figure 23 The population genomic approach for the study of TE phenotypic effects.

Individuals sampled from human populations are characterized using genome (DNA-seq) and transcriptome (RNA-seq) profiling techniques. Genome-wide TE insertion genotypes are compared to tissue-specific gene expression levels to uncover TE variants implicated in gene regulation. The linkage disequilibrium patterns (LD) among TE polymorphisms and SNPs are evaluated to identify TE insertions linked to genome-wide association study (GWAS) loci. Interrogation of functional information is used to hone in on likely TE causal variants.

5.3 Genome-enabled approaches for characterizing TE insertion variants

Two distinct classes of genome-enabled approaches for the characterization of TE insertion variants have emerged over the last several years [131]: (1) bioinformatics methods that rely on the analysis of whole genome sequence data to find TE insertions that differ from a reference sequence (Figure 24A), and (2) high-throughput experimental methods that leverage next-generation sequencing to pinpoint the locations of novel TE insertions (Figure 24B).

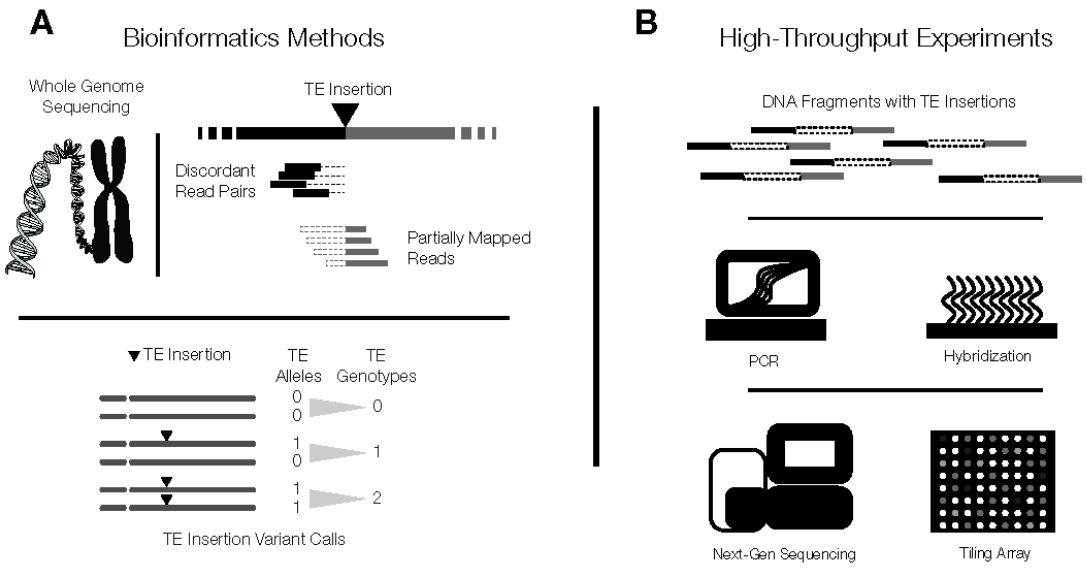


Figure 24 Genome-enabled approaches for the discovery and characterization of TE insertion variants.

(A) Bioinformatics methods rely on the computational analysis of whole genome sequence read data to characterize genome-wide patterns of TE insertion alleles and genotypes. (B) High-throughput experimental methods use enrichment of genomic fragments that contain known active TE sequences followed by sequence or array based characterization of their genomic locations.

Computational approaches for the discovery of TE insertion variants rely on one of two methods: (1) discordant read-pair mapping for short read sequencing technology, or (2) split read mapping for long read technology [120]. Our own group recently performed a benchmarking study on 21 bioinformatics tools designed for detecting human TE insertion variants from whole genome sequence data [64]. After an initial screen of tools that were found to be unreliable, or no longer maintained, our study focused on seven programs: ITIS [149], MELT [150], Mobster [151], RetroSeq [152], Tangram [153], TEMP [154], and T-lex2 [155]. We found MELT to have superior performance for human TE variant detection from whole genome sequence data, but also show how a combined approach using two or more methods, including Mobster and RetroSeq, could yield superior performance. Since the publication of our paper, two new computational tools for TE insertion discovery have been published. The program STEAK [156] claims superior performance compared to existing short read methods, whereas LoRTE [157] is designed for PacBio® long read sequence technology.

At this time, given the predominance of Illumina® short read sequencing technology, discordant read-pair mapping approaches are most widely used. However, these methods are still far from perfect and there is substantial room for additional development in the field. As long read sequencing technology becomes more widespread, split read approaches should become more popular. Perhaps more importantly, we expect that split read approaches will be inherently more accurate and reliable than discordant read pair mapping, since long reads that span entire TE insertions should be mapped with much less ambiguity than shorter reads.

High-throughput experimental techniques for TE variant detection also share several basic features: (1) DNA fragmentation, (2) TE enrichment, and (3) TE calling. The methods are distinguished by the approaches used for each step of the process. DNA fragmentation can be achieved via enzymatic digestion or by mechanical shearing. TE enrichment can be performed using PCR, with active TE-specific primers, or with hybridization to active TE-specific probes. Finally, TE calling is done using next-generation sequencing, for more recent methods, or with tiling arrays for the older methods. The most widely used experimental methods for TE variant detection include ME-Scan [158], L1-Seq [159], RC-seq [160], and Transposon-Seq [75]. One area of ongoing improvement for these methods entails the refinement of algorithms used to map enriched TE fragments to genome reference sequences. For example, the TIPseqHunter algorithm was recently developed to refine and improve human TE variant calls made by the existing TIP-seq experimental method [161].

Genome-scale experimental approaches of this kind have been most widely applied to the study of somatic TE variants that characterize cancer tissues. This is one of the most promising areas of recent human TE research, and it has been extensively reviewed elsewhere [162]. This mini-review is focused instead on germline mutations that yield inter-individual differences in TE insertion patterns and manifest themselves as human population genetic variations, *i.e.* TE polymorphisms.

5.4 TE polymorphisms and human genome regulation

Our own group recently published a population-level view of the regulatory consequences of recent human TE activity [163]. To do so, we adopted the expression quantitative trait loci (eQTL) analytical paradigm for human TE polymorphisms. eQTL are genomic variants associated with changes in gene expression levels [164]. The eQTL approach requires multiple individual samples that have been deeply characterized at both the genomic (DNA-seq) and transcriptomic (RNA-seq) levels. Gene expression levels for individual samples are regressed against locus-specific genotypes for matched individuals to uncover eQTL associations. This approach was developed for single nucleotide polymorphism (SNP) genotypes, whereas in our case, we used locus-specific TE insertion state genotypes. TE insertion genotypes at any locus can be encoded as 0 (homozygous - insertion absent), 1 (heterozygous - one insertion present), or 2 (homozygous - two insertions present). Differences in gene expression levels across these distinct TE insertion states are indicative of TE polymorphism-to-gene expression associations (Figure 23).

This approach was powered by the 1000 Genomes Project (1KGP), phase 3 of which entailed the genome-wide characterization of TE insertion genotypes for 2,504 individuals across 26 human populations [92, 99]. B-lymphocyte gene expression data for 445 of the same 1KGP individuals, representing one African population and four European populations, were taken from the Genetic European Variation in Health and Disease (GEUVEDIS) RNA-seq project [93]. Merging data from both projects allowed us to directly compare TE insertion site genotypes to gene expression levels from the same individuals. Furthermore, comparison of results for African and European populations allowed us to uncover population-specific regulatory effects of human TE polymorphisms.

Regression of gene expression against TE insertion site genotypes revealed hundreds of eQTL associations, and TE-eQTL were found both within and between the African and European populations. A number of TE polymorphisms were shown to be associated with expression differences between population groups. One advantage of using TE insertion site genotypes for eQTL analysis is that the relatively low number of common TE genotypes across the genome (~16,000) allows for both *cis* and *trans* eQTL analysis. This is because the number of possible eQTL associations is the product of the number of genes and the number of variants being compared; accordingly, the analysis of millions of SNPs times thousands of genes presents a combinatorically daunting bioinformatics analysis challenge. For this reason, most SNP eQTL studies focus exclusively on *cis* SNPs that are found within or in close proximity to individual genes. Since our study was not limited in this way, we were able to discover many *trans* effects of TE polymorphisms on human gene regulation. In fact, we were surprised to find that *trans* regulatory effects for TE polymorphisms were even more common than *cis* effects.

For one particular example, the B cell specific transcription factor PAX5, we uncovered a potential mechanism that could explain the numerous *trans* TE-eQTL that we observed (Figure 25). This example also underscores how individual TE loci can participate in the rewiring of entire regulatory networks. The *PAX5* gene has a *cis* Alu eQTL that is associated with increased expression in B lymphocytes. This same Alu insertion is associated with increased expression of numerous PAX5 target genes, presumably by virtue of a transitive effect whereby increased PAX5 expression in turn increases the expression of downstream targets in its regulatory network.

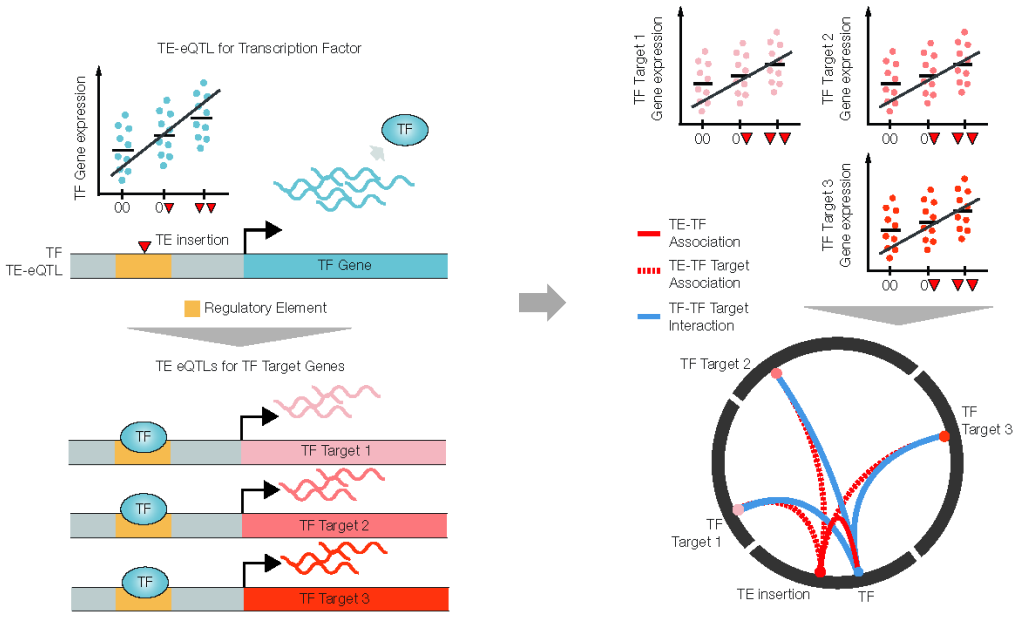


Figure 25 The impact of TE polymorphisms on gene regulatory networks

The eQTL approach is used to discover associations between TE insertion variants and tissue-specific gene expression levels (i.e. TE-eQTLs). A TE insertion variant found in cis to a transcription factor (TF) can lead to coordinated changes across a gene regulatory network via transitive effects on downstream targets of the TF. An example is shown, similar to what has been observed for the TF gene PAX5, where TE associated increase in the expression of a TF leads in turn to increased expression of the TF target genes. This will reveal itself as multiple trans TE-eQTL associations for the same TE insertion variant.

To our knowledge, this is the first and only study of its kind in humans. However, analogous genome-scale approaches have been used to discover TE associations with gene expression in the model organisms Arabidopsis [165] and maize [166].

5.5 TE polymorphisms and complex common disease

Two recent studies have taken a similar population-level view of the phenotypic effects of human TE polymorphisms [167, 168]. For each of these studies, associations

between TE insertion site genotypes and complex common diseases were explored. Both studies relied on the analysis of linkage disequilibrium (LD) patterns to discover TE polymorphisms linked to SNPs that were previously associated with health or disease related phenotypes via genome-wide association studies (GWAS). An implicit rationale for genome-scale surveys of this kind is the notion that TE insertions are expected to be more disruptive than SNP variations given the larger scale genomic changes that they entail. Interestingly, both studies report that TE polymorphisms are enriched at GWAS loci, highlighting their potential impact. The first study of this kind, from the group of Kathleen Burns, found 44 Alu insertions in tight LD with previously discovered GWAS trait associated SNPs [167]. The authors pointed out that this represents a >20-fold increase over the number of polymorphic Alu insertions that were previously known to be associated with human phenotypes, thereby underscoring the power of population genomic approaches for studies on the phenotypic impact of TE polymorphisms. Furthermore, the implicated Alu polymorphisms were found to be associated with a very broad range of health and disease related phenotypes.

Our own study on the impact of TE polymorphisms on complex common disease was designed to explore the connection between TE-mediated genome regulation and disease related phenotypic effects [168]. To achieve this aim, we used a progressive set of genome-wide bioinformatics screens that searched for polymorphic TE insertions that are: (1) found in LD with known GWAS SNPs, (2) located within tissue-specific enhancers, and (3) associated with tissue-specific gene expression levels. We further narrowed our search for candidate TE polymorphisms to those associated with genes with blood or immune related functions, consistent with the fact that the gene expression data we analyzed is from B lymphocytes. This progressive and stringent genomic screen uncovered six TE polymorphisms that are likely to be associated with disease phenotypes by virtue of their gene regulatory effects. These included both Alu elements, as previously reported, as

well as SVA elements. For example, we discovered an SVA insertion in the cell-type specific enhancer of the *B4GALT1* gene (Figure 26). B4GALT1 acts to convert the Immunoglobulin G (IgG) antibody from a pro-inflammatory to an anti-inflammatory form. The SVA insertion is associated with both down-regulation of the *B4GALT1* gene, thereby potentially leading to increased inflammation, and linked to a genomic region implicated by GWAS in both inflammatory conditions and autoimmune disease.

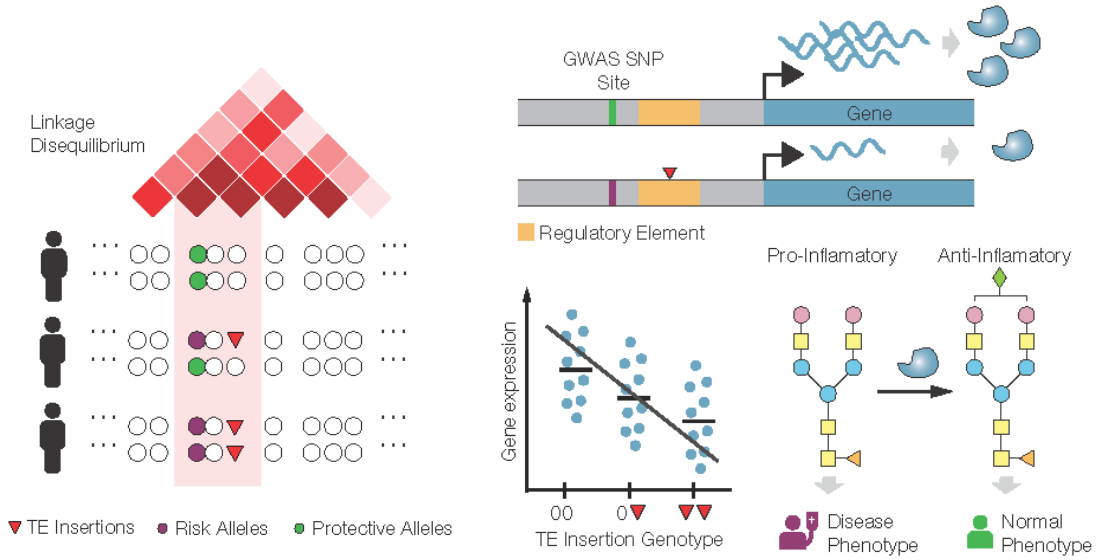


Figure 26 TE insertion variants impact on human disease via gene regulatory changes

TE insertion variants are found in tight linkage disequilibrium (LD) with previously characterized genome-wide association study (GWAS) SNP risk alleles. The linked TE insertion variant is associated with reduced gene expression, which is in turn associated with elevated disease risk. The scheme shown here corresponds to a TE insertion variant associated with reduced expression of the B4GALT1 gene, which leads to increased inflammation and related disease pathology.

5.6 Conclusions

The population genomics view of TEs exemplified by the recent studies reviewed here has the potential to expand our understanding of the phenotypic impact of human TEs. While ongoing human TE activity has widely been considered to be deleterious, the presence of TE insertion variants that segregate as common polymorphisms among human populations indicates that many novel TE insertions must have escaped the action of purifying selection. Accordingly, polymorphic human TE insertion variants comprise an important source of naturally occurring genetic variation with subtle effects on genome regulation and human health. Functionally relevant TE polymorphisms of this kind are likely to provide crucial source material for ongoing human evolution.

APPENDIX A.

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

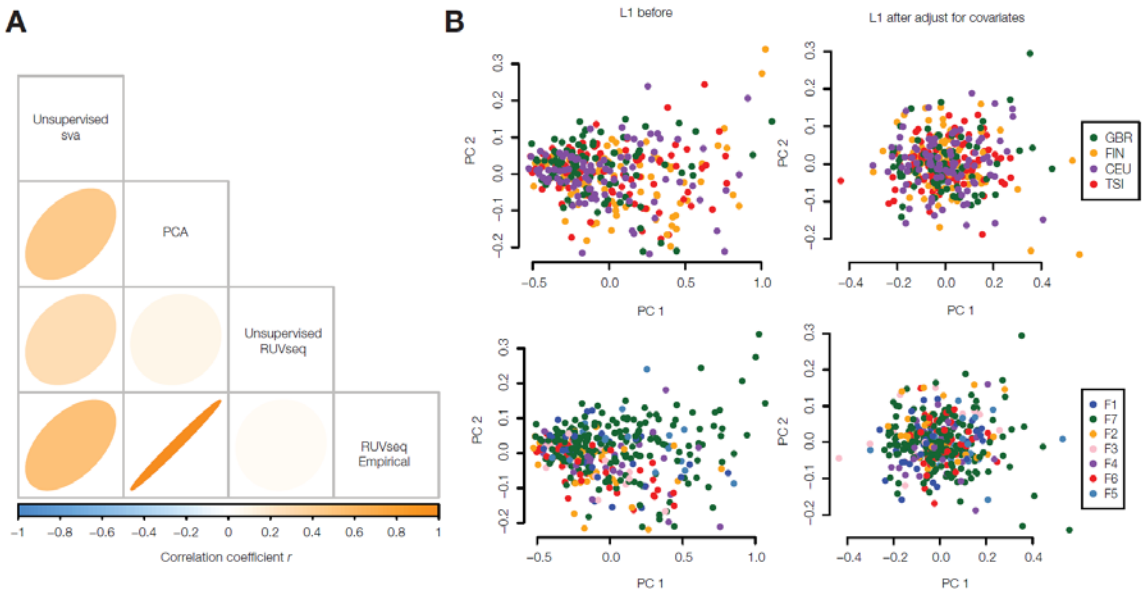


Figure 27 RNA-seq normalization and covariate adjustments

(A) Comparison of four RNA-seq normalization approaches that were evaluated here: unsupervised SVAseq, PCA, unsupervised RUVseq and empirical RUVseq. The matrix shows the results of pairwise comparisons between the latent factors influencing gene expression estimated by each method. For each cell in the pairwise comparison matrix, the shape and the color represent the magnitude of the correlation coefficient (r) observed between methods, based on the estimated latent factors. Rounder shapes and lighter colors show lower correlations; more narrow and brighter colors show higher correlations. (B) Comparison of the first two PCAs of the L1 expression matrices shown before and after adjustment for co-variates. L1 expression PCA values are color coded by population origin (top) and sequencing batch (bottom).

APPENDIX B.

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

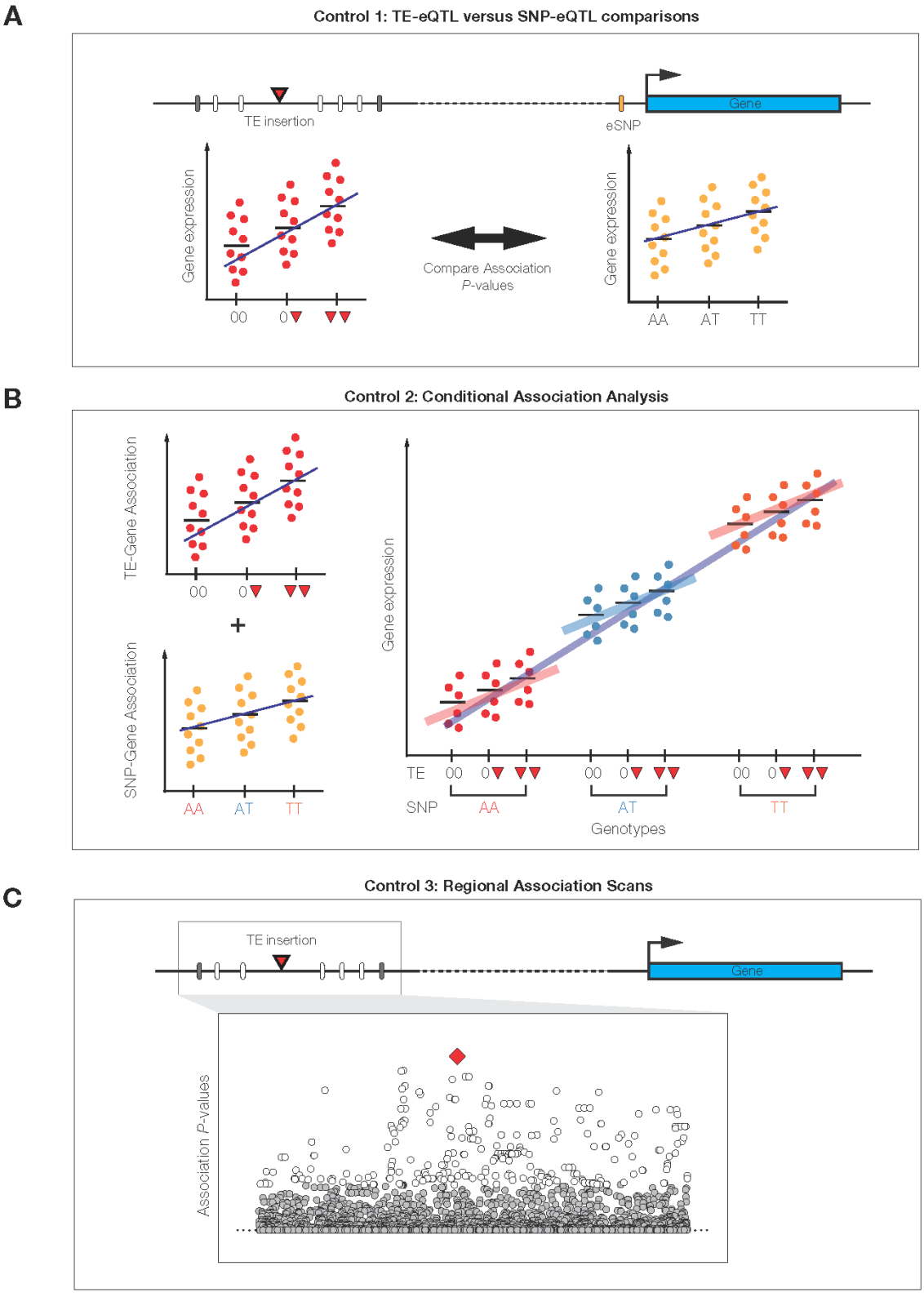

A
Control 1: TE-eQTL versus SNP-eQTL comparisons
TE insertion
eSNP
Gene
Gene expression
00
0
Compare Association
P-values
AA
AT
TT
B
Control 2: Conditional Association Analysis
TE-Gene Association
SNP-Gene Association
Gene expression
TE
SNP
Genotypes
C
Control 3: Regional Association Scans
TE insertion
Gene
Association P-values

Figure 28 Scheme of the three analyses used to control for the potential effects of regulatory SNPs on the TE-eQTL associations observed here.

(A) TE-eQTL versus SNP-eQTL comparisons. (B) Conditional association analysis. (C) Regional association scans.

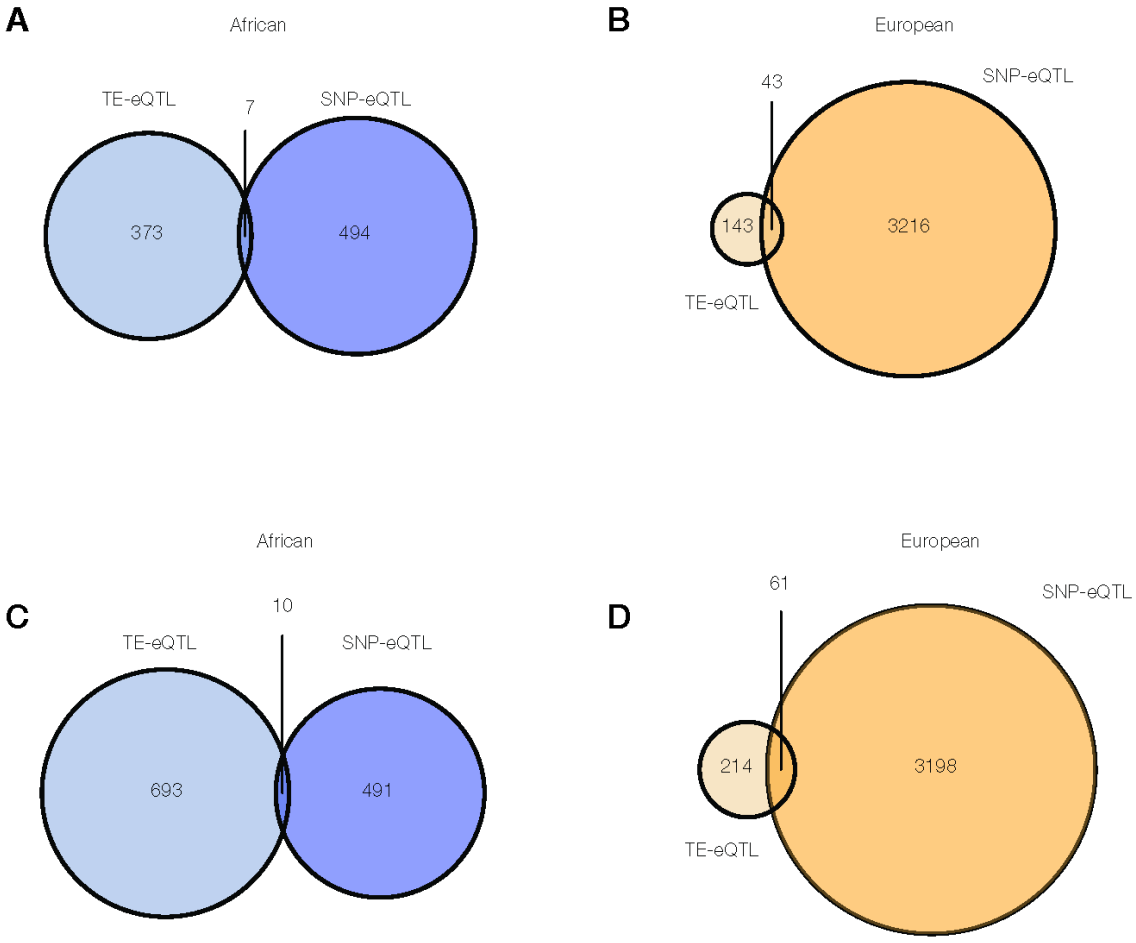
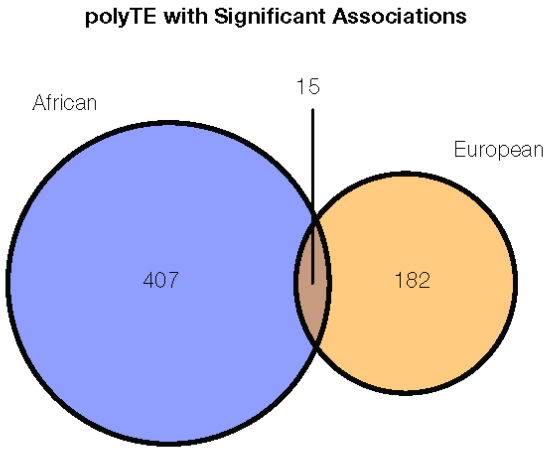


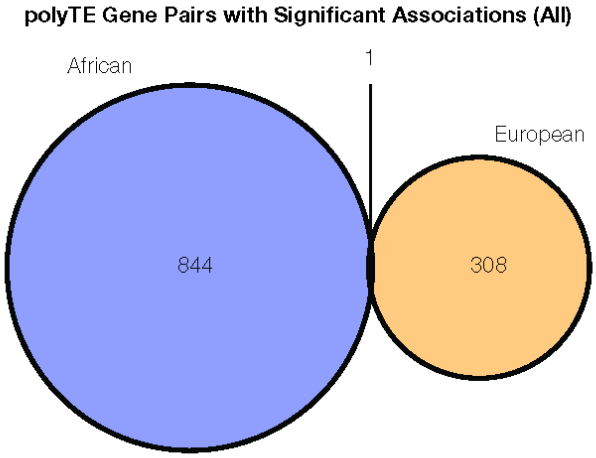
Figure 29 TE-eQTL versus SNP-eQTL comparisons.

Venn diagrams showing the numbers of genes found to be associated with TE-eQTL and/or SNP-eQTL for (A) African (blue) or (B) European (orange) populations. The numbers of genes associated with both TE-eQTL and SNP-eQTL for each population group are shown in the intersections.

A


polyTE with Significant Associations
African
15
European
407
182

B


polyTE Gene Pairs with Significant Associations (All)
African
1
European
844
308

C

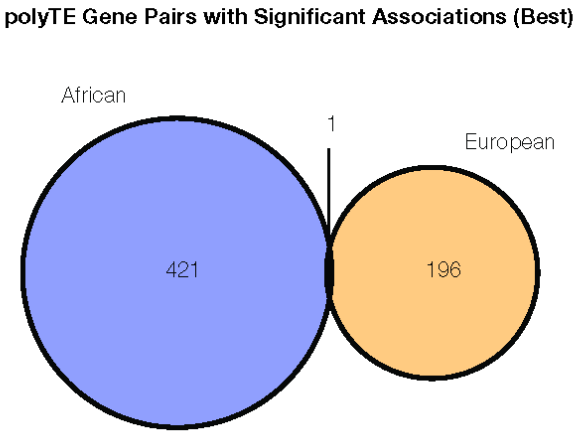

polyTE Gene Pairs with Significant Associations (Best)
African
1
European
421
196

Figure 30 Population-specific TE-eQTL associations.

Venn diagram showing the number of polyTEs with associations that are found in Africa (blue), Europe (orange) or both (intersection).

Table 3 Best TE eQTLs identified in the analysis

Chr	Pos	TE	Gene	t Statistic	P-value	FDR	TE subfamily	eQTL type
chr17	44153977	SVA_umary_SVA_706	RP11-259G18.3	22.63	7.83E-76	4.54E-68	SVA	trans
chr17	43670097	SVA_umary_SVA_705	RP11-259G18.3	14.37	1.11E-38	1.61E-31	SVA	trans
chr6	133160120	ALU_umary_ALU_5602	RPS12	-13.91	1.04E-36	1.21E-29	ALU	cis
chr17	43660599	SVA_umary_SVA_704	RP11-259G18.3	13.47	7.28E-35	7.02E-28	SVA	trans
chr8	71914591	ALU_umary_ALU_6806	XKR9	12.07	3.54E-29	2.05E-22	ALU	cis
chr1	75192907	L1_umary_LINE1_61	CRYZ	-11.02	4.12E-25	1.84E-18	L1	cis
chr22	19210913	L1_umary_LINE1_2986	CLTCL1	10.72	5.37E-24	2.22E-17	L1	cis
chr7	32831241	ALU_umary_ALU_5939	DPY19L1P1	10.62	1.28E-23	4.93E-17	ALU	cis
chr4	185651677	ALU_umary_ALU_3997	CENPU	-9.93	4.11E-21	1.40E-14	ALU	cis
chr9	33130564	SVA_umary_SVA_401	B4GALT1	-9.60	6.13E-20	1.87E-13	SVA	cis
chr6	130614798	ALU_umary_ALU_5583	TMEM200A	-9.54	9.93E-20	2.87E-13	ALU	cis
chr9	98696746	ALU_umary_ALU_7578	LINC00476	9.33	5.29E-19	1.46E-12	ALU	cis
chr6	32657952	ALU_umary_ALU_5075	HLA-DQB1-AS1	-9.11	2.86E-18	7.52E-12	ALU	cis
chr6	29818872	ALU_umary_ALU_5053	HLA-J	9.04	5.03E-18	1.27E-11	ALU	cis
chr11	43877448	ALU_umary_ALU_8559	HSD17B12	-8.98	7.96E-18	1.92E-11	ALU	cis
chr14	92619420	SVA_umary_SVA_615	NDUFB1	-8.96	9.00E-18	2.08E-11	SVA	cis
chr13	21894780	ALU_umary_ALU_9602	LINC00422	-8.92	1.22E-17	2.72E-11	ALU	cis
chr3	48372326	ALU_umary_ALU_2337	FCF1P2	-8.53	2.39E-16	5.12E-10	ALU	cis
chr6	33030313	SVA_umary_SVA_282	HLA-DPB2	8.29	1.36E-15	2.71E-09	SVA	cis
chr12	56753252	ALU_umary_ALU_9228	RP11-977G19.11	-8.03	8.77E-15	1.59E-08	ALU	cis
chr15	41100332	ALU_umary_ALU_10654	ZFYVE19	-8.02	9.90E-15	1.74E-08	ALU	cis
chr10	70973334	SVA_umary_SVA_452	HKDC1	7.97	1.40E-14	2.32E-08	SVA	cis
chr18	59713734	ALU_umary_ALU_11754	PIGN	-7.95	1.54E-14	2.47E-08	ALU	cis
chr7	23665685	ALU_umary_ALU_5888	CCDC126	7.90	2.29E-14	3.59E-08	ALU	cis
chr19	52888074	ALU_umary_ALU_11979	ZNF880	7.88	2.57E-14	3.81E-08	ALU	cis
chr2	191478717	L1_umary_LINE1_477	MFSD6	-7.72	7.82E-14	1.08E-07	L1	cis
chr9	37596348	ALU_umary_ALU_7420	POLR1E	-7.58	2.05E-13	2.76E-07	ALU	cis
chr6	32589834	ALU_umary_ALU_5072	HLA-DRB5	-7.47	4.31E-13	5.67E-07	ALU	cis
chr19	57023579	ALU_umary_ALU_11997	ZFP28	7.45	4.82E-13	6.20E-07	ALU	cis
chr1	11830220	SVA_umary_SVA_5	C1orf167	7.31	1.28E-12	1.61E-06	SVA	cis
chr13	50173371	ALU_umary_ALU_9777	ARL11	-7.21	2.41E-12	2.84E-06	ALU	cis
chr15	45569912	ALU_umary_ALU_10670	CTD-2651B20.3	7.18	2.90E-12	3.36E-06	ALU	cis

Chr	Pos	TE	Gene	t Statistic	P-value	FDR	TE subfamily	eQTL type
chr2	3630321	ALU_umary_ALU_970	RNASEH1-AS1	-7.18	3.03E-12	3.44E-06	ALU	cis
chr14	57744450	ALU_umary_ALU_10334	MAML2	7.08	5.87E-12	6.41E-06	ALU	trans
chr21	31411594	ALU_umary_ALU_12362	NCCRP1	7.00	9.38E-12	9.88E-06	ALU	trans
chr2	20085249	ALU_umary_ALU_1042	PATL1	6.84	2.64E-11	2.73E-05	ALU	trans
chr6	31250241	ALU_umary_ALU_5060	HCG4P7	6.71	5.88E-11	5.78E-05	ALU	trans
chr2	199988338	ALU_umary_ALU_1903	PIK3AP1	6.69	6.88E-11	6.64E-05	ALU	trans
chr4	56371962	ALU_umary_ALU_3295	SRD5A3	6.68	7.07E-11	6.71E-05	ALU	cis
chr11	65984338	ALU_umary_ALU_8622	RP11-755F10.1	6.65	8.76E-11	8.19E-05	ALU	cis
chr4	146749	SVA_umary_SVA_203	ZNF718	6.64	9.07E-11	8.34E-05	SVA	cis
chr7	125922888	ALU_umary_ALU_6374	IGLV3-27	6.64	9.47E-11	8.57E-05	ALU	trans
chr3	26906584	ALU_umary_ALU_2238	GPR128	6.61	1.11E-10	9.91E-05	ALU	trans
chr2	68733422	ALU_umary_ALU_1278	PVR	6.60	1.16E-10	1.02E-04	ALU	trans
chr12	58359071	ALU_umary_ALU_9234	XRCC6BP1	6.55	1.61E-10	1.40E-04	ALU	cis
chr9	76893085	ALU_umary_ALU_7481	PHF2	6.53	1.85E-10	1.58E-04	ALU	trans
chr5	124155349	L1_umary_LINE1_1205	IGKV2D-29	6.52	1.91E-10	1.60E-04	L1	trans
chr1	234805728	ALU_umary_ALU_884	RP4-781K5.6	-6.47	2.65E-10	2.17E-04	ALU	cis
chr17	6538671	ALU_umary_ALU_11197	CTC-281F24.1	6.36	4.95E-10	3.98E-04	ALU	cis
chr5	36792620	ALU_umary_ALU_4218	CTD-2353F22.1	-6.34	5.70E-10	4.52E-04	ALU	cis
chr15	55129551	L1_umary_LINE1_2652	CRYBB2	6.31	6.85E-10	5.36E-04	L1	trans
chr14	82223469	ALU_umary_ALU_10437	LRRC20	-6.29	7.57E-10	5.77E-04	ALU	trans
chr14	75499581	ALU_umary_ALU_10401	EIF2B2	6.29	7.79E-10	5.86E-04	ALU	cis
chr3	21234927	L1_umary_LINE1_558	MAML2	6.28	8.07E-10	5.91E-04	L1	trans
chr2	227473038	ALU_umary_ALU_2029	IGLV3-27	6.28	8.35E-10	6.04E-04	ALU	trans
chr9	87506228	ALU_umary_ALU_7533	ST8SIA1	6.27	8.83E-10	6.31E-04	ALU	trans
chr2	65163783	ALU_umary_ALU_1256	SLC1A4	-6.25	9.54E-10	6.66E-04	ALU	cis
chr8	133725454	ALU_umary_ALU_7169	FAM65B	6.23	1.06E-09	7.30E-04	ALU	trans
chr9	97710713	ALU_umary_ALU_7572	CROCCP2	6.22	1.15E-09	7.77E-04	ALU	trans
chr12	97526402	ALU_umary_ALU_9451	GJA4	6.22	1.15E-09	7.77E-04	ALU	trans
chr7	91751552	ALU_umary_ALU_6200	CYP51A1P2	-6.22	1.19E-09	7.89E-04	ALU	trans
chr7	63880180	ALU_umary_ALU_6074	TRIM60P18	6.20	1.32E-09	8.66E-04	ALU	cis
chr8	35244893	L1_umary_LINE1_1650	LRP2	6.18	1.48E-09	9.63E-04	L1	trans
chr13	60032561	ALU_umary_ALU_9825	HTR3A	6.17	1.56E-09	1.00E-03	ALU	trans
chr6	131028588	ALU_umary_ALU_5587	ZNF462	6.17	1.57E-09	1.00E-03	ALU	trans
chr15	53941096	ALU_umary_ALU_10715	TPCN2	-6.16	1.68E-09	1.04E-03	ALU	trans
chr10	6271127	ALU_umary_ALU_7761	FRAS1	6.15	1.78E-09	1.09E-03	ALU	trans
chr5	118097178	ALU_umary_ALU_4615	LRRC20	-6.13	1.91E-09	1.14E-03	ALU	trans
chr11	86195336	ALU_umary_ALU_8700	IGLV7-46	6.13	2.00E-09	1.18E-03	ALU	trans
chr7	22799818	ALU_umary_ALU_5885	SMPDL3A	6.11	2.15E-09	1.26E-03	ALU	trans
chr12	80480726	ALU_umary_ALU_9371	TPCN2	-6.07	2.68E-09	1.54E-03	ALU	trans
chr12	4742915	SVA_umary_SVA_532	AKAP3	-6.06	2.89E-09	1.61E-03	SVA	cis
chr4	185875956	ALU_umary_ALU_3998	TEPP	6.05	3.03E-09	1.66E-03	ALU	trans
chr6	32449301	ALU_umary_ALU_5065	HLA-DRB1	6.02	3.68E-09	1.97E-03	ALU	cis

Chr	Pos	TE	Gene	t Statistic	P-value	FDR	TE subfamily	eQTL type
chr3	86264950	L1_umary_LINE1_634	ATP6V0E2-AS1	6.02	3.71E-09	1.97E-03	L1	trans
chr7	18345324	ALU_umary_ALU_5870	RP11-777B9.5	6.01	3.82E-09	2.01E-03	ALU	trans
chr8	2275573	ALU_umary_ALU_6535	ENPEP	5.98	4.65E-09	2.40E-03	ALU	trans
chr4	95413280	ALU_umary_ALU_3517	FREM1	5.96	5.07E-09	2.55E-03	ALU	trans
chr1	40524704	ALU_umary_ALU_114	RP11-386M24.6	5.96	5.14E-09	2.56E-03	ALU	trans
chr4	74430433	ALU_umary_ALU_3402	CRYBB2	5.96	5.18E-09	2.56E-03	ALU	trans
chr5	95394266	ALU_umary_ALU_4487	CCL1	5.96	5.26E-09	2.57E-03	ALU	trans
chr3	165336826	ALU_umary_ALU_2883	LZIC	5.95	5.34E-09	2.57E-03	ALU	trans
chr5	37648672	ALU_umary_ALU_4220	CRYBB2P1	5.94	5.78E-09	2.72E-03	ALU	trans
chr17	35210358	ALU_umary_ALU_11294	ASB14	5.94	5.93E-09	2.77E-03	ALU	trans
chr12	84955216	L1_umary_LINE1_2324	FAM153B	5.91	6.74E-09	3.12E-03	L1	trans
chr1	36474694	ALU_umary_ALU_102	CLEC4A	5.90	7.13E-09	3.28E-03	ALU	trans
chr1	249191472	L1_umary_LINE1_242	LRRC20	-5.90	7.45E-09	3.40E-03	L1	trans
chr6	32514622	SVA_umary_SVA_280	HLA-DRB5	-5.89	7.54E-09	3.41E-03	SVA	cis
chr3	66951653	ALU_umary_ALU_2413	IGHV3-64	5.87	8.45E-09	3.78E-03	ALU	trans
chr2	213168133	ALU_umary_ALU_1964	IGLV4-60	5.87	8.48E-09	3.78E-03	ALU	trans
chr3	180190410	ALU_umary_ALU_2977	GAPT	5.86	8.97E-09	3.94E-03	ALU	trans
chr4	40235018	ALU_umary_ALU_3239	NR2F2	5.84	1.03E-08	4.50E-03	ALU	trans
chr6	80086769	ALU_umary_ALU_5318	RP4-756H11.3	5.83	1.05E-08	4.52E-03	ALU	trans
chr17	735172	ALU_umary_ALU_11176	XCL1	5.83	1.07E-08	4.57E-03	ALU	trans
chr2	129573988	L1_umary_LINE1_398	RCCD1	5.82	1.11E-08	4.67E-03	L1	trans
chr8	69587130	ALU_umary_ALU_6797	FAT1	5.82	1.14E-08	4.68E-03	ALU	trans
chr8	80466350	ALU_umary_ALU_6867	RP11-435B5.5	5.80	1.30E-08	5.31E-03	ALU	trans
chr3	132577737	ALU_umary_ALU_2718	VIL1	5.79	1.36E-08	5.48E-03	ALU	trans
chr7	125861744	ALU_umary_ALU_6372	ACSM1	5.78	1.43E-08	5.72E-03	ALU	trans
chr6	44114176	SVA_umary_SVA_289	RP11-793H13.8	5.78	1.44E-08	5.72E-03	SVA	trans
chr15	80662580	ALU_umary_ALU_10836	LGALS9B	5.77	1.47E-08	5.78E-03	ALU	trans
chr11	5126963	SVA_umary_SVA_482	HERC2P8	5.75	1.66E-08	6.33E-03	SVA	trans
chr6	103871467	L1_umary_LINE1_1371	FITM1	5.75	1.67E-08	6.33E-03	L1	trans
chr3	109697784	ALU_umary_ALU_2606	CD2	5.74	1.72E-08	6.37E-03	ALU	trans
chr11	82657872	ALU_umary_ALU_8680	XCL1	5.74	1.73E-08	6.37E-03	ALU	trans
chr7	123194557	ALU_umary_ALU_6352	VARS	5.74	1.75E-08	6.37E-03	ALU	trans
chr6	26962174	ALU_umary_ALU_5039	RP11-457M11.5	-5.73	1.86E-08	6.73E-03	ALU	cis
chr3	6525411	ALU_umary_ALU_2124	GCLM	5.73	1.87E-08	6.73E-03	ALU	trans
chr15	76826019	ALU_umary_ALU_10819	IGKV1D-13	5.71	2.05E-08	7.32E-03	ALU	trans
chr22	23872686	ALU_umary_ALU_12453	RP11-124N14.3	5.70	2.14E-08	7.60E-03	ALU	trans
chr5	148917332	ALU_umary_ALU_4754	WSCD1	5.70	2.19E-08	7.72E-03	ALU	trans
chr6	141414904	ALU_umary_ALU_5647	PSORS1C3	5.69	2.27E-08	7.97E-03	ALU	trans
chr8	111644268	ALU_umary_ALU_7045	GDAP1	5.69	2.31E-08	8.08E-03	ALU	trans
chr3	192376526	ALU_umary_ALU_3032	IGKV2-29	5.68	2.40E-08	8.26E-03	ALU	trans
chr12	58458223	ALU_umary_ALU_9236	XRCC6BP1	5.68	2.41E-08	8.27E-03	ALU	cis
chr15	81392597	ALU_umary_ALU_10841	PSD4	-5.68	2.47E-08	8.38E-03	ALU	trans

Chr	Pos	TE	Gene	t Statistic	P-value	FDR	TE subfamily	eQTL type
chr1	165553149	L1_umary_LINE1_157	IGLC2	5.67	2.57E-08	8.60E-03	L1	trans
chr13	36325342	SVA_umary_SVA_566	NEDD9	5.67	2.57E-08	8.60E-03	SVA	trans
chr10	68217408	ALU_umary_ALU_8046	CAND2	5.66	2.68E-08	8.83E-03	ALU	trans
chr12	61247559	ALU_umary_ALU_9258	CRYBB2P1	5.66	2.68E-08	8.83E-03	ALU	trans
chr6	104772669	ALU_umary_ALU_5447	HSPE1P8	-5.66	2.74E-08	8.92E-03	ALU	trans
chr12	46175166	ALU_umary_ALU_9193	IGLV3-27	-5.66	2.78E-08	8.96E-03	ALU	trans
chr1	97717644	ALU_umary_ALU_379	IGLV1-50	5.66	2.80E-08	8.96E-03	ALU	trans
chr5	5279799	ALU_umary_ALU_4046	WEE2	5.66	2.80E-08	8.96E-03	ALU	trans
chr10	9683707	ALU_umary_ALU_7780	IGHV1-58	5.65	2.89E-08	9.19E-03	ALU	trans
chr12	21006856	ALU_umary_ALU_9055	IGHV3-20	5.65	2.94E-08	9.30E-03	ALU	trans
chr3	169270604	ALU_umary_ALU_2911	BACE1	5.65	2.96E-08	9.30E-03	ALU	trans
chr1	38447717	ALU_umary_ALU_107	UBALD1	5.64	3.06E-08	9.42E-03	ALU	trans
chr4	5667998	L1_umary_LINE1_773	POC1B	5.64	3.09E-08	9.42E-03	L1	trans
chr6	126509049	ALU_umary_ALU_5563	RP11-288E14.2	5.64	3.11E-08	9.42E-03	ALU	trans
chr5	63430716	ALU_umary_ALU_4330	SEC61A1	5.64	3.13E-08	9.42E-03	ALU	trans
chr18	56819477	ALU_umary_ALU_11733	PPIC	5.63	3.14E-08	9.42E-03	ALU	trans
chr16	79190739	ALU_umary_ALU_11136	PPP4R4	5.63	3.29E-08	9.82E-03	ALU	trans
chr9	118509752	ALU_umary_ALU_7676	HOXB7	5.62	3.37E-08	9.90E-03	ALU	trans
chr10	17712792	SVA_umary_SVA_438	TMEM236	5.61	3.65E-08	1.06E-02	SVA	cis
chr4	186361924	ALU_umary_ALU_4000	RP11-276H1.3	-5.60	3.77E-08	1.08E-02	ALU	trans
chr2	4256152	ALU_umary_ALU_974	RAB11FIP5	5.60	3.86E-08	1.10E-02	ALU	trans
chr3	112466720	SVA_umary_SVA_179	LILRA1	5.59	3.92E-08	1.10E-02	SVA	trans
chr15	28179379	ALU_umary_ALU_10586	CA9	5.59	3.92E-08	1.10E-02	ALU	trans
chr8	17943136	ALU_umary_ALU_6606	ARHGEF12	5.59	3.94E-08	1.10E-02	ALU	trans
chr7	6314063	ALU_umary_ALU_5788	REL	5.59	3.96E-08	1.10E-02	ALU	trans
chr16	52412371	ALU_umary_ALU_11042	EPHB2	5.59	4.00E-08	1.11E-02	ALU	trans
chr3	168199075	ALU_umary_ALU_2905	IQCA1	5.59	4.08E-08	1.12E-02	ALU	trans
chr18	50097757	L1_umary_LINE1_2823	TP53BP2	5.58	4.13E-08	1.13E-02	L1	trans
chr3	144610083	L1_umary_LINE1_704	IGKV2-29	5.58	4.14E-08	1.13E-02	L1	trans
chr14	66934461	ALU_umary_ALU_10374	SSPN	5.58	4.24E-08	1.14E-02	ALU	trans
chr20	32793460	ALU_umary_ALU_12140	UBLCP1	5.57	4.54E-08	1.21E-02	ALU	trans
chr13	90038686	ALU_umary_ALU_10010	PIM1	5.57	4.55E-08	1.21E-02	ALU	trans
chr11	13092200	ALU_umary_ALU_8387	RP5-902P8.10	5.56	4.70E-08	1.24E-02	ALU	trans
chr2	192982697	ALU_umary_ALU_1861	VCAN-AS1	5.55	4.83E-08	1.26E-02	ALU	trans
chr9	80704885	ALU_umary_ALU_7506	ZNF667-AS1	5.55	5.07E-08	1.32E-02	ALU	trans
chr7	65045619	SVA_umary_SVA_343	RP11-146F11.1	5.53	5.37E-08	1.38E-02	SVA	trans
chr7	85013802	ALU_umary_ALU_6166	CLCN2	5.53	5.51E-08	1.41E-02	ALU	trans
chr8	128119620	ALU_umary_ALU_7140	DHRS4-AS1	-5.53	5.53E-08	1.41E-02	ALU	trans
chr11	25108568	ALU_umary_ALU_8453	U1	5.52	5.74E-08	1.43E-02	ALU	trans
chr4	134922359	ALU_umary_ALU_3724	EGFL7	5.52	5.89E-08	1.45E-02	ALU	trans
chr7	18273084	ALU_umary_ALU_5868	TIMP2	5.51	6.11E-08	1.47E-02	ALU	trans
chr6	72753379	ALU_umary_ALU_5270	MT3	5.51	6.15E-08	1.47E-02	ALU	trans

Chr	Pos	TE	Gene	t Statistic	P-value	FDR	TE subfamily	eQTL type
chr10	73193547	ALU_umary_ALU_8081	MAML2	-5.50	6.36E-08	1.51E-02	ALU	trans
chr5	116763098	ALU_umary_ALU_4607	SPTLC3	5.50	6.42E-08	1.52E-02	ALU	trans
chr14	30525258	L1_umary_LINE1_2526	TMEM63C	5.50	6.53E-08	1.54E-02	L1	trans
chr8	122155287	ALU_umary_ALU_7101	P2RY2	5.49	6.64E-08	1.56E-02	ALU	trans
chr10	21192833	ALU_umary_ALU_7830	TMEM38A	5.49	6.81E-08	1.59E-02	ALU	trans
chr8	8920127	ALU_umary_ALU_6560	AF131215.2	5.48	7.12E-08	1.64E-02	ALU	trans
chr20	3242418	ALU_umary_ALU_12024	IGKV2-29	5.48	7.17E-08	1.65E-02	ALU	trans
chr6	32773435	ALU_umary_ALU_5079	TAP2	5.47	7.70E-08	1.76E-02	ALU	cis
chr5	146369626	ALU_umary_ALU_4737	RALA	-5.46	7.88E-08	1.79E-02	ALU	trans
chr7	126661404	ALU_umary_ALU_6375	OTOF	5.46	7.94E-08	1.79E-02	ALU	trans
chr3	155818384	ALU_umary_ALU_2834	IGHV1-69	5.46	8.01E-08	1.79E-02	ALU	trans
chr3	44340304	ALU_umary_ALU_2324	MAPK9	5.46	8.02E-08	1.79E-02	ALU	trans
chr4	106550763	ALU_umary_ALU_3575	SEMA5A	5.45	8.45E-08	1.86E-02	ALU	trans
chr12	75931101	ALU_umary_ALU_9345	ENPP5	-5.44	8.76E-08	1.91E-02	ALU	trans
chr8	88410445	ALU_umary_ALU_6917	SNORA57	5.43	9.22E-08	1.99E-02	ALU	trans
chr11	114686177	ALU_umary_ALU_8857	RP11-566K19.6	5.43	9.30E-08	1.99E-02	ALU	trans
chr7	52789472	L1_umary_LINE1_1509	N6AMT2	5.43	9.33E-08	1.99E-02	L1	trans
chr2	212707197	L1_umary_LINE1_504	GNG12	5.43	9.36E-08	1.99E-02	L1	trans
chr14	51719000	ALU_umary_ALU_10302	TMX1	5.43	9.41E-08	2.00E-02	ALU	cis
chr4	86099317	ALU_umary_ALU_3472	PPFIBP1	5.43	9.49E-08	2.00E-02	ALU	trans
chr11	73562793	ALU_umary_ALU_8639	MRPL48	5.42	9.84E-08	2.04E-02	ALU	cis
chr1	51029967	ALU_umary_ALU_156	EPCAM	5.42	9.93E-08	2.05E-02	ALU	trans
chr5	122976151	ALU_umary_ALU_4639	RP11-575G13.2	5.42	9.95E-08	2.05E-02	ALU	trans
chr8	42039906	ALU_umary_ALU_6700	ENPP4	-5.41	1.03E-07	2.10E-02	ALU	trans
chr1	65572274	ALU_umary_ALU_214	MCOLN2	5.40	1.09E-07	2.16E-02	ALU	trans
chr2	189853875	ALU_umary_ALU_1846	SLC5A5	5.40	1.09E-07	2.16E-02	ALU	trans
chr1	169524859	L1_umary_LINE1_164	ATP5H	5.40	1.11E-07	2.17E-02	L1	trans
chr10	105817214	ALU_umary_ALU_8203	AC005740.3	-5.40	1.11E-07	2.17E-02	ALU	trans
chr5	116897151	ALU_umary_ALU_4608	LRRC20	-5.39	1.12E-07	2.17E-02	ALU	trans
chr14	30854683	L1_umary_LINE1_2529	ADAD2	5.39	1.13E-07	2.17E-02	L1	trans
chr3	5484122	ALU_umary_ALU_2116	RANBP17	5.39	1.15E-07	2.20E-02	ALU	trans
chr10	97668538	ALU_umary_ALU_8182	ENTPD1	5.39	1.16E-07	2.21E-02	ALU	cis
chr3	193354185	L1_umary_LINE1_769	RP11-175P19.2	5.39	1.16E-07	2.22E-02	L1	cis
chr4	86607956	L1_umary_LINE1_885	IGHV3-32	5.38	1.19E-07	2.26E-02	L1	trans
chr1	184950473	ALU_umary_ALU_643	MDK	5.38	1.19E-07	2.26E-02	ALU	trans
chr20	17860937	L1_umary_LINE1_2915	NPRL3	-5.37	1.26E-07	2.35E-02	L1	trans
chr22	35180750	ALU_umary_ALU_12488	CCDC3	5.37	1.26E-07	2.35E-02	ALU	trans
chr17	49408800	ALU_umary_ALU_11342	POTEE	5.37	1.27E-07	2.36E-02	ALU	trans
chr2	186001780	ALU_umary_ALU_1824	GPR183	5.37	1.28E-07	2.36E-02	ALU	trans
chr7	142673652	ALU_umary_ALU_6458	RP4-730K3.3	5.37	1.28E-07	2.36E-02	ALU	trans
chr12	78396948	ALU_umary_ALU_9357	TIMP2	5.37	1.30E-07	2.39E-02	ALU	trans
chr7	46504534	ALU_umary_ALU_6008	RP11-452L6.7	5.36	1.32E-07	2.39E-02	ALU	trans

Chr	Pos	TE	Gene	t Statistic	P-value	FDR	TE subfamily	eQTL type
chr15	79167169	ALU_umary_ALU_10832	ADAMTS7	5.36	1.35E-07	2.44E-02	ALU	cis
chr20	26190974	ALU_umary_ALU_12132	LINC00969	5.36	1.36E-07	2.44E-02	ALU	trans
chr11	90530723	ALU_umary_ALU_8730	FN1	5.36	1.37E-07	2.45E-02	ALU	trans
chr6	102220590	ALU_umary_ALU_5428	CLDN12	5.35	1.39E-07	2.48E-02	ALU	trans
chr11	88908952	ALU_umary_ALU_8717	HMGB1P1	5.35	1.40E-07	2.48E-02	ALU	trans
chr3	76222309	ALU_umary_ALU_2455	CREM	5.35	1.40E-07	2.48E-02	ALU	trans
chr11	100911793	ALU_umary_ALU_8790	USP32	-5.34	1.46E-07	2.57E-02	ALU	trans
chr21	28249247	ALU_umary_ALU_12342	LZIC	5.34	1.49E-07	2.59E-02	ALU	trans
chr21	26354237	ALU_umary_ALU_12333	NEDD9	-5.34	1.50E-07	2.59E-02	ALU	trans
chr13	50912089	ALU_umary_ALU_9780	FBN2	5.34	1.52E-07	2.61E-02	ALU	trans
chr15	53956115	L1_umary_LINE1_2651	TPCN2	-5.33	1.58E-07	2.68E-02	L1	trans
chr21	27163622	SVA_umary_SVA_808	RP11-22P6.3	5.32	1.65E-07	2.76E-02	SVA	trans
chr11	23572069	ALU_umary_ALU_8441	SEMA3A	5.31	1.73E-07	2.85E-02	ALU	trans
chr19	18835612	SVA_umary_SVA_743	SSTR3	5.31	1.73E-07	2.85E-02	SVA	trans
chr1	112992009	ALU_umary_ALU_454	CTTNBP2NL	-5.31	1.74E-07	2.85E-02	ALU	cis
chr17	70696176	ALU_umary_ALU_11447	TMEM159	5.31	1.74E-07	2.85E-02	ALU	trans
chr11	17683506	ALU_umary_ALU_8864	TINAG	5.31	1.77E-07	2.88E-02	ALU	trans
chr8	115604486	ALU_umary_ALU_7074	RASSF8	5.31	1.79E-07	2.89E-02	ALU	trans
chr3	195829642	SVA_umary_SVA_202	PDLIM4	5.30	1.79E-07	2.89E-02	SVA	trans
chr20	47206176	ALU_umary_ALU_12190	IGHV1OR15-2	5.30	1.80E-07	2.89E-02	ALU	trans
chr1	112103110	ALU_umary_ALU_444	NPIPB15	5.30	1.80E-07	2.89E-02	ALU	trans
chr11	47806655	ALU_umary_ALU_8566	TBC1D1	-5.30	1.83E-07	2.90E-02	ALU	trans
chr1	102920544	ALU_umary_ALU_412	RP11-670E13.5	5.30	1.84E-07	2.90E-02	ALU	trans
chr8	87770103	ALU_umary_ALU_6912	IGKV1-13	5.30	1.88E-07	2.94E-02	ALU	trans
chr5	56830734	ALU_umary_ALU_4298	CXCL9	5.29	1.95E-07	3.03E-02	ALU	trans
chr6	153633385	ALU_umary_ALU_5701	TRAPPC8	5.29	1.97E-07	3.04E-02	ALU	trans
chr4	190684483	SVA_umary_SVA_231	IGLV10-54	-5.28	2.02E-07	3.09E-02	SVA	trans
chr6	114091219	L1_umary_LINE1_1379	GVINP1	5.28	2.02E-07	3.09E-02	L1	trans
chr4	79289449	ALU_umary_ALU_3430	RP11-165P7.1	5.28	2.04E-07	3.11E-02	ALU	trans
chr2	114106446	ALU_umary_ALU_1441	ACY1	5.28	2.05E-07	3.12E-02	ALU	trans
chr10	83042769	ALU_umary_ALU_8107	NEK4	5.28	2.08E-07	3.14E-02	ALU	trans
chr1	212765848	ALU_umary_ALU_767	AC007551.2	5.27	2.10E-07	3.16E-02	ALU	trans
chr20	15513281	ALU_umary_ALU_12084	CTD-3065B20.2	5.27	2.19E-07	3.26E-02	ALU	trans
chr18	15325826	ALU_umary_ALU_11565	FXR2	5.26	2.23E-07	3.30E-02	ALU	trans
chr3	157369140	ALU_umary_ALU_2844	RGS22	5.26	2.26E-07	3.31E-02	ALU	trans
chr3	144308219	ALU_umary_ALU_2765	FITM1	5.26	2.27E-07	3.31E-02	ALU	trans
chr4	43399986	ALU_umary_ALU_3259	TPCN2	5.25	2.32E-07	3.34E-02	ALU	trans
chr8	87315040	L1_umary_LINE1_1704	EMBP1	-5.25	2.32E-07	3.34E-02	L1	trans
chr8	110101605	ALU_umary_ALU_7037	AC007881.4	5.25	2.32E-07	3.34E-02	ALU	trans
chr11	103736375	ALU_umary_ALU_8803	RP11-578F21.9	5.25	2.36E-07	3.37E-02	ALU	trans
chr8	75448823	ALU_umary_ALU_6841	HNRNPA1P10	5.25	2.39E-07	3.40E-02	ALU	trans
chr3	186372141	L1_umary_LINE1_761	TBC1D1	5.25	2.40E-07	3.40E-02	L1	trans

Chr	Pos	TE	Gene	t Statistic	P-value	FDR	TE subfamily	eQTL type
chr12	43256994	L1_umary_LINE1_2278	MAML2	-5.25	2.40E-07	3.40E-02	L1	trans
chr15	62606061	ALU_umary_ALU_10767	SNORA73B	5.24	2.44E-07	3.44E-02	ALU	trans
chr21	28221359	ALU_umary_ALU_12341	KANSL1-AS1	5.24	2.45E-07	3.46E-02	ALU	trans
chr12	70788935	ALU_umary_ALU_9313	CDC5L	-5.24	2.46E-07	3.46E-02	ALU	trans
chr6	126423676	ALU_umary_ALU_5562	WSCD1	5.24	2.47E-07	3.47E-02	ALU	trans
chr1	156705360	ALU_umary_ALU_528	CEP72	5.24	2.49E-07	3.48E-02	ALU	trans
chr11	7352462	ALU_umary_ALU_8361	PCDHGA7	5.24	2.52E-07	3.49E-02	ALU	trans
chr10	48419952	ALU_umary_ALU_7936	UAP1L1	5.24	2.55E-07	3.52E-02	ALU	trans
chr5	169376692	SVA_umary_SVA_268	PPFIA3	5.23	2.60E-07	3.54E-02	SVA	trans
chr7	42647660	ALU_umary_ALU_5994	RP11-336N8.4	-5.23	2.61E-07	3.54E-02	ALU	trans
chr4	26438630	ALU_umary_ALU_3159	RALA	-5.23	2.61E-07	3.54E-02	ALU	trans
chr2	126051106	ALU_umary_ALU_1497	ATP8B1	5.23	2.61E-07	3.54E-02	ALU	trans
chr8	17336021	ALU_umary_ALU_6603	PKIB	5.23	2.63E-07	3.56E-02	ALU	trans
chr6	124865521	ALU_umary_ALU_5553	SOX30	5.23	2.68E-07	3.60E-02	ALU	trans
chr5	137130159	L1_umary_LINE1_1215	RP11-578F21.10	5.23	2.68E-07	3.60E-02	L1	trans
chr4	100740869	ALU_umary_ALU_3545	AOAH	5.22	2.71E-07	3.64E-02	ALU	trans
chr6	122157421	SVA_umary_SVA_305	RP11-124N14.3	-5.22	2.79E-07	3.71E-02	SVA	trans
chr20	9605206	ALU_umary_ALU_12057	IGHV4-61	5.22	2.82E-07	3.74E-02	ALU	trans
chr1	100994221	ALU_umary_ALU_402	PHF2	-5.22	2.82E-07	3.74E-02	ALU	trans
chr3	29550384	ALU_umary_ALU_2253	ST3GAL6-AS1	5.21	2.86E-07	3.77E-02	ALU	trans
chr6	87678642	ALU_umary_ALU_5348	IGHV7-81	5.21	2.90E-07	3.82E-02	ALU	trans
chr3	47782873	ALU_umary_ALU_2333	PLA2G4C	-5.21	2.95E-07	3.87E-02	ALU	trans
chr2	226970099	ALU_umary_ALU_2028	FERMT1	5.21	2.96E-07	3.88E-02	ALU	trans
chr1	58071869	ALU_umary_ALU_182	CD86	-5.20	2.99E-07	3.90E-02	ALU	trans
chr10	14396698	ALU_umary_ALU_7798	ZBTB38	5.20	3.05E-07	3.95E-02	ALU	trans
chr16	74897618	ALU_umary_ALU_11115	GUCY1B3	5.20	3.05E-07	3.95E-02	ALU	trans
chr8	79813676	L1_umary_LINE1_1693	ABCA4	5.20	3.11E-07	3.97E-02	L1	trans
chr4	81899822	ALU_umary_ALU_3448	SAPCD2	-5.19	3.14E-07	3.99E-02	ALU	trans
chr8	77248370	ALU_umary_ALU_6854	SPRED2	5.19	3.15E-07	3.99E-02	ALU	trans
chr4	135130281	ALU_umary_ALU_3726	EGFL7	5.19	3.19E-07	4.03E-02	ALU	trans
chr5	89450997	L1_umary_LINE1_1164	ST8SIA1	-5.19	3.20E-07	4.03E-02	L1	trans
chr18	5411665	ALU_umary_ALU_11513	RHBDL1	5.19	3.20E-07	4.03E-02	ALU	trans
chr2	160159011	ALU_umary_ALU_1684	LINC00638	5.19	3.30E-07	4.11E-02	ALU	trans
chr6	46310306	L1_umary_LINE1_1293	MAML2	-5.17	3.49E-07	4.29E-02	L1	trans
chr5	33633348	L1_umary_LINE1_1089	MAML2	-5.17	3.55E-07	4.32E-02	L1	trans
chr8	50147620	ALU_umary_ALU_6712	RP11-124N14.3	-5.17	3.55E-07	4.32E-02	ALU	trans
chr5	91031026	ALU_umary_ALU_4463	NRIP1	5.17	3.56E-07	4.32E-02	ALU	trans
chr15	24591239	ALU_umary_ALU_10560	LINC00649	5.17	3.61E-07	4.36E-02	ALU	trans
chr1	64850193	SVA_umary_SVA_35	NMT2	-5.17	3.61E-07	4.36E-02	SVA	trans
chr11	25605138	ALU_umary_ALU_8456	AC078899.1	-5.16	3.70E-07	4.42E-02	ALU	trans
chr17	54947583	ALU_umary_ALU_11371	DGKE	-5.16	3.70E-07	4.42E-02	ALU	cis
chr14	22578806	ALU_umary_ALU_10141	PRIM1	-5.16	3.70E-07	4.42E-02	ALU	trans

Chr	Pos	TE	Gene	t Statistic	P-value	FDR	TE subfamily	eQTL type
chr3	82615586	SVA_umary_SVA_174	BTG1	5.16	3.70E-07	4.42E-02	SVA	trans
chr5	55167661	ALU_umary_ALU_4287	RP11-122K13.12	5.16	3.74E-07	4.45E-02	ALU	trans
chr1	64958078	ALU_umary_ALU_211	SAPCD2	5.16	3.76E-07	4.46E-02	ALU	trans
chr4	76993824	ALU_umary_ALU_3412	RAB3IP	5.16	3.83E-07	4.51E-02	ALU	trans
chr9	94058487	L1_umary_LINE1_1863	IGLV2-34	5.15	3.84E-07	4.51E-02	L1	trans
chr10	29105298	ALU_umary_ALU_7874	FTH1P12	-5.15	3.86E-07	4.52E-02	ALU	trans
chr1	178495286	ALU_umary_ALU_620	BRD7P3	5.15	3.91E-07	4.56E-02	ALU	trans
chr3	124568685	ALU_umary_ALU_2682	CHI3L1	5.15	3.92E-07	4.56E-02	ALU	trans
chr8	50550379	ALU_umary_ALU_6716	LGALS4	5.15	3.96E-07	4.58E-02	ALU	trans
chr10	58902214	ALU_umary_ALU_7994	DDI2	5.15	3.98E-07	4.59E-02	ALU	trans
chr10	94571333	ALU_umary_ALU_8168	ANKRD36BP2	5.14	4.13E-07	4.70E-02	ALU	trans
chr6	32687172	ALU_umary_ALU_5076	ATP4B	5.14	4.18E-07	4.73E-02	ALU	trans
chr3	137565396	ALU_umary_ALU_2733	FAM151B	5.14	4.18E-07	4.73E-02	ALU	trans
chr13	55438087	ALU_umary_ALU_9798	PCDHGA3	5.14	4.20E-07	4.73E-02	ALU	trans
chr6	153859178	ALU_umary_ALU_5702	PCBP3	5.14	4.20E-07	4.73E-02	ALU	trans
chr8	1310538	ALU_umary_ALU_6532	SAMD3	5.13	4.35E-07	4.83E-02	ALU	trans
chr11	9758001	ALU_umary_ALU_8374	SOCS2-AS1	5.13	4.36E-07	4.83E-02	ALU	trans
chr7	10215359	ALU_umary_ALU_5815	MEST	5.13	4.36E-07	4.83E-02	ALU	trans
chr8	107230078	ALU_umary_ALU_7010	TBC1D9B	5.13	4.37E-07	4.83E-02	ALU	trans
chr12	10234202	ALU_umary_ALU_8992	DHRS4-AS1	5.13	4.40E-07	4.85E-02	ALU	trans
chr7	128215395	SVA_umary_SVA_355	IGLL1	5.12	4.49E-07	4.91E-02	SVA	trans
chr4	176281836	ALU_umary_ALU_3956	SLC22A31	5.12	4.52E-07	4.93E-02	ALU	trans
chr17	43818955	ALU_umary_ALU_11327	AC022182.1	5.12	4.55E-07	4.94E-02	ALU	trans
chr4	180698483	ALU_umary_ALU_3979	CTSF	5.12	4.56E-07	4.94E-02	ALU	trans
chr9	72777422	L1_umary_LINE1_1833	SNORD3B-2	5.12	4.58E-07	4.94E-02	L1	trans
chr4	123796599	ALU_umary_ALU_3659	PNPLA3	5.12	4.60E-07	4.94E-02	ALU	trans
chr11	55771332	ALU_umary_ALU_8585	CXCL12	5.12	4.62E-07	4.94E-02	ALU	trans
chr2	212708356	ALU_umary_ALU_1959	CECR2	5.12	4.67E-07	4.97E-02	ALU	trans
chr6	101194045	SVA_umary_SVA_298	PDK2	5.11	4.71E-07	4.99E-02	SVA	trans
chr4	175534305	ALU_umary_ALU_3953	CDK11A	5.11	4.71E-07	4.99E-02	ALU	trans

Table 4 Functional enrichment of genes that are associated with TE-eQTL

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
KEGG INTESTINAL IMMUNE NETWORK FOR IGA PRODUCTION	48	Intestinal immune network for IgA production	11	0.2292	4.08E-14	4.40E-11
KEGG TYPE I DIABETES MELLITUS	44	Type I diabetes mellitus	8	0.1818	9.41E-10	5.07E-07
KEGG CELL ADHESION MOLECULES CAMS	134	Cell adhesion molecules (CAMs)	11	0.0821	4.29E-09	1.54E-06
KEGG ALLOGRAFT REJECTION	38	Allograft rejection	7	0.1842	9.83E-09	2.65E-06
KEGG GRAFT VERSUS HOST DISEASE	42	Graft-versus-host disease	7	0.1667	2.05E-08	4.42E-06
REACTOME TRANSLOCATION OF ZAP 70 TO IMMUNOLOGICAL SYNAPSE	14	Genes involved in Translocation of ZAP-70 to Immunological synapse	5	0.3571	3.57E-08	6.41E-06
KEGG LEISHMANIA INFECTION	72	Leishmania infection	8	0.1111	5.34E-08	8.21E-06
KEGG ASTHMA	30	Asthma	6	0.2	6.85E-08	9.21E-06
REACTOME PHOSPHORYLATION OF CD3 AND TCR ZETA CHAINS	16	Genes involved in Phosphorylation of CD3 and TCR zeta chains	5	0.3125	7.70E-08	9.21E-06
KEGG AUTOIMMUNE THYROID DISEASE	53	Autoimmune thyroid disease	7	0.1321	1.09E-07	1.18E-05
REACTOME PD1 SIGNALING	18	Genes involved in PD-1 signaling	5	0.2778	1.49E-07	1.46E-05
REACTOME COSTIMULATION BY THE CD28 FAMILY	63	Genes involved in Costimulation by the CD28 family	7	0.1111	3.70E-07	3.32E-05
KEGG VIRAL MYOCARDITIS	73	Viral myocarditis	7	0.0959	1.02E-06	8.48E-05
REACTOME GENERATION OF SECOND MESSENGER MOLECULES	27	Genes involved in Generation of second messenger molecules	5	0.1852	1.33E-06	1.03E-04
REACTOME ADAPTIVE IMMUNE SYSTEM	539	Genes involved in Adaptive Immune System	16	0.0297	2.40E-06	1.72E-04
KEGG ANTIGEN PROCESSING AND PRESENTATION	89	Antigen processing and presentation	7	0.0787	3.92E-06	2.64E-04
REACTOME MHC CLASS II ANTIGEN PRESENTATION	91	Genes involved in MHC class II antigen presentation	7	0.0769	4.55E-06	2.88E-04
REACTOME IMMUNE SYSTEM	933	Genes involved in Immune System	21	0.0225	5.29E-06	3.17E-04
REACTOME INTERFERON GAMMA SIGNALING	63	Genes involved in Interferon gamma signaling	6	0.0952	6.42E-06	3.64E-04

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
REACTOME DOWNSTREAM TCR SIGNALING	37	Genes involved in Downstream TCR signaling	5	0.1351	6.78E-06	3.65E-04
REACTOME TCR SIGNALING	54	Genes involved in TCR signaling	5	0.0926	4.45E-05	2.28E-03
KEGG SYSTEMIC LUPUS ERYTHEMATOSUS	140	Systemic lupus erythematosus	7	0.05	7.48E-05	3.66E-03
KEGG PYRIMIDINE METABOLISM	98	Pyrimidine metabolism	6	0.0612	8.05E-05	3.77E-03
KEGG PRIMARY IMMUNODEFICIENCY	35	Primary immunodeficiency	4	0.1143	1.16E-04	5.21E-03
BIOCARTA TH1TH2 PATHWAY	19	Th1/Th2 Differentiation	3	0.1579	3.29E-04	1.42E-02
BIOCARTA CTLA4 PATHWAY	21	The Co-Stimulatory Signal During T-cell Activation	3	0.1429	4.47E-04	1.85E-02
KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION	267	Cytokine-cytokine receptor interaction	8	0.03	7.67E-04	2.97E-02
REACTOME CHEMOKINE RECEPTORS BIND CHEMOKINES	57	Genes involved in Chemokine receptors bind chemokines	4	0.0702	7.73E-04	2.97E-02
REACTOME GAP JUNCTION TRAFFICKING	27	Genes involved in Gap junction trafficking	3	0.1111	9.53E-04	3.54E-02
KEGG PURINE METABOLISM	159	Purine metabolism	6	0.0377	1.08E-03	3.76E-02
REACTOME INTERFERON SIGNALING	159	Genes involved in Interferon Signaling	6	0.0377	1.08E-03	3.76E-02

Table 5 Results for conditional association controls

SNP	TE	Gene	Gene ID	Conditional P-value	Group
rs113175928	ALU_umary_ALU_6806	XKR9	ENSG00000221947	1.12E-22	EUR
rs56289286	SVA_umary_SVA_706	LRRC37A4P	ENSG00000214425	7.36E-15	EUR
rs140032472	ALU_umary_ALU_9602	LINC00422	ENSG00000224429	2.81E-14	EUR
rs150156751	ALU_umary_ALU_11979	ZNF880	ENSG00000221923	4.93E-13	EUR
rs11485298	L1_umary_LINE1_61	CRYZ	ENSG00000116791	1.67E-12	EUR
rs138311123	SVA_umary_SVA_706	LRRC37A	ENSG00000176681	1.21E-10	EUR
rs138311123	SVA_umary_SVA_705	LRRC37A	ENSG00000176681	2.81E-09	EUR
rs56289286	SVA_umary_SVA_705	LRRC37A4P	ENSG00000214425	8.64E-09	EUR
rs9892659	ALU_umary_ALU_11371	DGKE	ENSG00000153933	7.31E-08	EUR
rs138311123	SVA_umary_SVA_704	LRRC37A	ENSG00000176681	1.37E-07	EUR
rs56289286	SVA_umary_SVA_704	LRRC37A4P	ENSG00000214425	4.27E-07	EUR
rs116628019	SVA_umary_SVA_282	HLA-DPA1	ENSG00000231389	5.16E-07	EUR
rs12104210	ALU_umary_ALU_11997	ZFP28	ENSG00000196867	6.41E-07	EUR
rs12150699	ALU_umary_ALU_11754	PIGN	ENSG00000197563	7.27E-07	EUR
rs138311123	SVA_umary_SVA_706	LRRC37A2	ENSG00000238083	1.37E-06	EUR
rs35854157	ALU_umary_ALU_1256	SLC1A4	ENSG00000115902	3.79E-06	EUR
rs71664293	ALU_umary_ALU_3295	SRD5A3	ENSG00000128039	6.06E-06	EUR
rs138311123	SVA_umary_SVA_705	LRRC37A2	ENSG00000238083	7.40E-06	EUR
rs9399043	ALU_umary_ALU_5602	RPS12	ENSG00000112306	9.13E-06	EUR
rs140032472	ALU_umary_ALU_9603	LINC00422	ENSG00000224429	1.19E-05	EUR
rs142039218	SVA_umary_SVA_706	GOSR2	ENSG00000108433	1.44E-05	EUR
rs11714944	ALU_umary_ALU_2337	NME6	ENSG00000172113	2.11E-05	EUR
rs116405062	ALU_umary_ALU_5075	HLA-DRB1	ENSG00000196126	4.54E-05	EUR
rs9274660	ALU_umary_ALU_5075	HLA-DQB1	ENSG00000179344	6.26E-05	EUR
rs116786525	ALU_umary_ALU_5053	HLA-G	ENSG00000204632	9.13E-05	EUR
rs35801758	ALU_umary_ALU_12145	CPNE1	ENSG00000214078	1.03E-04	EUR
rs11070297	ALU_umary_ALU_10654	ZFYVE19	ENSG00000166140	3.16E-04	EUR
rs175037	ALU_umary_ALU_10401	EIF2B2	ENSG00000119718	3.72E-04	EUR
rs1660559	ALU_umary_ALU_884	RP4-781K5.6	ENSG00000230628	8.37E-04	EUR
rs399970	ALU_umary_ALU_11997	ZNF470	ENSG00000197016	9.49E-04	EUR
rs1061810	ALU_umary_ALU_8559	HSD17B12	ENSG00000149084	1.93E-03	EUR
rs11160042	SVA_umary_SVA_615	NDUFB1	ENSG00000183648	2.31E-03	EUR
rs2647071	SVA_umary_SVA_280	HLA-DRB5	ENSG00000198502	2.61E-03	EUR
rs10193212	L1_umary_LINE1_477	MFSD6	ENSG00000151690	2.88E-03	EUR
rs10431506	ALU_umary_ALU_9236	XRCC6BP1	ENSG00000166896	3.26E-03	EUR
rs2647071	ALU_umary_ALU_5072	HLA-DRB5	ENSG00000198502	4.17E-03	EUR
rs12629	ALU_umary_ALU_970	RNASEH1-AS1	ENSG00000234171	4.24E-03	EUR
rs3824458	SVA_umary_SVA_401	B4GALT1	ENSG00000086062	4.25E-03	EUR
rs2647071	ALU_umary_ALU_5079	HLA-DRB5	ENSG00000198502	5.18E-03	EUR

SNP	TE	Gene	Gene ID	Conditional P-value	Group
rs10431506	ALU_umary_ALU_9234	XRCC6BP1	ENSG00000166896	5.51E-03	EUR
rs139480590	SVA_umary_SVA_706	CRHR1	ENSG00000120088	7.85E-03	EUR
rs74687091	ALU_umary_ALU_7578	LINC00476	ENSG00000175611	1.19E-02	EUR
rs2621323	ALU_umary_ALU_5079	TAP2	ENSG00000204267	1.39E-02	EUR
rs4634177	SVA_umary_SVA_203	ZNF718	ENSG00000250312	2.40E-02	EUR
rs56298119	ALU_umary_ALU_10302	TMX1	ENSG00000139921	2.70E-02	EUR
rs7679215	ALU_umary_ALU_3997	CENPU	ENSG00000151725	3.72E-02	EUR
rs4309324	ALU_umary_ALU_10302	TRIM9	ENSG00000100505	5.14E-02	EUR
rs71526018	ALU_umary_ALU_5888	CCDC126	ENSG00000169193	5.61E-02	EUR
rs12200674	ALU_umary_ALU_5583	TMEM200A	ENSG00000164484	6.92E-02	EUR
rs74904447	ALU_umary_ALU_6178	TP53TG1	ENSG00000182165	7.76E-02	AFR
rs10973388	ALU_umary_ALU_7420	POLR1E	ENSG00000137054	1.38E-01	EUR
rs7326010	ALU_umary_ALU_9777	ARL11	ENSG00000152213	2.26E-01	AFR
rs3176891	ALU_umary_ALU_8182	ENTPD1	ENSG00000138185	2.53E-01	EUR
rs199444	SVA_umary_SVA_706	WNT3	ENSG00000108379	2.71E-01	EUR
rs2647071	ALU_umary_ALU_5075	HLA-DRB5	ENSG00000198502	3.07E-01	EUR
rs35084227	L1_umary_LINE1_353	PLGLB1	ENSG00000183281	4.77E-01	EUR
rs887307	SVA_umary_SVA_532	AKAP3	ENSG00000111254	4.90E-01	EUR
rs139348312	SVA_umary_SVA_5	C1orf167	ENSG00000215910	5.52E-01	EUR
rs3998867	SVA_umary_SVA_450	SLC25A16	ENSG00000122912	5.60E-01	EUR

Table 6 Results for regional association controls

Chr	Pos	TE	Gene	P-value	Global FDR	Regional FDR	TE subfamily	eQTL type
chr14	82223469	ALU_umary_ALU_10437	LRRC20	7.57E-10	1.09E-05	5.77E-04	ALU	trans
chr8	133725454	ALU_umary_ALU_7169	FAM65B	1.06E-09	9.31E-06	7.30E-04	ALU	trans
chr15	80662580	ALU_umary_ALU_10836	LGALS9B	1.47E-08	2.20E-04	5.78E-03	ALU	trans
chr3	6525411	ALU_umary_ALU_2124	GCLM	1.87E-08	3.24E-04	6.73E-03	ALU	trans
chr15	81392597	ALU_umary_ALU_10841	PSD4	2.47E-08	3.53E-04	8.38E-03	ALU	trans
chr6	104772669	ALU_umary_ALU_5447	HSPE1P8	2.74E-08	3.80E-04	8.92E-03	ALU	trans
chr7	6314063	ALU_umary_ALU_5788	REL	3.96E-08	4.24E-04	1.10E-02	ALU	trans
chr18	50097757	L1_umary_LINE1_2823	TP53BP2	4.13E-08	2.54E-04	1.13E-02	L1	trans
chr14	66934461	ALU_umary_ALU_10374	SSPN	4.24E-08	2.60E-04	1.14E-02	ALU	trans
chr3	44340304	ALU_umary_ALU_2324	MAPK9	8.02E-08	9.20E-04	1.79E-02	ALU	trans
chr12	75931101	ALU_umary_ALU_9345	ENPP5	8.76E-08	1.21E-03	1.91E-02	ALU	trans
chr7	52789472	L1_umary_LINE1_1509	N6AMT2	9.33E-08	1.57E-03	1.99E-02	L1	trans
chr4	86099317	ALU_umary_ALU_3472	PPFIBP1	9.49E-08	1.15E-03	2.00E-02	ALU	trans
chr10	105817214	ALU_umary_ALU_8203	AC005740.3	1.11E-07	1.46E-03	2.17E-02	ALU	trans
chr20	17860937	L1_umary_LINE1_2915	NPRL3	1.26E-07	9.48E-04	2.35E-02	L1	trans
chr11	88908952	ALU_umary_ALU_8717	HMGB1P1	1.40E-07	2.11E-03	2.48E-02	ALU	trans
chr11	100911793	ALU_umary_ALU_8790	USP32	1.46E-07	5.68E-04	2.57E-02	ALU	trans
chr19	18835612	SVA_umary_SVA_743	SSTR3	1.73E-07	2.56E-03	2.85E-02	SVA	trans
chr17	70696176	ALU_umary_ALU_11447	TMEM159	1.74E-07	2.73E-03	2.85E-02	ALU	trans
chr1	102920544	ALU_umary_ALU_412	RP11-670E13.5	1.84E-07	2.75E-03	2.90E-02	ALU	trans
chr8	87315040	L1_umary_LINE1_1704	EMBP1	2.32E-07	2.02E-03	3.34E-02	L1	trans
chr3	186372141	L1_umary_LINE1_761	TBC1D1	2.40E-07	3.56E-03	3.40E-02	L1	trans
chr21	28221359	ALU_umary_ALU_12341	KANSL1-AS1	2.45E-07	1.82E-03	3.46E-02	ALU	trans
chr12	70788935	ALU_umary_ALU_9313	CDC5L	2.46E-07	2.52E-03	3.46E-02	ALU	trans
chr7	42647660	ALU_umary_ALU_5994	RP11-336N8.4	2.61E-07	3.70E-03	3.54E-02	ALU	trans
chr10	14396698	ALU_umary_ALU_7798	ZBTB38	3.05E-07	2.90E-03	3.95E-02	ALU	trans
chr2	160159011	ALU_umary_ALU_1684	LINC00638	3.30E-07	4.47E-03	4.11E-02	ALU	trans
chr3	82615586	SVA_umary_SVA_174	BTG1	3.70E-07	5.18E-03	4.42E-02	SVA	trans
chr4	76993824	ALU_umary_ALU_3412	RAB3IP	3.83E-07	5.26E-03	4.51E-02	ALU	trans
chr1	178495286	ALU_umary_ALU_620	BRD7P3	3.91E-07	2.63E-03	4.56E-02	ALU	trans
chr8	107230078	ALU_umary_ALU_7010	TBC1D9B	4.37E-07	5.80E-03	4.83E-02	ALU	trans
chr6	101194045	SVA_umary_SVA_298	PDK2	4.71E-07	6.17E-03	4.99E-02	SVA	trans

Chr	Pos	TE	Gene	P-value	Global FDR	Regional FDR	TE subfamily	eQTL type
chr4	175534305	ALU_umary_ALU_3953	CDK11A	4.71E-07	6.61E-03	4.99E-02	ALU	trans

Table 7 eQTL results for known Pax5 target genes that are associated with Alu-7481

Gene	t Statistic	P-value	FDR	eQTL type
PIK3AP1	5.63	3.13E-08	9.42E-03	trans
ZSCAN23	5.45	8.54E-08	1.87E-02	trans
REL	5.43	9.55E-08	2.00E-02	trans
TBC1D1	5.03	7.15E-07	6.16E-02	trans
GPR183	4.84	1.80E-06	9.49E-02	trans
CCDC19	4.78	2.43E-06	1.08E-01	trans
RAB27A	4.74	2.93E-06	1.18E-01	trans
MYO7B	4.71	3.30E-06	1.24E-01	trans
ARHGEF7	4.67	3.92E-06	1.33E-01	trans
C7orf50	4.47	9.93E-06	1.89E-01	trans
STK17B	4.43	1.18E-05	2.03E-01	trans
CHMP4B	4.36	1.63E-05	2.26E-01	trans
TMEM38A	4.36	1.65E-05	2.26E-01	trans
ACTR2	4.30	2.08E-05	2.44E-01	trans
TMEM37	4.28	2.27E-05	2.51E-01	trans
BMPR1A	4.24	2.75E-05	2.66E-01	trans
NOL7	4.16	3.82E-05	2.91E-01	trans
SEMA7A	4.15	4.00E-05	2.94E-01	trans
B3GNT2	4.06	5.72E-05	3.26E-01	trans
PIKFYVE	4.06	5.77E-05	3.27E-01	trans
CAPZA1	4.06	5.86E-05	3.27E-01	trans
ESAM	4.06	5.88E-05	3.27E-01	trans
DOCK9	4.05	5.94E-05	3.28E-01	trans
UBL3	4.05	5.97E-05	3.28E-01	trans
TMEM123	4.04	6.25E-05	3.32E-01	trans
CD40	4.03	6.49E-05	3.36E-01	trans
PTPN2	4.01	7.00E-05	3.43E-01	trans
CD82	4.01	7.10E-05	3.44E-01	trans
CSK	4.00	7.28E-05	3.47E-01	trans
KLF3	4.00	7.48E-05	3.50E-01	trans
UBE3A	3.94	9.33E-05	3.69E-01	trans
ARHGEF1	3.93	9.81E-05	3.74E-01	trans
NT5C	3.93	1.00E-04	3.76E-01	trans
STRN	3.93	1.00E-04	3.76E-01	trans
HIVEP2	3.91	1.05E-04	3.80E-01	trans
STAP1	3.89	1.16E-04	3.90E-01	trans
MGAT5	3.88	1.23E-04	3.95E-01	trans
TSSC1	3.86	1.30E-04	4.02E-01	trans
LILRB2	3.83	1.44E-04	4.13E-01	trans

Gene	t Statistic	P-value	FDR	eQTL type
DNAJA2	3.82	1.52E-04	4.18E-01	trans
DHRS7	3.81	1.58E-04	4.21E-01	trans
LEMD3	3.79	1.69E-04	4.27E-01	trans
PRMT6	3.77	1.83E-04	4.35E-01	trans
RBM43	3.74	2.09E-04	4.47E-01	trans
TET3	3.74	2.11E-04	4.48E-01	trans
RAB28	3.73	2.18E-04	4.51E-01	trans
SH2D3A	3.73	2.18E-04	4.51E-01	trans
NXPH3	3.72	2.29E-04	4.56E-01	trans
HESX1	3.71	2.30E-04	4.57E-01	trans
CR2	3.71	2.31E-04	4.57E-01	trans
GADD45B	3.71	2.38E-04	4.60E-01	trans
JAK1	3.69	2.55E-04	4.67E-01	trans
DTNBP1	3.67	2.68E-04	4.72E-01	trans
AGGF1	3.66	2.86E-04	4.78E-01	trans
CHMP2A	3.65	2.96E-04	4.82E-01	trans
INO80C	3.64	3.03E-04	4.84E-01	trans
SLC46A3	3.63	3.13E-04	4.87E-01	trans
VANGL1	3.63	3.15E-04	4.88E-01	trans
COASY	3.62	3.25E-04	4.91E-01	trans
SFT2D2	3.61	3.35E-04	4.95E-01	trans
GPSM3	3.61	3.44E-04	4.97E-01	trans
KIF26B	3.60	3.55E-04	5.00E-01	trans
RCSD1	3.59	3.71E-04	5.04E-01	trans
MRPS18A	3.58	3.85E-04	5.08E-01	trans
IFT52	3.58	3.87E-04	5.08E-01	trans
SETBP1	3.56	4.12E-04	5.15E-01	trans
ASPHD2	3.54	4.37E-04	5.21E-01	trans
FRY	3.54	4.40E-04	5.22E-01	trans
PTPN22	3.54	4.49E-04	5.24E-01	trans
SH3RF1	3.53	4.51E-04	5.24E-01	trans
LNPEP	3.51	4.89E-04	5.33E-01	trans
FAM188B	3.50	5.08E-04	5.37E-01	trans
PSAP	3.50	5.20E-04	5.40E-01	trans
CD38	3.48	5.50E-04	5.45E-01	trans
RAB8B	3.48	5.54E-04	5.46E-01	trans
ATF6	3.48	5.57E-04	5.46E-01	trans
OXCT1	3.48	5.58E-04	5.47E-01	trans
XPR1	3.47	5.76E-04	5.50E-01	trans
DEK	3.46	5.90E-04	5.52E-01	trans
ZCCHC7	3.43	6.68E-04	5.65E-01	trans
CD19	3.42	6.88E-04	5.68E-01	trans

Gene	t Statistic	P-value	FDR	eQTL type
PHOSPHO1	3.41	7.03E-04	5.70E-01	trans
PLCL2	3.41	7.20E-04	5.72E-01	trans
P2RX5	3.39	7.64E-04	5.78E-01	trans
ASB7	3.39	7.68E-04	5.79E-01	trans
TNIP2	3.38	7.76E-04	5.80E-01	trans
RHOH	3.38	7.86E-04	5.82E-01	trans
SNAPC4	3.38	7.92E-04	5.82E-01	trans
MRPS23	3.38	8.01E-04	5.83E-01	trans
NAT14	3.37	8.06E-04	5.84E-01	trans
FCRL3	3.37	8.14E-04	5.85E-01	trans
T	3.37	8.15E-04	5.85E-01	trans
KIAA0226L	3.36	8.49E-04	5.89E-01	trans
PPP1R18	3.36	8.50E-04	5.89E-01	trans
KCNN1	3.36	8.58E-04	5.90E-01	trans
USP18	3.35	8.64E-04	5.91E-01	trans
ATP6V0E2	3.35	8.81E-04	5.93E-01	trans
AP1B1	3.34	8.96E-04	5.95E-01	trans
PLEKHF2	3.33	9.29E-04	5.99E-01	trans
ACTB	3.33	9.35E-04	5.99E-01	trans
BAD	3.29	1.07E-03	6.11E-01	trans
FAM96A	3.29	1.09E-03	6.13E-01	trans
PDCD2	3.29	1.09E-03	6.14E-01	trans
TMEM131	3.27	1.17E-03	6.20E-01	trans
DDX21	3.26	1.20E-03	6.23E-01	trans
FAM65B	3.26	1.22E-03	6.25E-01	trans
RALA	3.25	1.24E-03	6.26E-01	trans
ZNF837	3.25	1.26E-03	6.28E-01	trans
ILK	3.24	1.28E-03	6.30E-01	trans
RNF146	3.24	1.31E-03	6.32E-01	trans
MYBL1	3.23	1.33E-03	6.34E-01	trans
KIAA1147	3.22	1.36E-03	6.36E-01	trans
SLC35A4	3.22	1.37E-03	6.37E-01	trans
LAMP3	3.22	1.39E-03	6.38E-01	trans
SAMD10	3.21	1.42E-03	6.39E-01	trans
GEMIN7	3.20	1.47E-03	6.42E-01	trans
EBI3	3.20	1.48E-03	6.43E-01	trans
MYL12A	3.19	1.52E-03	6.46E-01	trans
LSM11	3.19	1.53E-03	6.47E-01	trans
TET2	3.18	1.56E-03	6.48E-01	trans
IPO8	3.17	1.61E-03	6.51E-01	trans
C19orf84	3.17	1.61E-03	6.51E-01	trans
BCL2A1	3.17	1.61E-03	6.52E-01	trans

Gene	t Statistic	P-value	FDR	eQTL type
CYCS	3.17	1.63E-03	6.52E-01	trans
ZNF608	3.16	1.66E-03	6.54E-01	trans
GUCY1B3	3.16	1.69E-03	6.55E-01	trans
ZNF783	3.15	1.73E-03	6.57E-01	trans
MRS2	3.15	1.76E-03	6.60E-01	trans
TMEM154	3.13	1.89E-03	6.66E-01	trans
MATN1	3.12	1.95E-03	6.69E-01	trans
TMEM230	3.12	1.96E-03	6.69E-01	trans
UBASH3A	3.11	2.02E-03	6.72E-01	trans
RGS3	3.10	2.03E-03	6.72E-01	trans
ZNF717	3.10	2.04E-03	6.73E-01	trans
DBNDD1	3.09	2.12E-03	6.76E-01	trans
PAN3	3.09	2.14E-03	6.77E-01	trans
ACTR3	3.09	2.15E-03	6.77E-01	trans
COPS3	3.09	2.15E-03	6.77E-01	trans
VPS13D	3.08	2.17E-03	6.78E-01	trans
BNIP1	3.08	2.21E-03	6.79E-01	trans
HEATR6	3.07	2.27E-03	6.82E-01	trans
MARCKS	3.07	2.31E-03	6.83E-01	trans
SLC37A1	3.06	2.33E-03	6.84E-01	trans
LATS2	3.06	2.34E-03	6.84E-01	trans
NUS1	3.06	2.36E-03	6.85E-01	trans
PIP5K1B	3.06	2.37E-03	6.85E-01	trans
MTPN	3.06	2.38E-03	6.86E-01	trans
CSNK1G2	3.03	2.62E-03	6.94E-01	trans
WDSUB1	3.03	2.62E-03	6.95E-01	trans
ADAP1	3.03	2.63E-03	6.95E-01	trans
MAPKAPK2	3.02	2.65E-03	6.95E-01	trans
ID3	3.01	2.77E-03	6.99E-01	trans
ATP5L	3.01	2.79E-03	6.99E-01	trans
VASP	3.00	2.81E-03	7.00E-01	trans
RBM38	3.00	2.83E-03	7.01E-01	trans
PPFIBP1	3.00	2.84E-03	7.01E-01	trans
FIG4	2.99	2.92E-03	7.04E-01	trans
ZNF621	2.99	2.92E-03	7.04E-01	trans
SCIMP	2.99	2.93E-03	7.04E-01	trans
RIN3	2.99	2.94E-03	7.04E-01	trans
EML3	2.99	2.94E-03	7.04E-01	trans
SMIM11	2.99	2.98E-03	7.05E-01	trans
GNG2	2.98	3.09E-03	7.07E-01	trans
KLF13	2.97	3.10E-03	7.08E-01	trans
SLC16A11	2.97	3.11E-03	7.08E-01	trans

Gene	t Statistic	P-value	FDR	eQTL type
TTC39B	2.97	3.15E-03	7.09E-01	trans
BTF3	2.96	3.22E-03	7.11E-01	trans
FBXW11	2.96	3.25E-03	7.11E-01	trans
LYRM9	2.96	3.26E-03	7.12E-01	trans
CMKLR1	2.96	3.27E-03	7.12E-01	trans
ZNF215	2.95	3.31E-03	7.13E-01	trans
TTC38	2.95	3.33E-03	7.14E-01	trans
ABHD17B	2.95	3.38E-03	7.15E-01	trans
TINF2	2.94	3.40E-03	7.16E-01	trans
OSBPL3	2.94	3.45E-03	7.17E-01	trans
LBH	2.94	3.49E-03	7.18E-01	trans
TRIP10	2.93	3.57E-03	7.19E-01	trans
DBT	2.92	3.73E-03	7.22E-01	trans
C20orf196	2.91	3.78E-03	7.23E-01	trans
C9orf9	2.91	3.79E-03	7.24E-01	trans
UBAC2	2.90	3.98E-03	7.28E-01	trans
ATP6V1E2	2.90	3.98E-03	7.28E-01	trans
MAP2K6	2.89	3.98E-03	7.28E-01	trans
POMP	2.89	3.99E-03	7.29E-01	trans
TWISTNB	2.89	4.10E-03	7.30E-01	trans
TTC24	2.88	4.13E-03	7.31E-01	trans
SLC10A7	2.88	4.14E-03	7.31E-01	trans
SC5D	2.88	4.22E-03	7.32E-01	trans
LYN	2.88	4.23E-03	7.33E-01	trans
FLYWCH2	2.87	4.29E-03	7.34E-01	trans
CNP	2.86	4.40E-03	7.36E-01	trans
CD83	2.86	4.40E-03	7.36E-01	trans
ALDH3A1	2.85	4.56E-03	7.39E-01	trans
BCL2L13	2.85	4.58E-03	7.39E-01	trans
IFITM3	2.84	4.72E-03	7.42E-01	trans
MTA2	2.84	4.77E-03	7.42E-01	trans
PHLDB3	2.83	4.80E-03	7.43E-01	trans
EIF1B	2.83	4.87E-03	7.44E-01	trans
VTI1A	2.83	4.89E-03	7.44E-01	trans
CSTF3	2.83	4.93E-03	7.45E-01	trans
RASD1	2.82	5.02E-03	7.46E-01	trans
PAK1IP1	2.82	5.02E-03	7.46E-01	trans
PARP12	2.82	5.07E-03	7.47E-01	trans
GRK5	2.81	5.14E-03	7.48E-01	trans
TIGIT	2.81	5.15E-03	7.48E-01	trans
FBXL14	2.81	5.16E-03	7.49E-01	trans
SNX3	2.81	5.16E-03	7.49E-01	trans

Gene	t Statistic	P-value	FDR	eQTL type
FASTKD2	2.81	5.20E-03	7.49E-01	trans
PIK3R5	2.80	5.29E-03	7.51E-01	trans
GHITM	2.80	5.34E-03	7.51E-01	trans
RFX3	2.80	5.36E-03	7.52E-01	trans
ST3GAL2	2.79	5.47E-03	7.53E-01	trans
ABHD3	2.79	5.47E-03	7.53E-01	trans
PSMD6	2.79	5.50E-03	7.53E-01	trans
CDKN2D	2.79	5.51E-03	7.53E-01	trans
CWC22	2.79	5.54E-03	7.54E-01	trans
MMD	2.78	5.61E-03	7.55E-01	trans
ACYP2	2.78	5.64E-03	7.55E-01	trans
MED13	2.78	5.66E-03	7.56E-01	trans
TAOK1	2.78	5.71E-03	7.56E-01	trans
SRP19	2.78	5.73E-03	7.57E-01	trans
DCXR	2.77	5.91E-03	7.59E-01	trans
PTPN12	2.77	5.92E-03	7.59E-01	trans
EFHD2	2.75	6.15E-03	7.62E-01	trans
CTDSPL2	2.75	6.18E-03	7.62E-01	trans
NEK7	2.75	6.29E-03	7.64E-01	trans
MAPK6	2.74	6.36E-03	7.65E-01	trans
KCNC3	2.74	6.43E-03	7.66E-01	trans
TAF10	2.74	6.44E-03	7.66E-01	trans
CHST2	2.74	6.46E-03	7.66E-01	trans
GRK4	2.73	6.64E-03	7.68E-01	trans
ANAPC13	2.73	6.65E-03	7.68E-01	trans
ILKAP	2.72	6.77E-03	7.70E-01	trans
PPP4R2	2.72	6.78E-03	7.70E-01	trans
STT3B	2.71	6.94E-03	7.72E-01	trans
MYO1G	2.71	6.96E-03	7.72E-01	trans
RCAN1	2.71	6.98E-03	7.72E-01	trans
DCK	2.71	7.03E-03	7.73E-01	trans
TBC1D12	2.71	7.06E-03	7.73E-01	trans
ELOVL6	2.70	7.10E-03	7.74E-01	trans
SRD5A1	2.70	7.25E-03	7.75E-01	trans
MFN1	2.70	7.28E-03	7.76E-01	trans
GIPR	2.69	7.32E-03	7.76E-01	trans
ETF1	2.69	7.40E-03	7.77E-01	trans
SPN	2.69	7.45E-03	7.77E-01	trans
SEPP1	2.68	7.54E-03	7.78E-01	trans
HECA	2.68	7.66E-03	7.79E-01	trans
CBX5	2.68	7.70E-03	7.80E-01	trans
NRSN2	2.67	7.83E-03	7.81E-01	trans

Gene	t Statistic	P-value	FDR	eQTL type
PRR14	2.67	7.84E-03	7.81E-01	trans
TIMM8B	2.67	7.88E-03	7.82E-01	trans
CCDC71	2.67	7.97E-03	7.82E-01	trans
MVD	2.66	8.01E-03	7.83E-01	trans
CDYL	2.65	8.27E-03	7.85E-01	trans
SLC50A1	2.65	8.40E-03	7.86E-01	trans
MYL12B	2.64	8.49E-03	7.87E-01	trans
C17orf62	2.64	8.49E-03	7.87E-01	trans
LDLR	2.64	8.53E-03	7.88E-01	trans
GPR63	2.63	8.80E-03	7.90E-01	trans
MCOLN2	2.63	8.86E-03	7.91E-01	trans
IFNGR1	2.63	8.93E-03	7.91E-01	trans
ALAD	2.62	8.99E-03	7.92E-01	trans
PNKP	2.62	9.00E-03	7.92E-01	trans
SNAPC2	2.62	9.04E-03	7.92E-01	trans
GTF3C1	2.62	9.05E-03	7.92E-01	trans
CRTC3	2.62	9.17E-03	7.93E-01	trans
PPP6R3	2.62	9.22E-03	7.94E-01	trans
MIOS	2.62	9.23E-03	7.94E-01	trans
MOB3C	2.60	9.58E-03	7.96E-01	trans
RSL24D1	2.60	9.64E-03	7.97E-01	trans
COX16	2.60	9.65E-03	7.97E-01	trans
MFSD3	2.60	9.68E-03	7.97E-01	trans
TBCA	2.59	9.89E-03	7.98E-01	trans
WDFY4	2.59	9.90E-03	7.98E-01	trans

APPENDIX C.

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

Table 8 Top LD results for polyTE for African population

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_64	rs4654899	Superior frontal gyrus grey matter volume	EIF4G3	0.46
ALU_umary_ALU_102	rs96067	Corneal structure	COL8A2	0.65
ALU_umary_ALU_153	rs17106184	Type 2 diabetes	FAF1	0.43
ALU_umary_ALU_216	rs10889569	C-reactive protein	LEPR	-0.59
ALU_umary_ALU_224	rs1417437	Orofacial clefts	LRRC7	0.50
ALU_umary_ALU_226	rs11809230	Cannabis use (initiation)	.	0.74
ALU_umary_ALU_250	rs10789369	Schizophrenia	KRT8P21,LRRIQ3	0.51
ALU_umary_ALU_267	rs1857353	MRI atrophy measures	SLC44A5	0.40
ALU_umary_ALU_335	rs17131547	Bone mineral density	TGFBR3	0.41
ALU_umary_ALU_368	rs6675668	Stearic acid (18:0) plasma levels	ALG14	0.45
L1_umary_LINE1_99	rs11578152	Menarche (age at onset)	DNAJA1P5,COL11A1	0.53
SVA_umary_SVA_45	rs7411387	Interferon alpha levels in systemic lupus erythematosus	CHIA	0.40
ALU_umary_ALU_542	rs6427528	Response to anti-TNF treatment in rheumatoid arthritis	CD84	0.40
SVA_umary_SVA_57	rs4657482	Testicular germ cell tumor	UCK2	0.41
ALU_umary_ALU_580	rs6687813	D-dimer levels	SLC19A2,F5	0.63
L1_umary_LINE1_164	rs6703865	Hippocampal atrophy	F5	0.58
ALU_umary_ALU_668	rs12720541	Epilepsy (generalized)	PLA2G4A	0.41
ALU_umary_ALU_676	rs10737562	Systemic lupus erythematosus	RNA5SP73,BRINP3	0.59
ALU_umary_ALU_689	rs10801047	Crohn's disease	HNRNPA1P46,RGS18	0.51
ALU_umary_ALU_699	rs6678275	Alzheimer's disease (late onset)	B3GALT2,RPL23AP22	0.53
ALU_umary_ALU_716	rs426736	Meningococcal disease	CFHR3	0.65
ALU_umary_ALU_717	rs426736	Meningococcal disease	CFHR3	0.88
ALU_umary_ALU_718	rs426736	Meningococcal disease	CFHR3	0.89
ALU_umary_ALU_842	rs12410462	Major depressive disorder	BTF3P9,TUBB8P10	0.64
ALU_umary_ALU_884	rs482329	Life threatening arrhythmia	LINC00184,LINC01132	-0.54
ALU_umary_ALU_897	rs2820037	Hypertension	RPL39P10,CHRM3	0.59
ALU_umary_ALU_903	rs476141	Diabetic retinopathy	LOC339529	0.45
ALU_umary_ALU_940	rs10802346	Fractional exhaled nitric oxide (childhood)	SMYD3	-0.46
ALU_umary_ALU_971	rs11123610	Response to inhaled corticosteroid treatment in asthma (percentage change of FEV1)	ALLC	-0.41

TE	GWAS hits	GWAS phenotype	GWAS gene	r
L1_umary_LINE1_266	rs17043947	Self-rated health	RNA5SP87,KLHL29	0.61
L1_umary_LINE1_267	rs2681019	Dialysis-related mortality	RNA5SP87,KLHL29	-0.53
ALU_umary_ALU_1078	rs7601155	Waist circumference	BRE	0.44
L1_umary_LINE1_286	rs6750486	Conduct disorder (symptom count)	SLC25A5P2,MIR548AD	0.43
ALU_umary_ALU_1156	rs4245791	LDL cholesterol	ABCG8	0.60
ALU_umary_ALU_1163	rs2341459	Height	CAMKMT	-0.58
ALU_umary_ALU_1184	rs12987465	Adverse response to chemotherapy (neutropenia/leucopenia) (etoposide)	MIR548BA,RPL7P13	-0.46
ALU_umary_ALU_1207	rs2163237	IgG glycosylation	PNPT1,EFEMP1	0.42
ALU_umary_ALU_1221	rs889956	Educational attainment	EIF2S2P7,VRK2	0.44
ALU_umary_ALU_1224	rs17552189	Cannabis dependence	LINC01122	0.45
ALU_umary_ALU_1231	rs17552189	Cannabis dependence	LINC01122	0.48
ALU_umary_ALU_1273	rs4141819	Endometriosis	ETAA1,C1D	0.71
ALU_umary_ALU_1275	rs2901879	Colorectal cancer (diet interaction)	MEIS1-AS2,DNMT3AP1	0.42
ALU_umary_ALU_1277	rs6759808	Obesity-related traits	PLEK,FBXO48	-0.45
ALU_umary_ALU_1278	rs10208940	Urate levels in lean individuals	.	-0.56
L1_umary_LINE1_337	rs13391552	Metabolic traits	ALMS1	0.43
ALU_umary_ALU_1324	rs2037723	Lung function (forced expiratory volume in 1 second to forced vital capacity ratio)	SNAR-H,CYCSP6	0.67
L1_umary_LINE1_366	rs4321386	Hormone measurements	IL1R2	0.42
ALU_umary_ALU_1406	rs2163349	Addiction	NCK2	-0.71
ALU_umary_ALU_1517	rs17015535	Coronary artery calcification	WDR33	0.56
ALU_umary_ALU_1557	rs13405020	Non-small cell lung cancer	THSD7B	0.64
ALU_umary_ALU_1585	rs17515225	Motion sickness	LRP1B	0.58
ALU_umary_ALU_1621	rs7584099	Response to statin therapy	PABPC1P2,RPL26P14	0.58
ALU_umary_ALU_1623	rs2307394	Urate levels	ORC4	0.70
ALU_umary_ALU_1693	rs10192369	Amyotrophic lateral sclerosis	MIR4785,TANK	0.43
ALU_umary_ALU_1736	rs2102808	Parkinson's disease	PHF5GP,CERS6	0.51
ALU_umary_ALU_1740	rs16856332	Liver enzyme levels (alkaline phosphatase)	ABCB11	0.41
ALU_umary_ALU_1756	rs836589	Erectile dysfunction in type 1 diabetes	PDK1,RAPGEF4-AS1	-0.57
ALU_umary_ALU_1778	rs9287989	Periodontal microbiota	EXTL2P1,KIAA1715	-0.60
ALU_umary_ALU_1802	rs16867321	Obesity	CWC22,SCHLAP1	0.68
ALU_umary_ALU_1824	rs11678036	IgG glycosylation	RPL23AP33,ELF2P4	-0.60
ALU_umary_ALU_1843	rs2675399	Obesity-related traits	DIRC1,COL3A1	0.65
ALU_umary_ALU_1869	rs801350	Response to anti-retroviral therapy (ddI/d4T) in HIV-1 infection (Grade 3 peripheral neuropathy)	HNRNPA1P47,AHCYP5	0.68
ALU_umary_ALU_1871	rs801350	Response to anti-retroviral therapy (ddI/d4T) in HIV-1 infection (Grade 3 peripheral neuropathy)	HNRNPA1P47,AHCYP5	0.49
ALU_umary_ALU_1894	rs6434928	Schizophrenia	SF3B1,COQ10B	0.48
L1_umary_LINE1_488	rs988583	Neutrophil count	PLCL1	0.60
ALU_umary_ALU_1903	rs12471454	Insomnia	PLCL1,SATB2	0.62
ALU_umary_ALU_1943	rs13383928	Lung cancer-asbestos exposure interaction	LOC101927960	0.77

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_1963	rs4673659	Asthma (sex interaction)	ERBB4	0.42
ALU_umary_ALU_1971	rs9789347	Obesity-related traits	SPAG16	0.64
ALU_umary_ALU_1987	rs16857609	Breast cancer	DIRC3	0.49
ALU_umary_ALU_2009	rs10170846	Schizophrenia (inflammation and infection response interaction)	.	0.52
ALU_umary_ALU_2069	rs2292873	Obesity-related traits	RAB17	0.47
ALU_umary_ALU_2101	rs7652782	Serum uric acid levels	CNTN4	0.62
ALU_umary_ALU_2109	rs2587949	Periodontitis (DPAL)	SUMF1	-0.72
ALU_umary_ALU_2114	rs1810320	Breast cancer	GRM7	0.43
SVA_umary_SVA_157	rs3729931	Cardiac hypertrophy	RAF1	0.56
ALU_umary_ALU_2164	rs1318937	Alcohol dependence	SH3BP5,SH3BP5-AS1	-0.66
ALU_umary_ALU_2289	rs13072940	Autism spectrum disorder attention deficit-hyperactivity disorder bipolar disorder major depressive disorder and schizophrenia (combined)	HSPD1P6,TRANK1	-0.60
L1_umary_LINE1_580	rs3922844	PR interval	SCN5A	0.82
ALU_umary_ALU_2333	rs319690	Blood pressure	MAP4	0.77
SVA_umary_SVA_170	rs9876781	Longevity	ATRIP	0.41
ALU_umary_ALU_2340	rs11719291	Cognitive function	IP6K2	0.47
ALU_umary_ALU_2346	rs7613875	Body mass index	.	0.48
ALU_umary_ALU_2349	rs11130248	Keloid	.	0.50
ALU_umary_ALU_2371	rs6764184	Optic cup area	.	0.63
L1_umary_LINE1_629	rs17518584	Cognitive function	CADM2	0.60
ALU_umary_ALU_2504	rs9883474	Brain connectivity	KRT8P25,APOOP2	-0.48
ALU_umary_ALU_2577	rs10511217	Economic and political preferences (environmentalism)	MIR548AB,RAP1BP2	0.66
ALU_umary_ALU_2578	rs2677247	IgG glycosylation	MIR548AB,RAP1BP2	0.70
ALU_umary_ALU_2616	rs1881681	Current cigarettes per day in onic obstructive pulmonary disease	PVRL3,CD96	0.41
ALU_umary_ALU_2627	rs13092825	Dental caries	.	0.46
ALU_umary_ALU_2637	rs9841504	Gastric cancer	LOC102723469,ZBTB20	0.57
ALU_umary_ALU_2654	rs6804441	Systemic lupus erythematosus	CD80	0.42
ALU_umary_ALU_2693	rs13075436	Response to angiotensin II receptor blocker therapy	C3orf56,TPRA1	0.42
ALU_umary_ALU_2696	rs2687729	Menarche (age at onset)	EEFSEC	0.50
ALU_umary_ALU_2697	rs2712381	Monocyte count	RPN1	-0.66
ALU_umary_ALU_2698	rs1534166	Alcohol consumption (transferrin glycosylation)	SRPRB	0.42
ALU_umary_ALU_2757	rs3773506	Type 2 diabetes	PLS1	0.78
ALU_umary_ALU_2795	rs13072552	Serum ceruloplasmin levels	CP	0.44
ALU_umary_ALU_2810	rs1351267	Schizophrenia	SUCNR1,MBNL1	-0.42
ALU_umary_ALU_2838	rs12638253	Multiple sclerosis (severity)	LEKR1	0.41
ALU_umary_ALU_2842	rs13064954	Smoking cessation in onic obstructive pulmonary disease	LINC00881,CCNL1	0.71
L1_umary_LINE1_723	rs2362965	Height	RSRC1	0.43
ALU_umary_ALU_2847	rs2362965	Height	RSRC1	0.49
ALU_umary_ALU_2911	rs2201862	Myeloproliferative neoplasms	.	0.42

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_2929	rs3913363	Response to angiotensin II receptor blocker therapy	TMEM212,FNDC3B	0.44
L1_umary_LINE1_761	rs9898	Activated partial thromboplastin time	HRG	0.46
ALU_umary_ALU_3022	rs10937470	Total ventricular volume	UTS2B	0.83
ALU_umary_ALU_3079	rs4619890	Glaucoma (primary open-angle)	AFAP1	0.46
ALU_umary_ALU_3085	rs13142500	Rheumatoid arthritis	CLNK,RNPS1P1	0.66
ALU_umary_ALU_3178	rs7442317	Attention deficit hyperactivity disorder motor coordination	EEF1A1P21,RPS3AP17	-0.65
ALU_umary_ALU_3235	rs35141484	Asthma (childhood onset)	KLHL5	0.56
ALU_umary_ALU_3236	rs11096990	Cognitive function	WDR19	0.42
ALU_umary_ALU_3244	rs10938397	Body mass index	PRDX4P1,PRKRIRP9	0.41
ALU_umary_ALU_3246	rs114070671	Opioid sensitivity	APBB2	0.45
L1_umary_LINE1_831	rs2055942	Type 2 diabetes	GABRA4	-0.54
ALU_umary_ALU_3274	rs13106975	Sphingolipid levels	ATP10D	0.42
ALU_umary_ALU_3295	rs13113518	Height	CLOCK	0.64
ALU_umary_ALU_3340	rs7656244	Kawasaki disease	TECRL	-0.40
ALU_umary_ALU_3393	rs7041	Serum vitamin D-binding protein levels	GC	0.40
ALU_umary_ALU_3402	rs1894292	Prostate cancer	AFM	-0.58
ALU_umary_ALU_3412	rs2273	Longevity	SDAD1	-0.42
ALU_umary_ALU_3430	rs1268789	Hair morphology	FRAS1	0.62
ALU_umary_ALU_3479	rs6834314	Liver enzyme levels (alanine transaminase)	GAPDHP60,MIR5705	0.60
ALU_umary_ALU_3594	rs10033900	Age-related macular degeneration	PLA2G12A,CFI	0.43
SVA_umary_SVA_222	rs4698790	Fasting insulin (interaction)	CFI,GAR1	0.40
L1_umary_LINE1_928	rs1585471	Myopia (pathological)	RPL36AP23,CCDC34P1	0.89
ALU_umary_ALU_3601	rs10034228	Myopia (pathological)	RPL36AP23,CCDC34P1	0.70
ALU_umary_ALU_3688	rs724950	Obesity-related traits	RBM48P1,INTU	0.43
ALU_umary_ALU_3796	rs1395821	Coronary heart disease	TTC29,MIR548G	0.71
ALU_umary_ALU_3957	rs2333163	Obesity-related traits	ADAM29,TSEN2P1	0.65
ALU_umary_ALU_3997	rs2130392	Kawasaki disease	CENPU	-0.41
ALU_umary_ALU_4046	rs16875288	Functional impairment in major depressive disorder bipolar disorder and schizophrenia	ADAMTS16	0.49
ALU_umary_ALU_4079	rs2607292	Body mass index	6-Mar	0.48
L1_umary_LINE1_1051	rs20476	PR interval in Tripanosoma cruzi seropositivity	CTNND2	0.49
ALU_umary_ALU_4117	rs4866334	IgG glycosylation	RPL36AP21,RPL32P14	0.52
L1_umary_LINE1_1097	rs10053502	Myopia (pathological)	INTS6P1,LINC00603	0.81
ALU_umary_ALU_4266	rs9291768	Classic bladder exstrophy	.	-0.50
ALU_umary_ALU_4267	rs4865673	Dental caries	HMGB1P47,KATNBL1P4	0.43
ALU_umary_ALU_4283	rs7716219	Height	SLC38A9	-0.46
ALU_umary_ALU_4289	rs16884711	IgG glycosylation	FLJ31104,ANKRD55	0.56
ALU_umary_ALU_4310	rs6859219	Rheumatoid arthritis	ANKRD55	0.40
ALU_umary_ALU_4329	rs1494630	Age-related hearing impairment	HTR1A,RNF180	0.50
ALU_umary_ALU_4333	rs1494630	Age-related hearing impairment	HTR1A,RNF180	0.48

TE	GWAS hits	GWAS phenotype	GWAS gene	r
SVA_umary_SVA_245	rs7729539	QT interval	CWC27,ADAMTS6	0.68
ALU_umary_ALU_4367	rs10515148	Hip geometry	YBX1P5,ZNF366	0.43
ALU_umary_ALU_4449	rs6870983	Body mass index	.	-0.66
ALU_umary_ALU_4535	rs1829883	Hemostatic factors and hematological phenotypes	RPS9P3,GUSBP8	0.43
ALU_umary_ALU_4563	rs112724034	Alzheimer's disease (cognitive decline)	LOC100289673	0.45
ALU_umary_ALU_4615	rs2376682	Diisocyanate-induced asthma	.	-0.41
ALU_umary_ALU_4650	rs1910003	Antibody status in Tripanosoma cruzi seropositivity	RPSAP37,GRAMD3	0.42
ALU_umary_ALU_4684	rs7735563	Diisocyanate-induced asthma	.	0.60
ALU_umary_ALU_4773	rs727809	Age-related hearing impairment (interaction)	TRNAC32P,GRIA1	0.62
ALU_umary_ALU_4807	rs2853694	Psoriasis	.	0.66
ALU_umary_ALU_4976	rs204247	Breast cancer	RANBP9,MCUR1	-0.67
ALU_umary_ALU_5017	rs4712652	Obesity	.	0.54
ALU_umary_ALU_5040	rs3129055	Nasopharyngeal carcinoma	TRNAI25	0.41
ALU_umary_ALU_5044	rs172166	Cardiac Troponin-T levels	TRNAI25	0.47
ALU_umary_ALU_5053	rs2523822	Drug-induced liver injury (amoxicillin-clavulanate)	TRNAI25	0.92
ALU_umary_ALU_5054	rs1061235	Adverse response to carbamapezine	HLA-A	0.79
ALU_umary_ALU_5055	rs259919	HIV-1 control	ZNRD1-AS1	0.58
ALU_umary_ALU_5056	rs6935053	Ulcerative colitis	TRNAI25	0.58
ALU_umary_ALU_5060	rs12175489	Visceral adipose tissue adjusted for BMI	MICA	0.46
SVA_umary_SVA_278	rs9368677	Atopic dermatitis	TRNAI25	0.82
ALU_umary_ALU_5064	rs1055569	Psychotic symptoms and prion disease	.	-0.64
ALU_umary_ALU_5075	rs7775228	Asthma	TRNAI25	0.65
ALU_umary_ALU_5076	rs2859113	IgG glycosylation	TRNAI25	0.62
ALU_umary_ALU_5077	rs7756516	Chronic hepatitis B infection	HLA-DQB2	0.56
ALU_umary_ALU_5079	rs2621416	Lymphoma	TRNAI25	0.80
SVA_umary_SVA_282	rs3077	Hepatitis B (viral clearance)	HLA-DPA1	0.65
ALU_umary_ALU_5127	rs10948222	Height	SUPT3H	0.57
ALU_umary_ALU_5132	rs10948222	Height	SUPT3H	0.81
L1_umary_LINE1_1293	rs9357506	Body mass index	.	0.42
L1_umary_LINE1_1317	rs9342616	QT interval	NUFIP1P,RNA5SP208	0.45
ALU_umary_ALU_5237	rs9354654	Classic bladder exstrophy	.	-0.47
ALU_umary_ALU_5248	rs9346353	Sleep duration	LMBRD1	0.43
ALU_umary_ALU_5278	rs9447004	Calcium levels	CD109	0.45
ALU_umary_ALU_5280	rs9447004	Calcium levels	CD109	0.72
ALU_umary_ALU_5298	rs7738636	Acute lymphoblastic leukemia (childhood)	IMPG1,HTR1B	-0.40
ALU_umary_ALU_5356	rs366676	Echocardiographic traits	AKIRIN2,SPACA1	-0.44
ALU_umary_ALU_5380	rs2506933	Cognitive performance	ATF1P1,COPS5P1	0.53
ALU_umary_ALU_5395	rs11757063	Migraine	FUT9,UFL1	0.82
ALU_umary_ALU_5415	rs4840097	Age-related macular degeneration (smoking status interaction)	PRDM13,MCHR2	0.40
ALU_umary_ALU_5429	rs484621	Glucose homeostasis traits	ATG5	0.41

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_5475	rs33980500	Psoriatic arthritis	TRAF3IP2,TRAF3IP2-AS1	0.42
ALU_umary_ALU_5495	rs9488822	Cholesterol total	FRK	0.42
ALU_umary_ALU_5509	rs89107	Cardiac structure and function	SLC35F1	-0.51
SVA_umary_SVA_305	rs12110693	Biomedical quantitative traits	RPL23AP48,HMGB3P18	-0.56
ALU_umary_ALU_5565	rs13209747	Blood pressure	RPS4XP9,RSPO3	0.44
ALU_umary_ALU_5575	rs6938574	Menarche (age at onset)	LOC101928140,PTPRK	0.42
ALU_umary_ALU_5647	rs225675	Thiazide-induced adverse metabolic effects in hypertensive patients	VTA1	0.40
L1_umary_LINE1_1418	rs225675	Thiazide-induced adverse metabolic effects in hypertensive patients	VTA1	0.84
SVA_umary_SVA_315	rs1933488	Prostate cancer	RGS17	-0.58
ALU_umary_ALU_5708	rs2275336	Parkinson's disease	CNKSR3	0.47
ALU_umary_ALU_5713	rs4269383	Pancreatic cancer	LOC101928923	0.43
SVA_umary_SVA_318	rs2451258	Rheumatoid arthritis	TAGAP,FNDC1	0.42
ALU_umary_ALU_5769	rs7762018	Thiazide-induced adverse metabolic effects in hypertensive patients	PHF10	0.60
ALU_umary_ALU_5882	rs12670798	Cholesterol total	DNAH11	0.62
ALU_umary_ALU_5886	rs2286503	Fibrinogen	TOMM7	-0.56
ALU_umary_ALU_5916	rs10486483	Crohn's disease	SKAP2	0.61
L1_umary_LINE1_1474	rs10486483	Crohn's disease	SKAP2	0.49
ALU_umary_ALU_5957	rs9648428	Obesity-related traits	EEPD1	0.53
ALU_umary_ALU_5967	rs2392510	Periodontitis	GPR141	-0.40
ALU_umary_ALU_5970	rs4723738	Treatment response for severe sepsis	STARD3NL	0.65
ALU_umary_ALU_6007	rs1722133	Sitting height ratio	.	0.44
ALU_umary_ALU_6013	rs10279826	Urate levels in obese individuals	.	0.42
ALU_umary_ALU_6059	rs7786410	Age-related hearing impairment	14-Sep	0.44
ALU_umary_ALU_6100	rs7794356	Response to montelukast in asthma (change in FEV1)	.	0.51
ALU_umary_ALU_6200	rs1133906	Systemic lupus erythematosus and Systemic sclerosis	SAMD9L	-0.42
L1_umary_LINE1_1577	rs10953730	Metabolite levels	LINC00998,PPP1R3A	0.68
ALU_umary_ALU_6320	rs41997	Response to platinum-based chemotherapy in non-small-cell lung cancer	ANKRD7,GTF3AP6	0.40
ALU_umary_ALU_6359	rs4731207	Cutaneous malignant melanoma	.	0.48
ALU_umary_ALU_6369	rs7458938	Response to efavirenz-containing treatment in HIV 1 infection (virologic failure)	RPL31P39,GRM8	0.53
ALU_umary_ALU_6372	rs2687481	Hearing function	RPL31P39,GRM8	0.73
ALU_umary_ALU_6373	rs2687481	Hearing function	RPL31P39,GRM8	0.70
ALU_umary_ALU_6375	rs17864092	Depression (quantitative trait)	GRM8	0.42
ALU_umary_ALU_6427	rs10250997	Autism spectrum disorder attention deficit-hyperactivity disorder bipolar disorder major depressive disorder and schizophrenia (combined)	MTPN,PSMC1P3	0.51
ALU_umary_ALU_6521	rs10274279	Myopia (pathological)	PTPRN2	0.58
ALU_umary_ALU_6531	rs11986414	Gaucher disease severity	CLN8,MIR3674	0.41

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_6563	rs6601327	Multiple myeloma (hyperdiploidy)	PPP1R3B,TNKS	-0.55
ALU_umary_ALU_6564	rs6601327	Multiple myeloma (hyperdiploidy)	PPP1R3B,TNKS	0.51
ALU_umary_ALU_6565	rs6601327	Multiple myeloma (hyperdiploidy)	PPP1R3B,TNKS	-0.45
ALU_umary_ALU_6611	rs920590	Acute lymphoblastic leukemia (childhood)	CSGALNACT1,INTS10	0.45
SVA_umary_SVA_369	rs7843479	Mean corpuscular volume	XPO7	0.41
ALU_umary_ALU_6662	rs2439312	Dialysis-related mortality	NRG1	-0.48
ALU_umary_ALU_6669	rs11997175	Body mass index	.	0.62
ALU_umary_ALU_6673	rs6990255	Autism spectrum disorder attention deficit-hyperactivity disorder bipolar disorder major depressive disorder and schizophrenia (combined)	CYCSP3,RPL10AP3	0.47
ALU_umary_ALU_6674	rs6987004	Pulmonary function decline	RPL10AP3,RPL21P80	0.78
ALU_umary_ALU_6695	rs7829127	Refractive error	ZMAT4	0.51
ALU_umary_ALU_6769	rs9650199	Response to amphetamines	PDCL3P1,RAB2A	-0.45
ALU_umary_ALU_6806	rs7017914	Bone mineral density	XKR9	0.62
ALU_umary_ALU_6820	rs11994034	Attention deficit hyperactivity disorder (combined symptoms)	TRPA1,RNA5SP271	0.54
ALU_umary_ALU_6893	rs7015622	Response to anti-depressant treatment in major depressive disorder	HNRNPA1P4,RALYL	0.81
ALU_umary_ALU_6951	rs278567	Bipolar disorder and schizophrenia	C8orf87	0.76
ALU_umary_ALU_6990	rs2033562	IgA nephropathy	.	0.52
ALU_umary_ALU_7003	rs284489	Glaucoma (primary open-angle)	LRP12,RPL23P9	-0.44
ALU_umary_ALU_7007	rs12541635	Age of smoking initiation	RPL12P24,SLC16A14P1	-0.50
SVA_umary_SVA_389	rs374810	Ossification of the posterior longitudinal ligament of the spine	RSPO2	0.68
ALU_umary_ALU_7037	rs7832552	Body mass (lean)	TRHR	0.60
ALU_umary_ALU_7106	rs3870371	Periodontal disease-related phenotypes	HAS2-AS1,MRPS36P3	0.53
ALU_umary_ALU_7140	rs7830341	Body mass index	.	0.57
ALU_umary_ALU_7143	rs13281615	Breast cancer	LOC101930033	0.47
ALU_umary_ALU_7218	rs10962181	Superior frontal gyrus grey matter volume	RNA5SP279,SMARCA2	-0.43
L1_umary_LINE1_1778	rs10814916	Type 2 diabetes	GLIS3	0.50
L1_umary_LINE1_1783	rs16924631	Periodontal microbiota	UHRF2	-0.42
ALU_umary_ALU_7256	rs4742269	Radiation response	KDM4C	0.42
ALU_umary_ALU_7311	rs3904778	Adolescent idiopathic scoliosis	.	-0.45
ALU_umary_ALU_7331	rs7867456	Axial length	HACD4,IFNNP1,PTPLAD2	-0.48
ALU_umary_ALU_7339	rs10738626	Atopic dermatitis	UBA52P6,DMRTA1	-0.58
ALU_umary_ALU_7381	rs10969853	Alcohol dependence (age at onset)	RBMXP2,KRT18P66	0.57
SVA_umary_SVA_401	rs10758189	IgG glycosylation	B4GALT1	-0.63
L1_umary_LINE1_1831	rs11145465	Refractive error	.	0.41
ALU_umary_ALU_7499	rs79460104	Response to amphetamines	.	0.50
L1_umary_LINE1_1854	rs12554999	Plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid)	CHCHD2P9,TLE4	0.87
ALU_umary_ALU_7555	rs883924	Hepatitis C induced liver fibrosis	LINC01508,LOC101927873	0.45
L1_umary_LINE1_1867	rs10122541	Thyroid cancer	.	0.57

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_7593	rs7028939	Preeclampsia	ERP44	-0.71
ALU_umary_ALU_7594	rs7028939	Preeclampsia	ERP44	-0.77
ALU_umary_ALU_7615	rs7848024	Tourette's syndrome or obsessive-compulsive disorder	ZYG11AP1,CYLC2	0.53
ALU_umary_ALU_7658	rs10980800	Monocyte count	RNY4P18,MIR7702	0.41
ALU_umary_ALU_7756	rs7086888	Multiple myeloma (IgH translocation)	LINC00703,LINC00704	0.41
ALU_umary_ALU_7783	rs11256676	Dental caries	CUX2P1,SFTA1P	0.48
ALU_umary_ALU_7866	rs927675	MRI atrophy measures	MPP7	0.48
ALU_umary_ALU_8021	rs442309	Vogt-Koyanagi-Harada syndrome	ZNF365,ALDH7A1P4	0.62
ALU_umary_ALU_8046	rs2441727	Interstitial lung disease	CTNNA3	0.42
ALU_umary_ALU_8048	rs2441727	Interstitial lung disease	CTNNA3	0.43
ALU_umary_ALU_8056	rs16926246	Hemoglobin	HK1	0.43
SVA_umary_SVA_449	rs58180147	Parasitemia in Tripanosoma cruzi seropositivity	MYPN	0.42
ALU_umary_ALU_8112	rs1414874	Self-reported allergy	MARK2P15,HMGN2P8	0.60
ALU_umary_ALU_8178	rs1934951	Osteonecrosis of the jaw	CYP2C8	-0.41
ALU_umary_ALU_8182	rs56322409	Blood metabolite levels	ALDH18A1	0.73
ALU_umary_ALU_8232	rs11195062	Multiple myeloma	MXI1	0.52
ALU_umary_ALU_8297	rs7069346	Migraine without aura	C10orf88,PSTK	-0.59
ALU_umary_ALU_8344	rs2213169	Hematology traits	LCRB	0.43
ALU_umary_ALU_8398	rs1330	Height	NUCB2	0.40
L1_umary_LINE1_2073	rs12788764	Age-related nuclear cataracts	LUZP2,RPL36AP40	0.52
ALU_umary_ALU_8456	rs10834691	IgG glycosylation	LUZP2,RPL36AP40	0.73
ALU_umary_ALU_8509	rs1355223	Systemic lupus erythematosus and Systemic sclerosis	LOC102723568	0.41
ALU_umary_ALU_8532	rs7951105	Free thyroxine concentration	RPL7AP56,RPL18P8	-0.44
L1_umary_LINE1_2104	rs1484948	RR interval (heart rate)	RPL9P23,HNRNPKP3	0.79
ALU_umary_ALU_8559	rs2176598	Body mass index	.	0.71
ALU_umary_ALU_8572	rs1351696	D-dimer levels	OR4C4P,OR4C5	0.58
ALU_umary_ALU_8574	rs1814175	Height	CBX3P8,TRIM51FP	0.81
ALU_umary_ALU_8590	rs11228719	Orofacial clefts	OR2AH1P,OR9G1	0.48
ALU_umary_ALU_8620	rs478304	Acne (severe)	RNASEH2C,KRT8P26	-0.73
ALU_umary_ALU_8692	rs17148090	Phospholipid levels (plasma)	DLG2	0.47
ALU_umary_ALU_8712	rs1386330	Multiple sclerosis (age of onset)	HMGB3P25,RAB38	0.58
ALU_umary_ALU_8716	rs2658782	Pulmonary function decline	CCDC67	0.42
ALU_umary_ALU_8717	rs10830228	Age-related macular degeneration	RNU6-16P,TYR	0.68
ALU_umary_ALU_8799	rs7947821	Tuberculosis	.	0.88
ALU_umary_ALU_8801	rs313426	Toenail selenium levels	DYNC2H1	-0.44
L1_umary_LINE1_2187	rs326946	Alzheimer's disease (cognitive decline)	ARHGAP20	0.46
L1_umary_LINE1_2191	rs2250417	Protein quantitative trait loci	BCO2	0.85
ALU_umary_ALU_8954	rs2007044	Schizophrenia	.	0.43
ALU_umary_ALU_8998	rs1031391	Bitter taste perception	PRH1-PRR4	0.46
ALU_umary_ALU_9001	rs2908835	Information processing speed	HIGD1AP8,IQSEC3P2	0.52

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_9108	rs1979679	Ossification of the posterior longitudinal ligament of the spine	CCDC91	0.81
ALU_umary_ALU_9143	rs11170468	Body mass index	.	0.48
ALU_umary_ALU_9156	rs11170468	Body mass index	.	0.41
ALU_umary_ALU_9158	rs76904798	Parkinson's disease	RPL30P13,LRRK2	0.54
ALU_umary_ALU_9177	rs275380	Adverse response to lamotrigine and phenytoin	ADAMTS20	0.52
ALU_umary_ALU_9214	rs59448276	Diisocyanate-induced asthma	.	0.51
ALU_umary_ALU_9233	rs1633360	Rheumatoid arthritis	OS9	-0.41
ALU_umary_ALU_9243	rs17121944	Temperament (bipolar disorder)	RPS6P22,METTL15P2	0.49
ALU_umary_ALU_9286	rs6581612	Hippocampal volume	APOOP3,LEMD3	0.54
ALU_umary_ALU_9348	rs2669010	Systemic lupus erythematosus	RPL7AP9,YWHAQP7	0.46
ALU_umary_ALU_9371	rs10778699	Thiazide-induced adverse metabolic effects in hypertensive patients	RPL7P38,RPL26P32	-0.47
ALU_umary_ALU_9388	rs1511589	Optic cup area	.	0.72
ALU_umary_ALU_9431	rs11106179	Expressive vocabulary in infants	DCN,C12orf79	0.45
ALU_umary_ALU_9450	rs7953959	IgG glycosylation	TRNAQ46P,RMST,TRQ-TTG9-1	0.65
ALU_umary_ALU_9636	rs509915	Urate levels (BMI interaction)	.	0.45
SVA_umary_SVA_566	rs7328278	Asthma (childhood onset)	DCLK1	0.44
L1_umary_LINE1_2401	rs4142110	Nephrolithiasis	DGKH	0.80
ALU_umary_ALU_9813	rs9537938	Educational attainment	RNA5SP30,CTAGE16P	0.61
ALU_umary_ALU_9857	rs1340490	Response to platinum-based chemotherapy (cisplatin)	RPL32P28,LINC00395	0.84
L1_umary_LINE1_2432	rs1340490	Response to platinum-based chemotherapy (cisplatin)	RPL32P28,LINC00395	0.58
ALU_umary_ALU_9861	rs11148643	Rheumatoid arthritis	NFYAP1,LGMNP1	-0.45
ALU_umary_ALU_9863	rs9540294	Recalcitrant atopic dermatitis	.	0.74
ALU_umary_ALU_9886	rs2991396	Male fertility	RPSAP53,NPM1P22	0.45
ALU_umary_ALU_9941	rs975739	Hair color	MIR3665,EDNRB-AS1	-0.41
ALU_umary_ALU_9995	rs4773460	Hippocampal atrophy	DDX6P2,TXNL1P1	-0.62
ALU_umary_ALU_9998	rs9589866	Diisocyanate-induced asthma	.	-0.51
ALU_umary_ALU_10028	rs4771859	Adverse response to chemotherapy (neutropenia/leucopenia) (all antimicrotubule drugs)	GPC5	0.46
ALU_umary_ALU_10095	rs1509091	Metabolite levels (Pyroglutamine)	FAM155A	0.46
ALU_umary_ALU_10151	rs1950500	Height	RIPK3,NFATC4	0.45
ALU_umary_ALU_10167	rs7493138	Longevity	RPL26P3,EIF4A1P12	0.41
ALU_umary_ALU_10255	rs2488856	Osteoprotegerin levels	YWHAQP1,TUBBP3	0.58
ALU_umary_ALU_10284	rs7159841	Hemostatic factors and hematological phenotypes	MDGA2	-0.45
ALU_umary_ALU_10285	rs7159841	Hemostatic factors and hematological phenotypes	MDGA2	0.45
ALU_umary_ALU_10299	rs1959536	Psychosis (atypical)	TRIM9	0.48
ALU_umary_ALU_10334	rs11851015	Alcohol consumption	EXOC5	0.56
ALU_umary_ALU_10353	rs7153648	Prostate cancer	SIX1,SIX4	0.46
ALU_umary_ALU_10368	rs10498514	Cognitive performance	MTHFD1	-0.49
ALU_umary_ALU_10374	rs8020095	Depression (quantitative trait)	GPHN	0.64
L1_umary_LINE1_2583	rs8017304	Age-related macular degeneration	RAD51B	0.57

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_10396	rs55951657	Hippocampal volume	ACOT4,ACOT6	0.44
ALU_umary_ALU_10401	rs910316	Height	TMED10	0.67
SVA_umary_SVA_613	rs6574644	Obesity-related traits	STON2	-0.53
ALU_umary_ALU_10482	rs17124955	Obesity-related traits	TTC8,TRA-AGC15-1,TRNAA17	0.53
ALU_umary_ALU_10518	rs7140601	Immune response to smallpox vaccine (IL-6)	C14orf64,C14orf177,LINC01550	0.81
ALU_umary_ALU_10580	rs61996546	Response to methotrexate in juvenile idiopathic arthritis	GABRB3	0.45
L1_umary_LINE1_2647	rs2414095	Follicule stimulating hormone	CYP19A1	0.47
SVA_umary_SVA_632	rs2553218	Immune response to smallpox vaccine (IL-6)	UNC13C	0.55
ALU_umary_ALU_10715	rs491567	Chronic kidney disease	WDR72	-0.50
L1_umary_LINE1_2651	rs491567	Chronic kidney disease	WDR72	-0.41
ALU_umary_ALU_10766	rs1436958	IgG glycosylation	VPS13C	0.60
SVA_umary_SVA_640	rs1549318	Proinsulin levels	RPL29P30,LARP6	0.41
ALU_umary_ALU_10812	rs8038465	Liver enzyme levels (gamma-glutamyl transferase)	CD276	0.74
ALU_umary_ALU_10819	rs3099143	Recalcitrant atopic dermatitis	.	0.60
ALU_umary_ALU_10885	rs7495052	Inattentive symptoms	SLCO3A1	0.40
ALU_umary_ALU_11100	rs8047014	Attention deficit hyperactivity disorder	RPS2P45,HAS3	0.53
ALU_umary_ALU_11101	rs12149862	Blood pressure (smoking interaction)	CYB5B	0.46
ALU_umary_ALU_11136	rs8050187	Anorexia nervosa	WWOX	0.48
ALU_umary_ALU_11154	rs12933472	Glucose homeostasis traits	CDH13	0.52
SVA_umary_SVA_683	rs781856	Glucose homeostasis traits	ZZEF1	0.72
ALU_umary_ALU_11196	rs73976923	Diisocyanate-induced asthma	.	0.59
ALU_umary_ALU_11238	rs7211756	Blood pressure (smoking interaction)	ZSWIM7	0.64
ALU_umary_ALU_11275	rs225212	Hypertension risk in short sleep duration	MYO1D	0.69
ALU_umary_ALU_11276	rs379123	Local histogram emphysema pattern	MYO1D	0.50
ALU_umary_ALU_11327	rs7207400	Alzheimer's disease in APOE e4-carriers	.	-0.54
ALU_umary_ALU_11330	rs2935183	Multiple sclerosis or amyotrophic lateral sclerosis	NPEPPS	-0.56
ALU_umary_ALU_11333	rs9303542	Ovarian cancer	SKAP1	-0.64
ALU_umary_ALU_11353	rs9635759	Menarche (age at onset)	RPL7P48,CA10	0.42
ALU_umary_ALU_11384	rs7224438	Immune reponse to smallpox (secreted IL-2)	BCAS3	0.41
ALU_umary_ALU_11398	rs4329	Metabolic traits	ACE	0.62
ALU_umary_ALU_11400	rs7223966	Body mass index	.	0.66
ALU_umary_ALU_11422	rs817565	Response to anti-retroviral therapy (ddI/d4T) in HIV-1 infection (Grade 1 peripheral neuropathy)	MAP2K6	0.48
ALU_umary_ALU_11427	rs10775360	QT interval	CALM2P1,CASC17	0.51
ALU_umary_ALU_11428	rs10775360	QT interval	CALM2P1,CASC17	0.42
ALU_umary_ALU_11462	rs16970672	Psychosis and Alzheimer's disease	FLJ45079,TNRC6C	0.42
ALU_umary_ALU_11474	rs7220048	Obesity-related traits	AATK	-0.45
ALU_umary_ALU_11492	rs2345595	PR interval in Tripanosoma cruzi seropositivity	LINC00470,METTL4	0.53
ALU_umary_ALU_11552	rs1893217	Celiac disease or Rheumatoid arthritis	PTPN2	0.77

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_11628	rs7239368	Response to citalopram treatment	NOL4	0.61
ALU_umary_ALU_11673	rs1398217	Menarche (age at onset)	SKOR2	0.45
ALU_umary_ALU_11704	rs4940203	Obesity-related traits	DCC	0.51
ALU_umary_ALU_11708	rs11876941	Body mass index (interaction)	DCC	0.42
ALU_umary_ALU_11760	rs11152166	Major depressive disorder	CCBE1	0.41
ALU_umary_ALU_11766	rs4553720	Adverse response to chemotherapy (neutropenia/leucopenia) (docetaxel)	LINC00305,CDH7	0.69
ALU_umary_ALU_11798	rs637644	Adverse response to chemotherapy in breast cancer (alopecia) (cyclophosphamide+doxorubicin+/-5FU)	LINC00305,CDH7	0.41
ALU_umary_ALU_11899	rs1865075	Dental caries	RPL34P34,ZNF98	0.44
ALU_umary_ALU_12007	rs6117615	Adverse response to chemotherapy in breast cancer (alopecia) (docetaxel)	SLC52A3,FAM110A	0.89
ALU_umary_ALU_12018	rs6139030	Response to hepatitis C treatment	ITPA	0.44
ALU_umary_ALU_12039	rs6054383	Optic cup area	.	0.40
ALU_umary_ALU_12074	rs6042314	Intelligence (childhood)	ESF1	0.51
ALU_umary_ALU_12101	rs6044112	Response to taxane treatment (docetaxel)	KIF16B	0.46
ALU_umary_ALU_12132	rs816535	Parkinson disease and lewy body pathology	.	0.87
ALU_umary_ALU_12143	rs6088765	Ulcerative colitis	PROCR	0.45
ALU_umary_ALU_12183	rs6065906	Triglycerides	PLTP,PCIF1	0.49
ALU_umary_ALU_12184	rs6065906	HDL cholesterol	PLTP,PCIF1	0.60
ALU_umary_ALU_12208	rs6091737	Calcium levels	RNU7-14P,SUMO1P1	0.56
ALU_umary_ALU_12366	rs458685	Breast cancer	GRIK1	0.44
L1_umary_LINE1_2980	rs12483205	HIV-1 replication	DYRK1A	0.41
ALU_umary_ALU_12465	rs132390	Breast cancer	EMID1	0.55
ALU_umary_ALU_12536	rs138880	Schizophrenia	BRD1	0.66

Table 9 Top LD results for polyTE for European population

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_14	rs11120822	Stearic acid (18:0) plasma levels	CAMTA1,LOC100129476	0.43
ALU_umary_ALU_19	rs12711517	Breast cancer	SLC45A1	-0.55
ALU_umary_ALU_20	rs301799	Thyroid peroxidase antibody positivity	LOC102724552,RERE	0.44
SVA_umary_SVA_5	rs17367504	Blood pressure	MTHFR	0.70
ALU_umary_ALU_33	rs2489260	Obesity-related traits	AADACL4,AADACL3	0.82
ALU_umary_ALU_64	rs4654899	Superior frontal gyrus grey matter volume	EIF4G3	0.70
SVA_umary_SVA_18	rs28411352	Rheumatoid arthritis	MTF1	-0.48
ALU_umary_ALU_127	rs2274465	Menarche (age at onset)	KDM4A	0.65
ALU_umary_ALU_139	rs11588062	Age-related hearing impairment (interaction)	UQCRH	0.51
ALU_umary_ALU_189	rs2811893	Diabetic retinopathy	MYSM1	0.45
ALU_umary_ALU_194	rs7534016	Obesity-related traits	FGGY	0.87
ALU_umary_ALU_216	rs1751492	Soluble leptin receptor levels	LEPR	0.45
ALU_umary_ALU_224	rs1417437	Orofacial clefts	LRRC7	0.59
ALU_umary_ALU_226	rs11809230	Cannabis use (initiation)	.	0.49
ALU_umary_ALU_244	rs2568958	Weight	GDI2P2,RPL31P12	-0.44
ALU_umary_ALU_288	rs12024204	Endometriosis	ADH5P2,HMGB1P18	-0.65
ALU_umary_ALU_342	rs12091709	Cognitive function	LRRC8D	0.65
ALU_umary_ALU_390	rs303386	Adverse response to chemotherapy (neutropenia/leucopenia) (all topoisomerase inhibitors)	LOC100129620	0.46
L1_umary_LINE1_98	rs1948368	Bipolar disorder	PPIAP7,RPSAP19	0.77
L1_umary_LINE1_99	rs11578152	Menarche (age at onset)	DNAJA1P5,COL11A1	0.63
ALU_umary_ALU_412	rs10874639	Protein quantitative trait loci	DNAJA1P5,COL11A1	0.55
ALU_umary_ALU_417	rs3934285	Obesity-related traits	AMY1C,FTLP17	0.74
SVA_umary_SVA_45	rs7411387	Interferon alpha levels in systemic lupus erythematosus	CHIA	0.52
ALU_umary_ALU_446	rs10776733	Obesity-related traits	ADORA3	0.46
ALU_umary_ALU_464	rs11102807	Autism	EIF2S2P5,PKMP1	0.42
ALU_umary_ALU_481	rs10802047	Relative hand skill in reading disability	RNA5SP56,PSMC1P12	-0.69
ALU_umary_ALU_510	rs12403795	Illicit drug use	MRPS21	0.49
ALU_umary_ALU_534	rs857684	Red blood cell traits	OR10Z1	0.45
SVA_umary_SVA_57	rs3790672	Testicular germ cell tumor	UCK2	0.83
ALU_umary_ALU_585	rs3903239	Atrial fibrillation	GORAB,PRRX1	-0.42
ALU_umary_ALU_589	rs28588043	Number of children (6+ vs. 0 or 1)	.	0.86
ALU_umary_ALU_602	rs17301853	Migraine - clinic-based	RABGAP1L	0.57
ALU_umary_ALU_618	rs12760731	Obesity-related traits	LINC00083,TEX35	0.62
ALU_umary_ALU_632	rs199950	Body mass index (change over time)	CACNA1E	0.55
ALU_umary_ALU_659	rs12125250	Economic and political preferences	SLC4A1APP2,RPS3AP9	0.43

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_660	rs12125250	Economic and political preferences	SLC4A1APP2,RPS3AP9	0.46
ALU_umary_ALU_664	rs10489764	Amyotrophic lateral sclerosis (sporadic)	SLC4A1APP2,RPS3AP9	0.49
ALU_umary_ALU_672	rs12408261	Number of pregnancies	.	-0.68
ALU_umary_ALU_689	rs10801047	Crohn's disease	HNRNPA1P46,RGS18	0.67
ALU_umary_ALU_713	rs1890645	Neonatal lupus	EEF1A1P14,KCNT2	0.43
ALU_umary_ALU_716	rs426736	Meningococcal disease	CFHR3	0.76
ALU_umary_ALU_717	rs426736	Meningococcal disease	CFHR3	0.71
ALU_umary_ALU_718	rs426736	Meningococcal disease	CFHR3	0.74
ALU_umary_ALU_795	rs12120588	Urate levels in overweight individuals	.	0.57
ALU_umary_ALU_796	rs12120588	Urate levels in overweight individuals	.	-0.50
ALU_umary_ALU_841	rs12410462	Major depressive disorder	BTF3P9,TUBB8P10	0.44
ALU_umary_ALU_842	rs12410462	Major depressive disorder	BTF3P9,TUBB8P10	0.40
ALU_umary_ALU_845	rs801114	Basal cell carcinoma	FTH1P2,ISCA1P2	0.57
ALU_umary_ALU_884	rs482329	Life threatening arrhythmia	LINC00184,LINC01132	-0.72
SVA_umary_SVA_79	rs12135191	Urate levels (BMI interaction)	.	-0.44
ALU_umary_ALU_897	rs2820037	Hypertension	RPL39P10,CHRM3	0.92
ALU_umary_ALU_958	rs10189761	Obesity	FAM150B,TMEM18	0.88
ALU_umary_ALU_971	rs11123610	Response to inhaled corticosteroid treatment in asthma (percentage change of FEV1)	ALLC	0.59
ALU_umary_ALU_1048	rs4635554	Hypertriglyceridemia	TDRD15,RNA5SP87	-0.46
L1_umary_LINE1_267	rs2681019	Dialysis-related mortality	RNA5SP87,KLHL29	-0.87
ALU_umary_ALU_1070	rs3795958	Metabolite levels (HVA/MHPG ratio)	DRC1	0.48
ALU_umary_ALU_1128	rs1863080	Anthropometric traits	MRPL50P1,RPL21P36	0.66
ALU_umary_ALU_1149	rs3816183	Hypospadias	HAAO	-0.50
ALU_umary_ALU_1163	rs2341459	Height	CAMKMT	-0.52
ALU_umary_ALU_1171	rs12474201	Height	CRIPT,SOCS5	0.57
ALU_umary_ALU_1173	rs34198350	QT interval in Tripanosoma cruzi seropositivity	RPS27AP7,VN1R18P	0.74
ALU_umary_ALU_1211	rs12713280	Economic and political preferences	EML6	0.51
ALU_umary_ALU_1216	rs6751715	HIV-1 control	MIR216B,CCDC85A	-0.40
ALU_umary_ALU_1221	rs889956	Educational attainment	EIF2S2P7,VRK2	0.73
SVA_umary_SVA_110	rs3845817	Bipolar disorder	RPS15AP15,KRT18P33	0.60
ALU_umary_ALU_1273	rs4141819	Endometriosis	ETAA1,C1D	0.48
ALU_umary_ALU_1287	rs432203	Longevity	.	-0.51
ALU_umary_ALU_1336	rs10496262	Aging traits	LRRTM1,MTND4P25	0.85
ALU_umary_ALU_1337	rs10496262	Aging traits	LRRTM1,MTND4P25	0.75
L1_umary_LINE1_346	rs12052359	Bilirubin levels	LRRTM1,MTND4P25	0.51
L1_umary_LINE1_348	rs10496289	Hypertension	MTND5P27,RPL37P10	0.61
ALU_umary_ALU_1350	rs7581224	Coronary artery calcification	SUCLG1,DNAH6	-0.41
ALU_umary_ALU_1388	rs7583877	Type 1 diabetes nephropathy	AFF3	-0.50

TE	GWAS hits	GWAS phenotype	GWAS gene	r
L1_umary_LINE1_366	rs1558648	Serum protein levels (sST2)	IL1RL2	0.53
ALU_umary_ALU_1406	rs2163349	Addiction	NCK2	0.88
ALU_umary_ALU_1429	rs13401811	Chronic lymphocytic leukemia	ACOXL	0.66
ALU_umary_ALU_1433	rs11122895	Allergic sensitization	RPL34P8,ANAPC1	0.61
ALU_umary_ALU_1441	rs1823125	Sleep duration	LOC101927400	0.63
ALU_umary_ALU_1555	rs13405020	Non-small cell lung cancer	THSD7B	-0.53
ALU_umary_ALU_1557	rs13405020	Non-small cell lung cancer	THSD7B	0.61
ALU_umary_ALU_1585	rs17515225	Motion sickness	LRP1B	-0.77
ALU_umary_ALU_1621	rs7584099	Response to statin therapy	PABPC1P2,RPL26P14	0.64
ALU_umary_ALU_1623	rs2307394	Urate levels	ORC4	0.71
ALU_umary_ALU_1628	rs10191411	Protein quantitative trait loci	RPS29P8,EPC2	0.43
L1_umary_LINE1_426	rs7594648	Age-related hearing impairment	MTND5P30,NR4A2	0.84
ALU_umary_ALU_1709	rs1424760	Phospholipid levels (plasma)	RNA5SP109,RPL7P61	-0.48
ALU_umary_ALU_1740	rs2268365	Blood pressure (smoking interaction)	LRP2	0.51
ALU_umary_ALU_1778	rs9287989	Periodontal microbiota	EXTL2P1,KIAA1715	-0.43
L1_umary_LINE1_455	rs13413635	Heart rate	PDE11A	0.73
ALU_umary_ALU_1824	rs6741522	Cervical artery dissection	RPL23AP33,ELF2P4	0.58
ALU_umary_ALU_1838	rs2675399	Obesity-related traits	DIRC1,COL3A1	0.56
ALU_umary_ALU_1843	rs2675399	Obesity-related traits	DIRC1,COL3A1	0.46
ALU_umary_ALU_1867	rs2176528	Bipolar disorder	RPS17P8,GLULP6	0.57
ALU_umary_ALU_1894	rs6738825	Crohn's disease	.	-0.81
L1_umary_LINE1_489	rs17266097	Menarche (age at onset)	SATB2	0.41
ALU_umary_ALU_1947	rs12478665	Hippocampal volume	MEAF6P1,MAP2	0.65
ALU_umary_ALU_1961	rs1464443	Amyotrophic lateral sclerosis (sporadic)	ERBB4	-0.48
ALU_umary_ALU_1964	rs4673659	Asthma (sex interaction)	ERBB4	0.66
ALU_umary_ALU_1987	rs16857609	Breast cancer	DIRC3	0.74
ALU_umary_ALU_1992	rs492400	Body mass index	.	0.72
ALU_umary_ALU_2011	rs12621643	Acute lymphoblastic leukemia (childhood)	KCNE4	-0.42
ALU_umary_ALU_2101	rs7652782	Serum uric acid levels	CNTN4	0.70
ALU_umary_ALU_2109	rs2587949	Periodontitis (DPAL)	SUMF1	-0.55
ALU_umary_ALU_2123	rs271066	Alzheimer's disease (age of onset)	MRPS35P1,MRPS36P1	0.50
SVA_umary_SVA_157	rs3729931	Cardiac hypertrophy	RAF1	-0.42
ALU_umary_ALU_2169	rs6771632	Lung function (forced expiratory flow between 25%25 and 75%25 of forced vital capacity)	IMPDH1P8,GALNT15	-0.42
ALU_umary_ALU_2247	rs7617877	Parkinson's disease	LINC00693	-0.59
ALU_umary_ALU_2251	rs4680719	Metabolite levels (HVA)	MESTP4,RBMS3-AS3	-0.41
ALU_umary_ALU_2289	rs75968099	Schizophrenia	HSPD1P6,TRANK1	0.83
L1_umary_LINE1_580	rs11708996	QT interval	SCN5A	0.57
ALU_umary_ALU_2319	rs10865924	Clozapine-induced agranulocytosis	ACKR2	0.88
ALU_umary_ALU_2333	rs319690	Blood pressure	MAP4	-0.52

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_2337	rs319690	Blood pressure	MAP4	0.47
ALU_umary_ALU_2340	rs11719291	Cognitive function	IP6K2	0.80
ALU_umary_ALU_2349	rs1031925	Melanoma	.	0.75
ALU_umary_ALU_2352	rs2029213	Heart rate	DCP1A	0.60
ALU_umary_ALU_2371	rs6764184	Optic cup area	.	0.72
L1_umary_LINE1_629	rs17518584	Cognitive function	CADM2	-0.51
ALU_umary_ALU_2515	rs7632427	Orofacial clefts	EPHA3,PROS2P,PROSP	0.48
ALU_umary_ALU_2575	rs1397924	Economic and political preferences (environmentalism)	MIR548AB,RAP1BP2	0.63
ALU_umary_ALU_2577	rs12485744	Economic and political preferences (environmentalism)	MIR548AB,RAP1BP2	0.81
ALU_umary_ALU_2578	rs2677247	IgG glycosylation	MIR548AB,RAP1BP2	0.65
ALU_umary_ALU_2591	rs12491921	Cannabis dependence	CBLB,FCF1P3	-0.43
ALU_umary_ALU_2627	rs7611694	Prostate cancer	SIDT1	0.43
ALU_umary_ALU_2637	rs9841504	Gastric cancer	LOC102723469,ZBTB20	0.86
ALU_umary_ALU_2667	rs13077101	Obesity-related traits	RABL3	0.84
ALU_umary_ALU_2697	rs2712381	Monocyte count	RPN1	-0.55
ALU_umary_ALU_2733	rs16847609	Alzheimer's disease in APOE e4- carriers	.	0.82
ALU_umary_ALU_2743	rs908821	Multiple sclerosis	TRIM42,RPL23AP41	0.50
ALU_umary_ALU_2757	rs9826463	QRS duration in Tripanosoma cruzi seropositivity	PLS1	0.65
ALU_umary_ALU_2770	rs345013	Prostate cancer	RNA5SP144,LARP7P4	0.90
ALU_umary_ALU_2810	rs1351267	Schizophrenia	SUCNR1,MBNL1	-0.52
ALU_umary_ALU_2845	rs7646881	Tetralogy of Fallot	LOC100287290	0.52
ALU_umary_ALU_2909	rs9864370	Multiple myeloma (hyperdiploidy)	MECOM	0.67
ALU_umary_ALU_2913	rs13097028	Melanoma	SDHDP3,TERC	0.83
ALU_umary_ALU_2925	rs3913363	Response to angiotensin II receptor blocker therapy	TMEM212,FNDC3B	0.79
ALU_umary_ALU_3022	rs10937470	Total ventricular volume	UTS2B	-0.77
SVA_umary_SVA_203	rs11723261	Immune response to smallpox vaccine (IL-6)	.	0.61
ALU_umary_ALU_3058	rs11723261	Immune response to smallpox vaccine (IL-6)	.	0.61
ALU_umary_ALU_3062	rs13108904	Obesity-related traits	LOC101928676,MAEA	-0.77
ALU_umary_ALU_3085	rs16872571	Vitiligo	CLNK,RNPS1P1	-0.69
L1_umary_LINE1_786	rs1503874	Illicit drug use	KRT18P63,RPL21P46	0.47
ALU_umary_ALU_3139	rs4697263	Age-related hearing impairment (interaction)	KCNIP4-IT1,GPR125	0.84
ALU_umary_ALU_3178	rs7442317	Attention deficit hyperactivity disorder motor coordination	EEF1A1P21,RPS3AP17	0.87
ALU_umary_ALU_3195	rs2177312	Very long-chain saturated fatty acid levels (fatty acid 22:0)	PCDH7,MAPRE1P2	0.53
L1_umary_LINE1_818	rs10010758	Periodontal microbiota	TBC1D1	0.44
L1_umary_LINE1_831	rs2055942	Type 2 diabetes	GABRA4	0.52
ALU_umary_ALU_3287	rs6820391	Cervical artery dissection	LNX1	-0.45

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_3295	rs13113518	Height	CLOCK	0.71
ALU_umary_ALU_3340	rs7656244	Kawasaki disease	TECRL	0.49
ALU_umary_ALU_3360	rs1155865	Cognitive test performance	RPS23P3,CENPC	0.77
ALU_umary_ALU_3378	rs4356975	Obesity-related traits	UGT2B7	0.44
ALU_umary_ALU_3409	rs6854845	Survival in rectal cancer	.	0.45
ALU_umary_ALU_3486	rs4416442	Chronic obstructive pulmonary disease (moderate to severe)	FAM13A	0.40
ALU_umary_ALU_3554	rs17030795	Anorexia nervosa	PPP3CA	0.80
L1_umary_LINE1_913	rs6533014	Homeostasis model assessment of beta-cell function (interaction)	SLC39A8,NFKB1	0.54
L1_umary_LINE1_915	rs4699052	Testicular germ cell tumor	CENPE,DDX3P3,DDX3YP3	0.45
L1_umary_LINE1_928	rs10034228	Myopia (pathological)	RPL36AP23,CCDC34P1	0.93
ALU_umary_ALU_3601	rs10034228	Myopia (pathological)	RPL36AP23,CCDC34P1	0.82
L1_umary_LINE1_942	rs6838310	Cognitive function	NT5C3AP1,NDST3	0.45
ALU_umary_ALU_3643	rs10028773	Educational attainment	KLHL2P1	-0.40
ALU_umary_ALU_3782	rs1512281	Percentage gas trapping	.	0.71
ALU_umary_ALU_3796	rs1395821	Coronary heart disease	TTC29,MIR548G	0.54
ALU_umary_ALU_3944	rs6835098	Dementia and core Alzheimer's disease neuropathologic changes	GALNT7,LOC101930370	-0.40
ALU_umary_ALU_3997	rs2130392	Kawasaki disease	CENPU	0.61
ALU_umary_ALU_4041	rs11748327	Myocardial infarction	IRX1,LINC01020	0.72
ALU_umary_ALU_4046	rs16875288	Functional impairment in major depressive disorder%2C bipolar disorder and schizophrenia	ADAMTS16	0.54
ALU_umary_ALU_4055	rs7729273	Cognitive performance	RNA5SP176,ADCY2	0.59
ALU_umary_ALU_4195	rs1173766	Blood pressure	NPR3,RPS8P8	-0.54
ALU_umary_ALU_4218	rs293748	Obesity-related traits	NIPBL	0.74
ALU_umary_ALU_4266	rs9291768	Classic bladder exstrophy	.	0.43
ALU_umary_ALU_4267	rs4865673	Dental caries	HMGB1P47,KATNBL1P4	0.64
ALU_umary_ALU_4270	rs4348174	Serum thyroid-stimulating hormone levels	KATNBL1P4,RPS17P11	0.42
ALU_umary_ALU_4283	rs7716219	Height	SLC38A9	0.60
ALU_umary_ALU_4294	rs16886364	Breast cancer (early onset)	MAP3K1	0.87
ALU_umary_ALU_4375	rs295688	Dysphagia	.	0.48
ALU_umary_ALU_4403	rs672413	Blood and toenail selenium levels	ARSB	0.52
SVA_umary_SVA_248	rs506500	Blood trace element (Se levels)	BHMT	-0.50
ALU_umary_ALU_4406	rs567754	Toenail selenium levels	BHMT	0.40
ALU_umary_ALU_4449	rs6452790	Cognitive function	RPS3AP22,LINC00461	0.68
ALU_umary_ALU_4509	rs10067427	Non-alcoholic fatty liver disease histology (lobular)	EEF1A1P20,MTND5P10	0.68
ALU_umary_ALU_4510	rs10067427	Non-alcoholic fatty liver disease histology (lobular)	EEF1A1P20,MTND5P10	0.65
ALU_umary_ALU_4514	rs4702982	Panic disorder	FAM174A,ST8SIA4	0.68
ALU_umary_ALU_4562	rs4388249	Schizophrenia	MAN2A1	0.89

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_4571	rs3853750	Asthma and hay fever	SLC25A46,TSLP	0.44
ALU_umary_ALU_4589	rs55670112	Epilepsy	KCNN2,TRIM36	0.45
ALU_umary_ALU_4624	rs255788	Response to platinum-based chemotherapy (carboplatin)	FAM170A,PRR16	0.55
ALU_umary_ALU_4646	rs6595551	Type 2 diabetes (young onset) and obesity	ZNF608,RPL28P3	0.59
ALU_umary_ALU_4684	rs6890695	Alzheimer's disease in APOE e4- carriers	.	0.68
ALU_umary_ALU_4717	rs3776331	Uric acid levels	ARHGAP26	0.44
ALU_umary_ALU_4737	rs9325032	Cognitive test performance	PPP2R2B	-0.51
ALU_umary_ALU_4773	rs727809	Age-related hearing impairment (interaction)	TRNAC32P,GRIA1	0.77
ALU_umary_ALU_4796	rs411174	Personality traits in bipolar disorder	ITK	0.47
ALU_umary_ALU_4807	rs2082412	Psoriasis	UBLCP1,IL12B	-0.48
L1_umary_LINE1_1242	rs17504106	Post-traumatic stress disorder	.	0.81
ALU_umary_ALU_4959	rs2236212	Phospholipid levels (plasma)	ELOVL2	0.47
ALU_umary_ALU_4976	rs204247	Breast cancer	RANBP9,MCUR1	-0.75
ALU_umary_ALU_4993	rs2274136	Obesity-related traits	NUP153	0.56
ALU_umary_ALU_5002	rs664154	Information processing speed	.	-0.42
ALU_umary_ALU_5053	rs2523822	Drug-induced liver injury (amoxicillin-clavulanate)	TRNAI25	0.88
ALU_umary_ALU_5055	rs259919	HIV-1 control	ZNRD1-AS1	0.56
ALU_umary_ALU_5058	rs12526186	Response to antipsychotic treatment	HCG20,TRNAI25	0.58
ALU_umary_ALU_5060	rs9263739	Ulcerative colitis	CCHCR1	0.58
SVA_umary_SVA_278	rs9368677	Atopic dermatitis	TRNAI25	0.75
ALU_umary_ALU_5064	rs1055569	Psychotic symptoms and prion disease	.	0.69
ALU_umary_ALU_5065	rs2157337	Rheumatoid arthritis	TRNAI25	0.52
SVA_umary_SVA_280	rs10484561	Follicular lymphoma	TRNAI25	0.78
ALU_umary_ALU_5072	rs4530903	Lymphoma	TRNAI25	0.88
ALU_umary_ALU_5075	rs2858870	Nodular sclerosis Hodgkin lymphoma	TRNAI25	0.69
ALU_umary_ALU_5076	rs7756516	Chronic hepatitis B infection	HLA-DQB2	-0.71
ALU_umary_ALU_5077	rs7756516	Chronic hepatitis B infection	HLA-DQB2	-0.66
ALU_umary_ALU_5079	rs2621416	Lymphoma	TRNAI25	0.62
SVA_umary_SVA_282	rs3077	Chronic hepatitis B infection	HLA-DPA1	0.90
ALU_umary_ALU_5106	rs9296295	Obesity-related traits	KIF6	0.48
ALU_umary_ALU_5120	rs9472155	Vascular endothelial growth factor levels	LINC01512,LOC100132354	0.58
ALU_umary_ALU_5132	rs10948222	Height	SUPT3H	0.87
L1_umary_LINE1_1293	rs9357506	Body mass index	.	-0.87
ALU_umary_ALU_5165	rs1161397	Overweight status	TRNAI25	0.59
ALU_umary_ALU_5209	rs9475752	Menarche (age at onset)	DST	0.58
ALU_umary_ALU_5213	rs9500256	Eosinophilic esophagitis (pediatric)	GAPDHP15,RBBP4P4	0.48
ALU_umary_ALU_5237	rs4710654	Response to amphetamines	RNA5SP208,ADGRB3,BAI3	0.72
ALU_umary_ALU_5239	rs875033	Response to amphetamines	RNA5SP208,ADGRB3,BAI3	0.55

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_5259	rs1048886	Type 2 diabetes	C6orf57,SDHAF4	0.76
ALU_umary_ALU_5262	rs6922893	Obesity-related traits	B3GAT2	0.79
ALU_umary_ALU_5264	rs9351814	Coronary artery disease or ischemic stroke	LINC00472,KRT19P1	0.42
ALU_umary_ALU_5280	rs9447004	Calcium levels	CD109	0.78
ALU_umary_ALU_5296	rs12198063	Capecitabine sensitivity	IMPG1,HTR1B	0.42
ALU_umary_ALU_5324	rs10943724	Thiazide-induced adverse metabolic effects in hypertensive patients	RPL17P25,FAM46A	0.54
ALU_umary_ALU_5395	rs11757063	Migraine	FUT9,UFL1	0.85
ALU_umary_ALU_5402	rs6924808	Response to inhaled glucocorticoid treatment in asthma (percentage change of FEV1)	.	0.49
SVA_umary_SVA_298	rs239198	Menarche (age at onset)	ASCC3	0.44
L1_umary_LINE1_1370	rs1416280	Longevity (90 years and older)	GRIK2,R3HDM2P2	0.43
ALU_umary_ALU_5490	rs9488343	Gray matter volume (schizophrenia interaction)	HS3ST5	0.55
ALU_umary_ALU_5509	rs11153730	Heart rate	RPL29P4,CEP85L	-0.58
ALU_umary_ALU_5582	rs7749983	Periodontal disease-related phenotypes	LOC102723409	0.53
ALU_umary_ALU_5583	rs10447419	PR interval	SAMD3	0.68
L1_umary_LINE1_1418	rs225675	Thiazide-induced adverse metabolic effects in hypertensive patients	VTA1	0.79
ALU_umary_ALU_5657	rs10979	Hypospadias	LOC285740	-0.45
SVA_umary_SVA_315	rs1933488	Prostate cancer	RGS17	0.73
ALU_umary_ALU_5713	rs1449672	Trans fatty acid levels	LOC101928923	0.41
SVA_umary_SVA_320	rs1620921	Lipoprotein (a) - cholesterol levels	PLG,MAP3K4	-0.60
ALU_umary_ALU_5742	rs13191362	Body mass index	.	0.69
ALU_umary_ALU_5746	rs9364687	Body mass index	.	0.55
L1_umary_LINE1_1448	rs59072263	Intraocular pressure	GLCCI1,ICA1	0.51
ALU_umary_ALU_5809	rs9918508	Hippocampal atrophy	RPL9P19,GAPDHP68	0.59
L1_umary_LINE1_1460	rs6961860	Adverse response to chemotherapy (neutropenia/leucopenia) (all antimicrotubule drugs)	RAD17P1,AHR	0.75
ALU_umary_ALU_5868	rs12666612	Obesity-related traits	HDAC9	0.89
ALU_umary_ALU_5886	rs2286503	Fibrinogen	TOMM7	0.69
ALU_umary_ALU_5967	rs2392510	Periodontitis	GPR141	-0.40
ALU_umary_ALU_5969	rs16879765	Dupuytren's disease	EPDR1	0.62
ALU_umary_ALU_5970	rs4723738	Treatment response for severe sepsis	STARD3NL	0.47
ALU_umary_ALU_6007	rs1722133	Sitting height ratio	.	0.47
ALU_umary_ALU_6015	rs1551277	Anxiety disorder	PKD1L1	0.41
ALU_umary_ALU_6027	rs4132601	Acute lymphoblastic leukemia (childhood)	IKZF1	0.42
ALU_umary_ALU_6074	rs10266483	Response to statin therapy	ZNF679,VN1R39P	0.69
ALU_umary_ALU_6114	rs2245368	Body mass index	.	0.67
ALU_umary_ALU_6127	rs62468577	Bronchopulmonary dysplasia	MAGI2	0.72

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_6148	rs2715148	Major depressive disorder	PCLO	0.55
ALU_umary_ALU_6185	rs17301259	Heschl's gyrus morphology	ZNF804B	0.71
ALU_umary_ALU_6272	rs1404697	Smoking behavior	C7orf66,EIF3IP1	0.90
ALU_umary_ALU_6316	rs757278	Response to methotrexate in juvenile idiopathic arthritis	CTTNBP2	0.67
ALU_umary_ALU_6320	rs41997	Response to platinum-based chemotherapy in non-small-cell lung cancer	ANKRD7,GTF3AP6	-0.62
ALU_umary_ALU_6359	rs4731207	Cutaneous malignant melanoma	.	-0.65
ALU_umary_ALU_6372	rs2687481	Hearing function	RPL31P39,GRM8	0.46
ALU_umary_ALU_6373	rs2687481	Hearing function	RPL31P39,GRM8	0.74
ALU_umary_ALU_6374	rs2687481	Hearing function	RPL31P39,GRM8	0.45
ALU_umary_ALU_6427	rs10250997	Autism spectrum disorder%2C attention deficit-hyperactivity disorder%2C bipolar disorder%2C major depressive disorder%2C and schizophrenia (combined)	MTPN,PSMC1P3	0.52
ALU_umary_ALU_6482	rs2708240	QT interval (interaction)	CNTNAP2	-0.46
ALU_umary_ALU_6493	rs17173637	HDL cholesterol	AOC1	0.65
ALU_umary_ALU_6560	rs1045529	Myopia (pathological)	ERI1	0.76
ALU_umary_ALU_6563	rs12545912	Multiple myeloma (hyperdiploidy)	TNKS	0.82
ALU_umary_ALU_6564	rs6601327	Multiple myeloma (hyperdiploidy)	PPP1R3B,TNKS	-0.53
ALU_umary_ALU_6594	rs4831760	Pulmonary function decline	TUSC3	0.75
ALU_umary_ALU_6595	rs4831760	Pulmonary function decline	TUSC3	-0.41
ALU_umary_ALU_6611	rs920590	Acute lymphoblastic leukemia (childhood)	CSGALNACT1,INTS10	0.78
ALU_umary_ALU_6645	rs4732957	Response to amphetamines	ADRA1A	0.43
ALU_umary_ALU_6669	rs11997175	Body mass index	.	-0.58
ALU_umary_ALU_6674	rs6987004	Pulmonary function decline	RPL10AP3,RPL21P80	0.74
ALU_umary_ALU_6760	rs1387221	Clozapine-induced cytotoxicity	.	-0.60
ALU_umary_ALU_6766	rs6984242	Schizophrenia	NUDT15P1,CA8	0.56
ALU_umary_ALU_6805	rs13272623	IgG glycosylation	LACTB2-AS1,LOC286190	0.60
ALU_umary_ALU_6806	rs7017914	Bone mineral density	XKR9	-0.74
ALU_umary_ALU_6814	rs13263568	Migraine	EYA1	0.59
ALU_umary_ALU_6846	rs16939046	Information processing speed	CASC9	0.85
ALU_umary_ALU_6919	rs9969524	Optic disc area	.	-0.43
ALU_umary_ALU_6932	rs160451	Leprosy	RNA5SP272,RIPK2	-0.73
ALU_umary_ALU_6951	rs278567	Bipolar disorder and schizophrenia	C8orf87	0.76
ALU_umary_ALU_6959	rs7818688	Vincristine-induced peripheral neuropathy in acute lymphoblastic leukemia	.	0.68
ALU_umary_ALU_6965	rs3104964	Colorectal cancer	C8orf37-AS1,LOC100616530	0.56
ALU_umary_ALU_6990	rs2033562	IgA nephropathy	.	-0.42

TE	GWAS hits	GWAS phenotype	GWAS gene	r
SVA_umary_SVA_389	rs374810	Ossification of the posterior longitudinal ligament of the spine	RSPO2	0.56
ALU_umary_ALU_7045	rs36068923	Schizophrenia	RPSAP48,EEF1A1P37	0.62
ALU_umary_ALU_7125	rs13268726	Amyotrophic lateral sclerosis	SQLE	0.51
ALU_umary_ALU_7143	rs13281615	Breast cancer	LOC101930033	0.62
ALU_umary_ALU_7145	rs4733601	Diffuse large B cell lymphoma	MIR1208,LINC00824,LINC01263	0.53
L1_umary_LINE1_1761	rs16904191	Migraine	MIR5194,ASAP1	0.66
ALU_umary_ALU_7166	rs7004484	Survival in rectal cancer	.	0.83
L1_umary_LINE1_1787	rs4626664	Restless legs syndrome	PTPRD	0.82
ALU_umary_ALU_7311	rs3904778	Adolescent idiopathic scoliosis	.	0.54
ALU_umary_ALU_7320	rs10810865	Cognitive performance	PABPC1P11,PUS7P1	0.54
ALU_umary_ALU_7331	rs7867456	Axial length	HACD4,IFNNP1,PTPLAD2	-0.56
ALU_umary_ALU_7381	rs10969853	Alcohol dependence (age at onset)	RBMXP2,KRT18P66	0.40
SVA_umary_SVA_401	rs10758189	IgG glycosylation	B4GALT1	0.82
SVA_umary_SVA_402	rs11574914	Rheumatoid arthritis	CCL21,LOC101929761	0.84
ALU_umary_ALU_7420	rs4878712	HIV-1 susceptibility	.	-0.45
ALU_umary_ALU_7547	rs2814828	Height	SPATA31C2,RPSAP49	0.44
ALU_umary_ALU_7555	rs883924	Hepatitis C induced liver fibrosis	LINC01508,LOC101927873	-0.43
L1_umary_LINE1_1863	rs4743820	Inflammatory bowel disease	SYK,AUH	0.45
ALU_umary_ALU_7564	rs944990	Body mass index	.	0.55
L1_umary_LINE1_1867	rs755109	Quantitative traits	HEMGN	0.56
ALU_umary_ALU_7593	rs7028939	Preeclampsia	ERP44	0.72
ALU_umary_ALU_7594	rs7028939	Preeclampsia	ERP44	0.88
ALU_umary_ALU_7615	rs10990268	Tourette syndrome	ZYG11AP1,CYLC2	0.54
ALU_umary_ALU_7620	rs144649413	Metabolite levels (MHPG)	CYLC2,RNA5SP291	0.59
ALU_umary_ALU_7645	rs7048146	Vascular brain injury	YBX1P6,PALM2	0.46
ALU_umary_ALU_7653	rs1889321	Pulmonary function decline	SVEP1	0.56
ALU_umary_ALU_7654	rs10980508	Type 2 diabetes (dietary heme iron intake interaction)	SVEP1,MUSK	0.46
ALU_umary_ALU_7709	rs888219	Response to antipsychotic treatment	PBX3,MVB12B	-0.40
ALU_umary_ALU_7750	rs7092929	Coronary artery calcification	PITRM1-AS1,KLF6	-0.44
ALU_umary_ALU_7864	rs10508727	Immune response to smallpox vaccine (IL-6)	MKX	0.57
ALU_umary_ALU_7994	rs11005694	Antibody status in Tripanosoma cruzi seropositivity	ZWINT,MIR3924	0.57
ALU_umary_ALU_8012	rs10761482	Schizophrenia	ANK3	-0.51
ALU_umary_ALU_8021	rs224136	Crohn's disease	ZNF365,ALDH7A1P4	0.58
ALU_umary_ALU_8022	rs442309	Vogt-Koyanagi-Harada syndrome	ZNF365,ALDH7A1P4	-0.52
ALU_umary_ALU_8067	rs1900005	Vertical cup-disc ratio	ATOH7,KRT19P4	-0.55
SVA_umary_SVA_450	rs12571093	Optic nerve measurement (disc area)	ATOH7,KRT19P4	0.43
ALU_umary_ALU_8138	rs791888	Magnesium levels	.	0.64

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_8178	rs1934955	Blood metabolite levels	CYP2C59P,CYP2C8	0.74
ALU_umary_ALU_8182	rs56322409	Blood metabolite levels	ALDH18A1	0.68
ALU_umary_ALU_8202	rs7069733	Autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia (combined)	.	0.47
ALU_umary_ALU_8232	rs11195062	Multiple myeloma	MXI1	0.48
ALU_umary_ALU_8297	rs7069346	Migraine without aura	C10orf88,PSTK	0.66
SVA_umary_SVA_488	rs1472189	Metabolite levels (Dihydroxy docosatrienoic acid)	USP47,DKK3	0.67
ALU_umary_ALU_8387	rs2727405	Obesity-related traits	RASSF10,ARNTL	0.55
ALU_umary_ALU_8392	rs12287212	Vitamin D levels	RRAS2,COPB1	0.68
L1_umary_LINE1_2073	rs12788764	Age-related nuclear cataracts	LUZP2,RPL36AP40	-0.48
ALU_umary_ALU_8454	rs11602337	Vascular brain injury	LUZP2,RPL36AP40	0.74
ALU_umary_ALU_8457	rs10834691	IgG glycosylation	LUZP2,RPL36AP40	0.43
ALU_umary_ALU_8468	rs12295638	Obesity (extreme)	ANO3	0.72
ALU_umary_ALU_8489	rs2057178	Tuberculosis	RCN1,WT1	0.41
ALU_umary_ALU_8490	rs2057178	Tuberculosis	RCN1,WT1	0.44
ALU_umary_ALU_8492	rs10767971	Parkinson's disease (age of onset)	PRRG4,QSER1	-0.49
L1_umary_LINE1_2092	rs331463	Rheumatoid arthritis	PRR5L,TRAF6	0.81
L1_umary_LINE1_2104	rs10768747	Post-traumatic stress disorder	.	0.94
ALU_umary_ALU_8549	rs9300039	Type 2 diabetes	RPL9P23,HNRNPKP3	0.58
ALU_umary_ALU_8559	rs2176598	Body mass index	.	0.72
ALU_umary_ALU_8566	rs10838725	Alzheimer's disease (late onset)	CELF1	0.53
ALU_umary_ALU_8572	rs11246602	HDL cholesterol	OR4C46,OR4C7P	0.84
ALU_umary_ALU_8574	rs1814175	Height	CBX3P8,TRIM51FP	-0.74
ALU_umary_ALU_8583	rs2220004	Odorant perception (&beta%3B-damascenone)	OR8H3,OR5BN1P	0.56
ALU_umary_ALU_8585	rs7927370	Systemic lupus erythematosus	OR4A15	0.72
L1_umary_LINE1_2125	rs7927370	Systemic lupus erythematosus	OR4A15	0.70
ALU_umary_ALU_8590	rs11228719	Orofacial clefts	OR2AH1P,OR9G1	0.49
ALU_umary_ALU_8591	rs7927370	Systemic lupus erythematosus	OR4A15	0.61
ALU_umary_ALU_8620	rs478304	Acne (severe)	RNASEH2C,KRT8P26	-0.79
ALU_umary_ALU_8622	rs524281	Electroencephalogram traits	PACS1	0.77
ALU_umary_ALU_8629	rs12808519	Urate levels in overweight individuals	.	0.42
ALU_umary_ALU_8698	rs17817600	Alzheimer's disease	PICALM	0.47
ALU_umary_ALU_8717	rs10830228	Age-related macular degeneration	RNU6-16P,TYR	-0.72
ALU_umary_ALU_8801	rs313426	Toenail selenium levels	DYNC2H1	0.56
ALU_umary_ALU_8804	rs10895547	LDL cholesterol	PDGFD	0.80
L1_umary_LINE1_2187	rs7945071	Cognitive function	RDX,FDX1	0.42
L1_umary_LINE1_2191	rs2250417	Inflammatory biomarkers	BCO2	-0.57

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_8951	rs11062040	Response to gemcitabine in pancreatic cancer	.	-0.59
ALU_umary_ALU_9052	rs7134375	HDL cholesterol	TCP1P3,PDE3A	0.51
ALU_umary_ALU_9104	rs12371778	Breast size	PTHLH,CCDC91	0.90
ALU_umary_ALU_9105	rs11049611	Height	CCDC91	0.76
ALU_umary_ALU_9107	rs11049611	Height	CCDC91	0.84
ALU_umary_ALU_9108	rs1979679	Ossification of the posterior longitudinal ligament of the spine	CCDC91	0.86
ALU_umary_ALU_9150	rs826838	Heart rate	CPNE8	0.53
ALU_umary_ALU_9158	rs10467147	Obesity-related traits	LRRK2,MUC19	0.58
ALU_umary_ALU_9169	rs285575	Body mass index	.	0.76
ALU_umary_ALU_9177	rs7978895	Type 2 diabetes	.	-0.53
ALU_umary_ALU_9215	rs17655565	Plasma amyloid beta peptide concentrations (ABx-42)	KRT86	0.50
ALU_umary_ALU_9228	rs11575234	Inflammatory skin disease	.	0.91
ALU_umary_ALU_9233	rs10876993	Celiac disease or Rheumatoid arthritis	B4GALNT1,RPL13AP23	0.67
ALU_umary_ALU_9268	rs7301016	IgG glycosylation	MON2	0.69
ALU_umary_ALU_9310	rs2904524	Amyotrophic lateral sclerosis (age of onset)	CNOT2	0.78
ALU_umary_ALU_9320	rs1495377	Creutzfeldt-Jakob disease (variant)	TSPAN8,LGR5	-0.54
ALU_umary_ALU_9331	rs7964120	Obesity-related traits	RPL31P48,VENTXP3	0.52
ALU_umary_ALU_9355	rs17788937	Myopia (pathological)	NAV3	0.89
ALU_umary_ALU_9388	rs1511589	Optic disc area	.	0.78
L1_umary_LINE1_2324	rs1545843	Major depressive disorder	RPL6P25,SLC6A15	0.43
ALU_umary_ALU_9398	rs7132746	Lewy body disease	N/A	0.76
ALU_umary_ALU_9450	rs7953959	IgG glycosylation	TRNAQ46P,RMST,TRQ-TTG9-1	0.74
ALU_umary_ALU_9499	rs10444533	Social autistic-like traits	RIC8B	0.53
ALU_umary_ALU_9509	rs59227481	Age-related nuclear cataracts	MMAB	0.43
ALU_umary_ALU_9517	rs6490294	Mean platelet volume	ACAD10	0.53
ALU_umary_ALU_9537	rs11064768	Schizophrenia	CCDC60	0.64
ALU_umary_ALU_9553	rs1716403	Response to fenofibrate (adiponectin levels)	ZNF664-FAM101A	0.47
ALU_umary_ALU_9589	rs12282	Immune response to smallpox vaccine (IL-6)	GOLGA3	0.44
ALU_umary_ALU_9602	rs9788333	Thiazide-induced adverse metabolic effects in hypertensive patients	MIPEPP3	0.53
ALU_umary_ALU_9625	rs17079928	Orofacial clefts	SPATA13	0.45
SVA_umary_SVA_560	rs1816752	Obesity-related traits	CYCSP33,PARP4	0.41
ALU_umary_ALU_9639	rs10507349	Type 2 diabetes	RNF6	-0.41
ALU_umary_ALU_9670	rs7331540	IgG glycosylation	FRY	0.43
ALU_umary_ALU_9701	rs6563569	Tourette's syndrome or obsessive-compulsive disorder	TRPC4	0.59
ALU_umary_ALU_9724	rs7336933	Calcium levels	VWA8-AS1,RPS28P8	0.41
ALU_umary_ALU_9726	rs4142110	Nephrolithiasis	DGKH	0.50
L1_umary_LINE1_2401	rs4142110	Nephrolithiasis	DGKH	0.69

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_9727	rs9594738	Bone mineral density	FABP3P2,TNFSF11	-0.42
ALU_umary_ALU_9775	rs9568281	Multiple sclerosis	RCBTB1,ARL11	0.52
ALU_umary_ALU_9789	rs9536318	Airflow obstruction	PCDH8,OLFM4	0.65
ALU_umary_ALU_9798	rs1116255	Post-traumatic stress disorder	.	0.78
ALU_umary_ALU_9803	rs9527419	Response to platinum-based chemotherapy (cisplatin)	MIR5007,HNF4GP1	0.60
ALU_umary_ALU_9813	rs9537938	Educational attainment	RNA5SP30,CTAGE16P	-0.60
ALU_umary_ALU_9815	rs9563576	Body mass index	.	0.68
ALU_umary_ALU_9834	rs4886238	Menopause (age at onset)	TDRD3	0.52
L1_umary_LINE1_2422	rs1847505	Polychlorinated biphenyl levels	.	0.50
ALU_umary_ALU_9847	rs9528384	Verbal declarative memory	PCDH20,RAC1P8	0.54
ALU_umary_ALU_9863	rs9540294	Recalcitrant atopic dermatitis	.	0.83
ALU_umary_ALU_9869	rs1333026	Body mass index	STARP1,HNRNPA3P5	0.54
L1_umary_LINE1_2451	rs1324913	Menarche (age at onset)	KLF12	0.40
ALU_umary_ALU_9941	rs975739	Hair color	MIR3665,EDNRB-AS1	-0.61
ALU_umary_ALU_9943	rs975739	Hair color	MIR3665,EDNRB-AS1	-0.50
ALU_umary_ALU_9951	rs9601248	Major depressive disorder	NDFIP2,LINC00382	0.56
ALU_umary_ALU_9959	rs11149178	Major depressive disorder	PWWP2AP1,ARF4P4	0.54
ALU_umary_ALU_9960	rs6563199	Height	ARF4P4,HIGD1AP2	-0.47
ALU_umary_ALU_10026	rs2352028	Lung cancer	GPC5	-0.48
ALU_umary_ALU_10031	rs4771859	Adverse response to chemotherapy (neutropenia/leucopenia) (all antimicrotubule drugs)	GPC5	-0.66
ALU_umary_ALU_10059	rs285098	Migraine	FARP1	0.55
ALU_umary_ALU_10096	rs12871532	Autism spectrum disorder%2C attention deficit-hyperactivity disorder%2C bipolar disorder%2C major depressive disorder%2C and schizophrenia (combined)	FAM155A-IT1,LIG4	0.65
ALU_umary_ALU_10121	rs1278769	Interstitial lung disease	ATP11A	-0.42
ALU_umary_ALU_10152	rs10147992	White blood cell types	STXBP6	0.48
ALU_umary_ALU_10246	rs1612141	QT interval (interaction)	FBXO33,LRFN5	0.66
ALU_umary_ALU_10250	rs2154294	Alcoholism (12-month weekly alcohol consumption)	OR10V7P,YWHAQP1	0.47
ALU_umary_ALU_10258	rs2488856	Osteoprotegerin levels	YWHAQP1,TUBBP3	0.57
ALU_umary_ALU_10284	rs7144383	Idiopathic pulmonary fibrosis	MDGA2	0.48
ALU_umary_ALU_10312	rs12434047	Economic and political preferences (fairness)	DDHD1,RPS3AP46	0.47
ALU_umary_ALU_10325	rs2274273	Protein biomarker	DLGAP5	0.72
ALU_umary_ALU_10401	rs910316	Height	TMED10	0.75
SVA_umary_SVA_613	rs6574644	Obesity-related traits	STON2	-0.47
ALU_umary_ALU_10518	rs4900384	Type 1 diabetes	C14orf64,C14orf177,LINC01550	0.69
ALU_umary_ALU_10560	rs11858159	Platelet thrombus formation	.	-0.41

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_10562	rs35600665	Obesity-related traits	PWRN3,PWRN1	0.48
L1_umary_LINE1_2634	rs587847	Intraocular pressure	MIR8063,RPS15P8	0.66
ALU_umary_ALU_10670	rs2467853	Renal function and chronic kidney disease	SPATA5L1	-0.51
SVA_umary_SVA_630	rs12594515	Weight	SQRDL,SEMA6D	0.57
ALU_umary_ALU_10674	rs11633886	Diisocyanate-induced asthma	.	0.54
ALU_umary_ALU_10693	rs8023445	Major depressive disorder	SHC4	0.88
ALU_umary_ALU_10695	rs10519227	Thyroid hormone levels	FAM227B,FGF7	0.81
L1_umary_LINE1_2647	rs1124769	Cognitive performance	TNFAIP8L3	0.72
ALU_umary_ALU_10749	rs7179456	Asperger disorder	SLTM	0.54
L1_umary_LINE1_2659	rs7172342	Schizophrenia	RORA	0.59
ALU_umary_ALU_10766	rs1436958	IgG glycosylation	VPS13C	0.84
ALU_umary_ALU_10778	rs7170930	Response to cytadine analogues (cytosine arabinoside)	MIR4311,DIS3L	0.65
ALU_umary_ALU_10786	rs2241423	Body mass index	MAP2K5	0.44
ALU_umary_ALU_10812	rs8038465	Liver enzyme levels (gamma-glutamyl transferase)	CD276	0.78
ALU_umary_ALU_10819	rs3099143	Recalcitrant atopic dermatitis	.	0.72
ALU_umary_ALU_10821	rs2404602	Blood metabolite levels	SCAPER	-0.67
ALU_umary_ALU_10824	rs2137111	Anticoagulant levels	HMG20A,LINGO1	0.47
ALU_umary_ALU_10830	rs950776	Sudden cardiac arrest	CHRNB4	0.50
ALU_umary_ALU_10832	rs2289700	Bipolar disorder	CTSH	0.61
ALU_umary_ALU_10841	rs2663905	QT interval (interaction)	MESDC1,ANP32BP3	-0.44
ALU_umary_ALU_10855	rs6496044	Interstitial lung disease	AKAP13,LOC101929656	0.48
ALU_umary_ALU_10921	rs4533267	Height	ADAMTS17	-0.42
ALU_umary_ALU_10963	rs7200786	Multiple sclerosis	CLEC16A	0.65
ALU_umary_ALU_11003	rs7404095	Inflammatory bowel disease	PRKCB	-0.71
ALU_umary_ALU_11054	rs17291845	Information processing speed	IRX5,IRX6	0.73
SVA_umary_SVA_673	rs2865531	Pulmonary function	CFDP1	-0.49
ALU_umary_ALU_11154	rs12933472	Glucose homeostasis traits	CDH13	0.69
ALU_umary_ALU_11164	rs17789174	Dysphagia	.	0.62
ALU_umary_ALU_11244	rs8082590	Schizophrenia	GID4	-0.55
ALU_umary_ALU_11275	rs225212	Hypertension risk in short sleep duration	MYO1D	0.83
ALU_umary_ALU_11298	rs6607284	Bipolar disorder and schizophrenia	.	0.45
SVA_umary_SVA_704	rs199533	Parkinson's disease	NSF	0.65
SVA_umary_SVA_705	rs12185268	Parkinson's disease	MAPT-AS1,SPPL2C	0.71
SVA_umary_SVA_706	rs12373124	Male-pattern baldness	MAPT-AS1,SPPL2C	0.92
ALU_umary_ALU_11330	rs8070463	Ankylosing spondylitis	KPNB1,TBKBP1	0.43
ALU_umary_ALU_11333	rs9303542	Ovarian cancer	SKAP1	0.85
ALU_umary_ALU_11336	rs2411984	Sex hormone-binding globulin levels	LOC102724596	0.41
ALU_umary_ALU_11398	rs4351	Blood metabolite levels	ACE	0.68

TE	GWAS hits	GWAS phenotype	GWAS gene	r
ALU_umary_ALU_11400	rs11658329	Height	LOC101927898,MAP3K3	0.65
ALU_umary_ALU_11428	rs10775360	QT interval	CALM2P1,CASC17	0.65
L1_umary_LINE1_2778	rs8066985	Waist-to-hip ratio adjusted for body mass index	.	0.56
ALU_umary_ALU_11447	rs8066857	Amyotrophic lateral sclerosis	SLC39A11	0.41
ALU_umary_ALU_11484	rs1291183	Pulmonary function in asthmatics	YES1,ADCYAP1	0.46
ALU_umary_ALU_11492	rs1992269	Alzheimer's disease (late onset)	.	0.44
ALU_umary_ALU_11525	rs7244245	Response to anti-retroviral therapy (ddI/d4T) in HIV-1 infection (Grade 1 peripheral neuropathy)	MTCL1,RPS4XP19	0.45
ALU_umary_ALU_11576	rs7235440	Obesity-related traits	HRH4,RAC1P1	0.44
ALU_umary_ALU_11598	rs11083271	Non-alcoholic fatty liver disease histology (lobular)	CDH2,ARIH2P1	0.72
ALU_umary_ALU_11708	rs11876941	Body mass index (interaction)	DCC	-0.52
ALU_umary_ALU_11719	rs12959570	Tourette's syndrome or obsessive-compulsive disorder	WDR7	0.55
ALU_umary_ALU_11754	rs62096106	Response to abacavir-containing treatment in HIV-1 infection (virologic failure)	PIGN	0.43
ALU_umary_ALU_11766	rs4553720	Adverse response to chemotherapy (neutropenia/leucopenia) (docetaxel)	LINC00305,CDH7	-0.78
L1_umary_LINE1_2860	rs2406342	Adverse response to chemotherapy (neutropenia/leucopenia) (cisplatin)	ARL2BPP1,ZNF236	0.48
ALU_umary_ALU_11946	rs11673344	Obesity-related traits	ZNF585B	0.48
L1_umary_LINE1_2883	rs2288912	Very long-chain saturated fatty acid levels (fatty acid 20:0)	APOC2,APOC4,APOC4-APOC2	0.52
ALU_umary_ALU_12024	rs965469	IFN-related cytopenia	C20orf194	0.44
ALU_umary_ALU_12050	rs6077414	Estradiol plasma levels (breast cancer)	PLCB1	0.42
ALU_umary_ALU_12052	rs6056209	Cognitive performance	PLCB1	-0.60
ALU_umary_ALU_12098	rs932541	Intelligence	KIF16B	0.41
ALU_umary_ALU_12132	rs816535	Parkinson disease and lewy body pathology	.	0.90
ALU_umary_ALU_12143	rs17310467	Hemostatic factors and hematological phenotypes	MYH7B	0.50
ALU_umary_ALU_12145	rs11906854	Migraine - clinic-based	PHF20	0.58
ALU_umary_ALU_12207	rs2041278	Obesity-related traits	ZNF217,RNU7-14P	0.49
ALU_umary_ALU_12382	rs2833607	Vitiligo	HUNK,LINC00159	0.64
L1_umary_LINE1_2987	rs11089937	Periodontitis (PAL4Q3)	IGL	0.69
ALU_umary_ALU_12449	rs11089937	Periodontitis (PAL4Q3)	IGL	0.70
ALU_umary_ALU_12461	rs739310	Obesity-related traits	ISCA2P1,MIAT	0.50
ALU_umary_ALU_12480	rs12530	IgG glycosylation	RTCB	0.87
ALU_umary_ALU_12481	rs12530	IgG glycosylation	RTCB	0.75
ALU_umary_ALU_12536	rs138880	Schizophrenia	BRD1	-0.42

Table 10 Genome-wide significant TE-eQTL for African population

Chr	Pos	TE	Gene	t Statistic	P-value	FDR
13	49536621	ALU_umary_ALU_9771	SLC7A2	7.98	7.27E-12	6.29E-05
6	29892872	ALU_umary_ALU_5054	JPH1	7.64	3.37E-11	1.46E-04
20	52273671	ALU_umary_ALU_12207	EPB41L4B	7.51	6.18E-11	1.78E-04
19	30850007	ALU_umary_ALU_11922	LILRA1	7.20	2.49E-10	4.75E-04
12	92483455	ALU_umary_ALU_9431	IGHV3-20	7.18	2.75E-10	4.75E-04
11	7435902	L1_umary_LINE1_2054	IMPG1	7.12	3.59E-10	5.17E-04
2	161430249	ALU_umary_ALU_1693	IGKV1D-12	7.01	5.90E-10	6.80E-04
1	116548031	L1_umary_LINE1_118	IGHV2-26	6.99	6.28E-10	6.80E-04
2	161430249	ALU_umary_ALU_1693	IGKV1-12	6.84	1.24E-09	1.20E-03
18	12884841	ALU_umary_ALU_11552	RP4-614O4.11	6.66	2.84E-09	2.42E-03
19	30850007	ALU_umary_ALU_11922	RP11-304L19.5	6.64	3.08E-09	2.42E-03
19	30850007	ALU_umary_ALU_11922	TEX22	6.53	5.02E-09	3.59E-03
4	9704849	SVA_umary_SVA_206	KIAA1462	6.51	5.39E-09	3.59E-03
19	30850007	ALU_umary_ALU_11922	SCNN1D	6.46	6.62E-09	4.09E-03
11	7435902	L1_umary_LINE1_2054	SLC35F4	6.45	7.13E-09	4.11E-03
19	30850007	ALU_umary_ALU_11922	ZNF667	6.34	1.15E-08	6.05E-03
11	35348849	ALU_umary_ALU_8504	ABCC3	6.33	1.19E-08	6.05E-03
19	30850007	ALU_umary_ALU_11922	RP11-122K13.12	6.31	1.29E-08	6.20E-03
19	30850007	ALU_umary_ALU_11922	HAUS3	6.26	1.65E-08	7.51E-03
8	122731020	ALU_umary_ALU_7106	RP4-669L17.2	6.20	2.11E-08	8.94E-03
19	30850007	ALU_umary_ALU_11922	ZNF667-AS1	6.19	2.17E-08	8.94E-03
4	9704849	SVA_umary_SVA_206	IGLV3-21	6.17	2.41E-08	9.08E-03
10	78496693	ALU_umary_ALU_8096	SYT5	6.17	2.41E-08	9.08E-03
13	49536621	ALU_umary_ALU_9771	FERMT1	6.14	2.73E-08	9.85E-03
20	52273671	ALU_umary_ALU_12207	LHFPL3	6.07	3.62E-08	1.25E-02
6	32657952	ALU_umary_ALU_5075	HLA-DQB1-AS1	-6.06	3.89E-08	1.30E-02
13	49536621	ALU_umary_ALU_9771	SNORD116-20	6.03	4.45E-08	1.38E-02
20	2371733	ALU_umary_ALU_12019	ZCCHC17	6.02	4.46E-08	1.38E-02
20	42310903	ALU_umary_ALU_12175	ALPL	6.01	4.68E-08	1.40E-02
6	29892872	ALU_umary_ALU_5054	GHRLOS	6.00	5.05E-08	1.44E-02
11	7435902	L1_umary_LINE1_2054	NEB	5.99	5.17E-08	1.44E-02
12	92483455	ALU_umary_ALU_9431	PXMP2	5.98	5.32E-08	1.44E-02
3	72016139	ALU_umary_ALU_2429	MCEE	5.96	5.88E-08	1.54E-02
11	7435902	L1_umary_LINE1_2054	RBM44	5.92	6.98E-08	1.69E-02

Chr	Pos	TE	Gene	t Statistic	P-value	FDR
11	7435902	L1_umary_LINE1_2054	RP4-738P11.3	5.92	6.98E-08	1.69E-02
2	102912450	L1_umary_LINE1_366	DPPA3P2	5.92	7.05E-08	1.69E-02
2	56373531	ALU_umary_ALU_1216	TNS1	5.90	7.59E-08	1.78E-02
11	7435902	L1_umary_LINE1_2054	SLC15A2	5.86	9.04E-08	2.06E-02
1	116548031	L1_umary_LINE1_118	TMEM56	5.83	1.02E-07	2.26E-02
11	7435902	L1_umary_LINE1_2054	RP11-409I10.2	5.81	1.11E-07	2.39E-02
15	66921115	ALU_umary_ALU_10780	RP11-829H16.2	5.80	1.15E-07	2.43E-02
19	30850007	ALU_umary_ALU_11922	HMX2	5.80	1.19E-07	2.45E-02
10	73193547	ALU_umary_ALU_8081	NRP2	5.75	1.42E-07	2.86E-02
2	161430249	ALU_umary_ALU_1693	GALNT9	5.73	1.60E-07	3.05E-02
1	116548031	L1_umary_LINE1_118	RP11-303E16.3	5.72	1.61E-07	3.05E-02
20	42310903	ALU_umary_ALU_12175	IGKV2D-28	5.72	1.63E-07	3.05E-02
11	7435902	L1_umary_LINE1_2054	RP4-738P11.4	5.72	1.67E-07	3.05E-02
20	52369852	ALU_umary_ALU_12208	CTC-248O19.1	5.70	1.75E-07	3.05E-02
4	9704849	SVA_umary_SVA_206	STRC	5.70	1.79E-07	3.05E-02
11	7435902	L1_umary_LINE1_2054	IGDCC4	5.70	1.82E-07	3.05E-02
20	52369852	ALU_umary_ALU_12208	BDKRB2	5.69	1.83E-07	3.05E-02
3	72016139	ALU_umary_ALU_2429	EEF1E1	5.69	1.88E-07	3.05E-02
2	144010793	L1_umary_LINE1_410	EPS8L1	5.68	1.93E-07	3.05E-02
19	30850007	ALU_umary_ALU_11922	RP11-166B2.1	5.68	1.93E-07	3.05E-02
11	77905200	ALU_umary_ALU_8655	VPS13B	5.68	1.94E-07	3.05E-02
19	30850007	ALU_umary_ALU_11922	ARVCF	5.68	1.98E-07	3.05E-02
8	115604486	ALU_umary_ALU_7074	RPS15AP1	5.62	2.53E-07	3.84E-02
20	52369852	ALU_umary_ALU_12208	PRKAG1	5.61	2.58E-07	3.84E-02
17	43383714	ALU_umary_ALU_11326	TAT	5.61	2.62E-07	3.84E-02
15	22797908	ALU_umary_ALU_10550	RP11-481K16.2	5.60	2.66E-07	3.84E-02
19	30850007	ALU_umary_ALU_11922	ZNF10	5.60	2.71E-07	3.85E-02
7	86968519	ALU_umary_ALU_6178	TP53TG1	-5.57	3.03E-07	4.21E-02
10	130888837	ALU_umary_ALU_8317	FREM1	5.57	3.11E-07	4.21E-02
11	7435902	L1_umary_LINE1_2054	CTA-134P22.2	5.56	3.15E-07	4.21E-02
8	122731020	ALU_umary_ALU_7106	AC139099.5	5.56	3.16E-07	4.21E-02
8	132942724	ALU_umary_ALU_7166	AC019221.4	5.56	3.22E-07	4.23E-02
6	29892872	ALU_umary_ALU_5054	AF165138.7	5.54	3.52E-07	4.48E-02
8	21879122	SVA_umary_SVA_369	RMI2	5.53	3.57E-07	4.48E-02
11	35348849	ALU_umary_ALU_8504	AS3MT	5.53	3.61E-07	4.48E-02
19	30850007	ALU_umary_ALU_11922	SNHG9	5.53	3.63E-07	4.48E-02
6	29892872	ALU_umary_ALU_5054	GGTA1P	5.52	3.75E-07	4.53E-02
22	44324605	L1_umary_LINE1_2991	IGHV3-65	5.52	3.81E-07	4.53E-02
11	77905200	ALU_umary_ALU_8655	RP11-234B24.2	5.52	3.82E-07	4.53E-02
11	72381238	ALU_umary_ALU_8635	C9orf57	5.51	3.89E-07	4.53E-02
20	42310903	ALU_umary_ALU_12175	LAMB1	5.51	3.93E-07	4.53E-02
3	111271467	ALU_umary_ALU_2617	ROCK1P1	-5.50	4.14E-07	4.72E-02

Chr	Pos	TE	Gene	t Statistic	P-value	FDR
5	8749528	L1_umary_LINE1_1049	CLDN6	5.49	4.21E-07	4.74E-02
11	7435902	L1_umary_LINE1_2054	C20orf203	5.49	4.35E-07	4.81E-02
11	7435902	L1_umary_LINE1_2054	FREM2	5.48	4.41E-07	4.81E-02
6	114078807	ALU_umary_ALU_5489	IGLV2-5	5.48	4.45E-07	4.81E-02
6	114078807	ALU_umary_ALU_5489	BAMBI	5.48	4.55E-07	4.83E-02
6	29892872	ALU_umary_ALU_5054	FITM1	5.47	4.58E-07	4.83E-02
17	55926227	ALU_umary_ALU_11375	TMEM56	5.47	4.68E-07	4.88E-02

Table 11 Genome-wide significant TE-eQTL for European population

Chr	Pos	TE	Gene	t Statistic	P-value	FDR
17	43660599	SVA_umary_SVA_704	RP11-259G18.3	11.77	3.32E-27	2.47E-20
22	19210913	L1_umary_LINE1_2986	CLTCL1	11.05	1.43E-24	5.30E-18
17	43660599	SVA_umary_SVA_704	KANSL1-AS1	10.44	2.08E-22	5.16E-16
17	43660599	SVA_umary_SVA_704	RP11-259G18.2	9.89	1.55E-20	2.88E-14
9	33130564	SVA_umary_SVA_401	B4GALT1	-9.76	4.47E-20	6.64E-14
6	33030313	SVA_umary_SVA_282	HLA-DPB2	7.74	1.05E-13	1.30E-07
12	58359071	ALU_umary_ALU_9234	XRCC6BP1	7.50	5.21E-13	5.54E-07
6	32589834	ALU_umary_ALU_5072	HLA-DRB5	-7.43	8.49E-13	7.89E-07
17	43660599	SVA_umary_SVA_704	LRRC37A4P	-7.01	1.22E-11	1.01E-05
6	32657952	ALU_umary_ALU_5075	HLA-DQB1-AS1	-6.99	1.36E-11	1.01E-05
17	43660599	SVA_umary_SVA_704	LRRC37A	6.50	2.71E-10	1.83E-04
3	154966214	ALU_umary_ALU_2829	LILRA1	6.37	5.94E-10	3.68E-04
6	32657952	ALU_umary_ALU_5075	HLA-DQB1	-6.33	7.45E-10	4.26E-04
13	32328983	SVA_umary_SVA_565	CCDC58	6.29	9.26E-10	4.92E-04
8	5506286	ALU_umary_ALU_6548	CDH23	6.10	2.82E-09	1.40E-03
6	33030313	SVA_umary_SVA_282	HLA-DPA1	-6.02	4.40E-09	2.04E-03
2	65163783	ALU_umary_ALU_1256	SLC1A4	-5.85	1.11E-08	4.87E-03
8	5506286	ALU_umary_ALU_6548	TMEM132B	5.82	1.33E-08	5.50E-03
8	5506286	ALU_umary_ALU_6548	RP11-492E3.2	5.80	1.46E-08	5.71E-03
8	5506286	ALU_umary_ALU_6548	APOC2	5.77	1.76E-08	6.53E-03
7	123194557	ALU_umary_ALU_6352	VARS	5.74	2.03E-08	7.17E-03
11	86195336	ALU_umary_ALU_8700	IGLV7-46	5.72	2.21E-08	7.48E-03
8	5506286	ALU_umary_ALU_6548	CD2	5.70	2.56E-08	8.27E-03
2	227473038	ALU_umary_ALU_2029	IGLV3-27	5.68	2.81E-08	8.69E-03
2	114106446	ALU_umary_ALU_1441	ACY1	5.66	3.07E-08	9.14E-03
14	106756949	ALU_umary_ALU_10537	NPR3	5.57	5.08E-08	1.42E-02
20	9477936	L1_umary_LINE1_2901	FAM110C	5.54	5.78E-08	1.42E-02
14	106756949	ALU_umary_ALU_10537	AC026703.1	5.54	5.79E-08	1.42E-02

14	106756949	ALU_umary_ALU_10537	PRODH	5.54	5.90E-08	1.42E-02
4	106636862	ALU_umary_ALU_3576	PIK3CA	5.54	5.96E-08	1.42E-02
6	32589834	ALU_umary_ALU_5072	HLA-DRB6	-5.54	5.98E-08	1.42E-02
8	5506286	ALU_umary_ALU_6548	FBXO27	5.53	6.12E-08	1.42E-02
6	32657952	ALU_umary_ALU_5075	HLA-DRB1	-5.53	6.38E-08	1.44E-02
10	105817214	ALU_umary_ALU_8203	RP1-37N7.3	-5.49	7.59E-08	1.66E-02
11	100781393	ALU_umary_ALU_8788	HMGN2P19	5.48	8.25E-08	1.75E-02
14	106756949	ALU_umary_ALU_10537	CLSTN2	5.40	1.21E-07	2.50E-02
10	17712792	SVA_umary_SVA_438	TMEM236	5.39	1.30E-07	2.62E-02
1	65346960	ALU_umary_ALU_212	VANGL2	5.38	1.39E-07	2.72E-02
1	97717644	ALU_umary_ALU_379	IGLV1-50	5.36	1.53E-07	2.91E-02
9	112908065	ALU_umary_ALU_7651	AC020571.3	5.33	1.74E-07	3.23E-02
14	106756949	ALU_umary_ALU_10537	CDH2	5.27	2.32E-07	4.21E-02
6	32657952	ALU_umary_ALU_5075	HLA-DRB5	-5.25	2.57E-07	4.54E-02

PUBLICATIONS

1. **Wang, L.**, Wang, J., Mariño-Ramírez, L., and Jordan, I. K. Genome-wide screen for genetic modifiers of human LINE-1 expression. *In Prep*.
2. **Wang, L.** and Jordan, I. K. Transposable element activity, genome regulation and human health. Current Opinion in Genetics & Development. *In Review*.
3. **Wang, L.**, Norris, E.T., Jordan, I.K., 2017. Human retrotransposon insertion polymorphisms are associated with health and disease via gene regulatory phenotypes. Frontiers in Microbiology. 8: 1418.
4. **Wang, L.**, Rishishwar, L., Mariño-Ramírez, L., and Jordan, I. K., 2016. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. Nucleic acids research.

5. Norris, E.T., Rishishwar, L., **Wang, L.**, Conley, A.B., Chande, A.T., Dabrowski, A.M., Valderrama-Aguirre, A. and Jordan, I.K., 2017. Assortative mating on ancestry-variant traits in admixed Latin American populations. bioRxiv, p.177634.
6. Rishishwar, L., **Wang, L.**, Wang, J., Yi, S.V., Lachance, J. Jordan, I.K. Positive selection on recent human transposable element insertions. *In Review*.
7. Tollervey, J., Liu, T., **Wang, L.**, Lopez, M. F., Jordan, I. K., Lunyak, V. V. Novel H1.0 Lysine 180 dimethylation is a hallmark of cellular senescence in human aging and neurodegenerative disease. *In Prep*.
8. Gaur, M., **Wang, L.**, Amaro-Ortiz, A., Dobke, M., Jordan, I.K., Lunyak, V.V., 2017. Acute genotoxic stress-induced senescence in human mesenchymal cells

drives a unique composition of senescence messaging secretome (SMS). Journal of Stem Cell Research & Therapy. 7: 396.

9. Lopez, M.F., Niu, P., **Wang, L.**, Vogelsang, M., Gaur, M., Krastins, B., Zhao, Y., Smagul, A., Nussupbekova, A., Akanov, A.A., Jordan, I.K., and Lunyak, V.V., 2017. Opposing activities of oncogenic MIR17HG and tumor suppressive MIR100HG clusters and their gene targets regulate replicative senescence in human adult stem cells. NPJ Aging Mech Dis. 3: 7.
10. Rishishwar, L., **Wang, L.**, Clayton, E.A., Mariño-Ramírez, L., McDonald, J.F. and Jordan I.K., 2017. Population and clinical genetics of human transposable elements in the (post) genomic era. Mob Genet Elements. 7: 1-20.
11. Clayton, E.A., **Wang, L.**, Rishishwar, L., Wang, J., McDonald, J.F. and Jordan I.K., 2016. Patterns of transposable element expression and insertion in cancer. Front. Mol. Biosci. 3: 76.
12. Niu, P., Smagul, A., **Wang, L.**, Sadvakas, A., Sha, Y., Pérez, L. M., ... & Lunyak, V. V. (2015). Transcriptional profiling of interleukin-2-primed human adipose derived mesenchymal stem cells revealed dramatic changes in stem cells response imposed by replicative senescence. Oncotarget, 5.
13. Rishishwar, L., Conley, A. B., Wigington, C. H., **Wang, L.**, Valderrama-Aguirre, A., & Jordan, I. K. (2015). Ancestry, admixture and fitness in Colombian genomes. Scientific reports, 5.
14. Castillo, A., **Wang, L.**, Koriyama, C., Eizuru, Y., Jordan, I.K., and Akiba, S., 2014. A systems biology analysis of the changes in gene expression via silencing of HPV-18 E1 expression in HeLa cells. Open Biol. 4: 130119

15. Lopes, F.R., Jjingo, D., da Silva, C.R.M., Andrade, A.C., Marraccini, P., Teixeira, J.B., Carazzolle, M.F., Pereira, G.A.G., Pereira, L.P.F., Vanzela, A.L.L., **Wang, L.**, Jordan, I.K., and Carareto, C.M.A., 2013. Transcriptional activity, chromosomal distribution and expression effects of transposable elements in Coffea genomes. PLoS ONE 8: e78931

REFERENCES

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921. doi: 10.1038/35057062. PubMed PMID: 11237011.

2. de Koning APJ, Gu WJ, Castoe TA, Batzer MA, Pollock DD. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. Plos Genet. 2011;7(12). doi: ARTN e100238410.1371/journal.pgen.1002384. PubMed PMID: WOS:000299167900005.

3. Batzer MA, Deininger PL. A human-specific subfamily of Alu sequences. Genomics. 1991;9(3):481-7. PubMed PMID: 1851725.

4. Batzer MA, Gudi VA, Mena JC, Foltz DW, Herrera RJ, Deininger PL. Amplification dynamics of human-specific (HS) Alu family members. Nucleic Acids Res. 1991;19(13):3619-23. PubMed PMID: 1649453; PubMed Central PMCID: PMC328388.

5. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, et al. Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A. 2003;100(9):5280-5. doi: 10.1073/pnas.0831042100. PubMed PMID: 12682288; PubMed Central PMCID: PMC154336.

6. Kazazian HH, Jr., Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature. 1988;332(6160):164-6. doi: 10.1038/332164a0. PubMed PMID: 2831458.

7. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH, Jr. SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet. 2003;73(6):1444-51. doi: 10.1086/380207. PubMed PMID: 14628287; PubMed Central PMCID: PMC1180407.

8. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, et al. SVA elements: a hominid-specific retroposon family. Journal of molecular biology. 2005;354(4):994-1007. doi: 10.1016/j.jmb.2005.09.085. PubMed PMID: 16288912.

9. Kazazian HH. Mobile DNA: finding treasure in junk: FT Press; 2011.

10. Yang J, Malik HS, Eickbush TH. Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. Proceedings of the National Academy of Sciences of the United States of America. 1999;96(14):7847-52. Epub 1999/07/08. PubMed PMID: 10393910; PubMed Central PMCID: PMCPMC22150.

11. Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell. 1996;87(5):905-16. Epub 1996/11/29. PubMed PMID: 8945517.

12. Luan DD, Korman MH, Jakubczak JL, Eickbush TH. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell. 1993;72(4):595-605. Epub 1993/02/26. PubMed PMID: 7679954.

13. Jurka J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proceedings of the National Academy of Sciences of the United States of America. 1997;94(5):1872-7. Epub 1997/03/04. PubMed PMID: 9050872; PubMed Central PMCID: PMCPMC20010.

14. Okada N, Hamada M, Ogiwara I, Ohshima K. SINEs and LINEs share common 3' sequences: a review. Gene. 1997;205(1-2):229-43. Epub 1998/02/14. PubMed PMID: 9461397.

15. Mc CB. The origin and behavior of mutable loci in maize. Proceedings of the National Academy of Sciences of the United States of America. 1950;36(6):344-55. Epub 1950/06/01. PubMed PMID: 15430309; PubMed Central PMCID: PMCPMC1063197.

16. Kazazian HH, Jr. Mobile elements: drivers of genome evolution. Science. 2004;303(5664):1626-32. Epub 2004/03/16. doi: 10.1126/science.1089670. PubMed PMID: 15016989.

17. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008;9(5):397-405. doi: 10.1038/nrg2337. PubMed PMID: 18368054; PubMed Central PMCID: PMC2596197.

18. Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annual review of genetics. 2012;46:21-42. doi: 10.1146/annurev-genet-110711-155621. PubMed PMID: 22905872.

19. Conley AB, Piriyapongsa J, Jordan IK. Retroviral promoters in the human genome. Bioinformatics. 2008;24(14):1563-7. doi: 10.1093/bioinformatics/btn243. PubMed PMID: 18535086.

20. Jordan IK, Rogozin IB, Glazko GV, Koonin EV. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends in Genetics. 2003;19(2):68-72. doi: Pii S0168-9525(02)00006-9

Doi 10.1016/S0168-9525(02)00006-9. PubMed PMID: WOS:000180845900005.

21. Marino-Ramirez L, Lewis KC, Landsman D, Jordan IK. Transposable elements donate lineage-specific regulatory sequences to host genomes. Cytogenet Genome Res. 2005;110(1-4):333-41. doi: 10.1159/000084965. PubMed PMID: WOS:000231064600033.

22. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 2006;441(7089):87-90. doi: 10.1038/nature04696. PubMed PMID: 16625209.

23. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351(6277):1083-7. doi: 10.1126/science.aad5497. PubMed PMID: 26941318.

24. Chuong EB, Rumi MA, Soares MJ, Baker JC. Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat Genet. 2013;45(3):325-9. doi: 10.1038/ng.2553. PubMed PMID: 23396136; PubMed Central PMCID: PMC3789077.

25. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010;42(7):631-4. doi: 10.1038/ng.600. PubMed PMID: 20526341.

26. Notwell JH, Chung T, Heavner W, Bejerano G. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. Nat Commun. 2015;6:6644. doi: 10.1038/ncomms7644. PubMed PMID: 25806706; PubMed Central PMCID: PMC4438107.

27. Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, et al. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev. 2007;21(6):708-18. Epub 2007/03/21. doi: 10.1101/gad.1525507. PubMed PMID: 17369403; PubMed Central PMCID: PMCPMC1820944.

28. Conley AB, Jordan IK. Cell type-specific termination of transcription by transposable element sequences. Mobile DNA. 2012;3(1):15. doi: 10.1186/1759-8753-3-15. PubMed PMID: 23020800; PubMed Central PMCID: PMC3517506.

29. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo XY, Ramsay L, Bourque G, et al. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. Plos Genetics. 2013;9(4). doi: ARTN e100347010.1371/journal.pgen.1003470. PubMed PMID: WOS:000318073300063.

30. Piriyapongsa J, Marino-Ramirez L, Jordan IK. Origin and evolution of human microRNAs from transposable elements. Genetics. 2007;176(2):1323-37. doi: 10.1534/genetics.107.072553. PubMed PMID: WOS:000247870700047.

31. Weber MJ. Mammalian small nucleolar RNAs are mobile genetic elements. Plos Genet. 2006;2(12):1984-97. doi: ARTN e20510.1371/journal.pgen.0020205. PubMed PMID: WOS:000243482100005.

32. Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A, et al. MIR retrotransposon sequences provide insulators to the human genome. P Natl Acad Sci USA. 2015;112(32):E4428-E37. doi: 10.1073/pnas.1507253112. PubMed PMID: WOS:000359285100014.

33. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, et al. 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. 2007;104(47):18613-8. Epub 2007/11/16. doi: 10.1073/pnas.0703637104. PubMed PMID: 18003932; PubMed Central PMCID: PMCPMC2141825.

34. Hancks DC, Kazazian HH. Active human retrotransposons: variation and disease. Curr Opin Genet Dev. 2012;22(3):191-203. doi: 10.1016/j.gde.2012.02.006. PubMed PMID: WOS:000306160500002.

35. Solyom S, Kazazian HH. Mobile elements in the human genome: implications for disease. Genome Med. 2012;4. PubMed PMID: WOS:000314566100001.

36. Belancio VP, Roy-Engel AM, Deininger PL. All y'all need to know 'bout retroelements in cancer. Semin Cancer Biol. 2010;20(4):200-10. doi: 10.1016/j.semcancer.2010.06.001. PubMed PMID: 20600922; PubMed Central PMCID: PMCPMC2943028.

37. Carreira PE, Richardson SR, Faulkner GJ. L1 retrotransposons, cancer stem cells and oncogenesis. The FEBS journal. 2014;281(1):63-73. doi: 10.1111/febs.12601. PubMed PMID: 24286172; PubMed Central PMCID: PMC4160015.

38. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007;8(4):272-85. doi: 10.1038/nrg2072. PubMed PMID: WOS:000245036100012.

39. Walsh CP, Chaillet JR, Bestor TH. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet. 1998;20(2):116-7. Epub 1998/10/15. doi: 10.1038/2413. PubMed PMID: 9771701.

40. Alves G, Tatro A, Fanning T. Differential methylation of human LINE-1 retrotransposons in malignant cells. Gene. 1996;176(1-2):39-44. Epub 1996/10/17. PubMed PMID: 8918229.

41. Kitkumthorn N, Tuangsintanakul T, Rattanatanyong P, Tiwawech D, Mutirangura A. LINE-1 methylation in the peripheral blood mononuclear cells of cancer patients. Clin Chim Acta. 2012;413(9-10):869-74. Epub 2012/02/14. doi: 10.1016/j.cca.2012.01.024. PubMed PMID: 22326975.

42. Park SY, Seo AN, Jung HY, Gwak JM, Jung N, Cho NY, et al. Alu and LINE-1 hypomethylation is associated with HER2 enriched subtype of breast cancer. PLoS One. 2014;9(6):e100429. Epub 2014/06/28. doi: 10.1371/journal.pone.0100429. PubMed PMID: 24971511; PubMed Central PMCID: PMCPMC4074093.

43. Leung DC, Lorincz MC. Silencing of endogenous retroviruses: when and why do histone marks predominate? Trends Biochem Sci. 2012;37(4):127-33. Epub 2011/12/20. doi: 10.1016/j.tibs.2011.11.006. PubMed PMID: 22178137.

44. Martens JH, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, Steinlein P, et al. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. EMBO J. 2005;24(4):800-12. Epub 2005/01/29. doi: 10.1038/sj.emboj.7600545. PubMed PMID: 15678104; PubMed Central PMCID: PMCPMC549616.

45. Kondo Y, Issa JP. Enrichment for histone H3 lysine 9 methylation at Alu repeats in human cells. J Biol Chem. 2003;278(30):27658-62. Epub 2003/05/02. doi: 10.1074/jbc.M304072200. PubMed PMID: 12724318.

46. Goodier JL, Ostertag EM, Engleka KA, Seleme MC, Kazazian HH, Jr. A potential role for the nucleolus in L1 retrotransposition. Hum Mol Genet. 2004;13(10):1041-8. Epub 2004/03/19. doi: 10.1093/hmg/ddh118. PubMed PMID: 15028673.

47. Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, et al. Role of transposable elements in heterochromatin and epigenetic control. Nature. 2004;430(6998):471-6. Epub 2004/07/23. doi: 10.1038/nature02651. PubMed PMID: 15269773.

48. Sijen T, Plasterk RH. Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. Nature. 2003;426(6964):310-4. Epub 2003/11/25. doi: 10.1038/nature02107. PubMed PMID: 14628056.

49. Yang N, Kazazian HH, Jr. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nature structural & molecular biology. 2006;13(9):763-71. Epub 2006/08/29. doi: 10.1038/nsmb1141. PubMed PMID: 16936727.

50. Peddigari S, Li PW, Rabe JL, Martin SL. hnRNPL and nucleolin bind LINE-1 RNA and function as host factors to modulate retrotransposition. Nucleic Acids Res. 2013;41(1):575-85. Epub 2012/11/20. doi: 10.1093/nar/gks1075. PubMed PMID: 23161687; PubMed Central PMCID: PMCPMC3592465.

51. deHaro D, Kines KJ, Sokolowski M, Dauchy RT, Streva VA, Hill SM, et al. Regulation of L1 expression and retrotransposition by melatonin and its receptor: implications for cancer risk associated with light exposure at night. Nucleic Acids Res. 2014;42(12):7694-707. Epub 2014/06/11. doi: 10.1093/nar/gku503. PubMed PMID: 24914052; PubMed Central PMCID: PMCPMC4081101.

52. Goodier JL, Cheung LE, Kazazian HH, Jr. MOV10 RNA helicase is a potent inhibitor of retrotransposition in cells. PLoS genetics. 2012;8(10):e1002941. Epub 2012/10/25. doi: 10.1371/journal.pgen.1002941. PubMed PMID: 23093941; PubMed Central PMCID: PMCPMC3475670.

53. Guo H, Chitiprolu M, Gagnon D, Meng L, Perez-Iratxeta C, Lagace D, et al. Autophagy supports genomic stability by degrading retrotransposon RNA. Nature communications. 2014;5:5276. Epub 2014/11/05. doi: 10.1038/ncomms6276. PubMed PMID: 25366815.

54. Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(16):E2326-34. doi: 10.1073/pnas.1602336113. PubMed PMID: 27001843; PubMed Central PMCID: PMCPMC4843416.

55. Swergold GD. Identification, characterization, and cell specificity of a human LINE-1 promoter. Mol Cell Biol. 1990;10(12):6718-29. PubMed PMID: 1701022; PubMed Central PMCID: PMCPMC362950.

56. Babushok DV, Kazazian HH, Jr. Progress in understanding the biology of the human mutagen LINE-1. Hum Mutat. 2007;28(6):527-39. Epub 2007/02/20. doi: 10.1002/humu.20486. PubMed PMID: 17309057.

57. Penzkofer T, Dandekar T, Zemojtel T. L1Base: from functional annotation to prediction of active LINE-1 elements. Nucleic Acids Res. 2005;33(Database issue):D498-500. Epub 2004/12/21. doi: 10.1093/nar/gki044. PubMed PMID: 15608246; PubMed Central PMCID: PMCPMC539998.

58. Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet. 2002;3(5):370-9. doi: 10.1038/nrg798. PubMed PMID: 11988762.

59. Shaikh TH, Roy AM, Kim J, Batzer MA, Deininger PL. cDNAs derived from primary and small cytoplasmic Alu (scAlu) transcripts. Journal of molecular biology. 1997;271(2):222-34. Epub 1997/08/15. doi: 10.1006/jmbi.1997.1161. PubMed PMID: 9268654.

60. Cordaux R, Hedges DJ, Herke SW, Batzer MA. Estimating the retrotransposition rate of human Alu elements. Gene. 2006;373:134-7. Epub 2006/03/09. doi: 10.1016/j.gene.2006.01.019. PubMed PMID: 16522357.

61. Batzer MA, Rubin CM, Hellmann-Blumberg U, Alegria-Hartman M, Leeflang EP, Stern JD, et al. Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. Journal of molecular biology. 1995;247(3):418-27. Epub 1995/03/31. doi: 10.1006/jmbi.1994.0150. PubMed PMID: 7714898.

62. Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shaikh TH, et al. African origin of human-specific polymorphic Alu insertions. Proceedings of the National Academy of Sciences of the United States of America. 1994;91(25):12288-92. Epub 1994/12/06. PubMed PMID: 7991620; PubMed Central PMCID: PMCPMC45422.

63. Stoneking M, Fontius JJ, Clifford SL, Soodyall H, Arcot SS, Saha N, et al. Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. Genome research. 1997;7(11):1061-71. Epub 1998/01/10. PubMed PMID: 9371742; PubMed Central PMCID: PMCPMC310683.

64. Rishishwar L, Marino-Ramirez L, Jordan IK. Benchmarking computational tools for polymorphic transposable element detection. Brief Bioinform. 2016. doi: 10.1093/bib/bbw072. PubMed PMID: 27524380.

65. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75-81. doi: 10.1038/nature15394. PubMed PMID: WOS:000362095100037.

66. Hancks DC, Kazazian HH, Jr. SVA retrotransposons: Evolution and genetic instability. Semin Cancer Biol. 2010;20(4):234-45. doi: 10.1016/j.semcancer.2010.04.001. PubMed PMID: 20416380; PubMed Central PMCID: PMCPMC2945828.

67. Hancks DC, Kazazian HH, Jr. Roles for retrotransposon insertions in human disease. Mobile DNA. 2016;7:9. doi: 10.1186/s13100-016-0065-9. PubMed PMID: 27158268; PubMed Central PMCID: PMCPMC4859970.

68. Meischl C, Boer M, Ahlin A, Roos D. A new exon created by intronic insertion of a rearranged LINE-1 element as the cause of chronic granulomatous disease. Eur J Hum Genet. 2000;8(9):697-703. Epub 2000/09/12. doi: 10.1038/sj.ejhg.5200523. PubMed PMID: 10980575.

69. Musova Z, Hedvicakova P, Mohrmann M, Tesarova M, Krepelova A, Zeman J, et al. A novel insertion of a rearranged L1 element in exon 44 of the dystrophin gene: further evidence for possible bias in retroposon integration. Biochem Biophys Res Commun. 2006;347(1):145-9. Epub 2006/07/01. doi: 10.1016/j.bbrc.2006.06.071. PubMed PMID: 16808900.

70. Li X, Scaringe WA, Hill KA, Roberts S, Mengos A, Careri D, et al. Frequency of recent retrotransposition events in the human factor IX gene. Hum Mutat. 2001;17(6):511-9. Epub 2001/06/01. doi: 10.1002/humu.1134. PubMed PMID: 11385709.

71. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, 3rd, et al. Landscape of somatic retrotransposition in human cancers. Science. 2012;337(6097):967-71. doi: 10.1126/science.1222077. PubMed PMID: 22745252; PubMed Central PMCID: PMCPMC3656569.

72. Solyom S, Ewing AD, Rahrmann EP, Doucet T, Nelson HH, Burns MB, et al. Extensive somatic L1 retrotransposition in colorectal tumors. Genome Res. 2012;22(12):2328-38. doi: 10.1101/gr.145235.112. PubMed PMID: 22968929; PubMed Central PMCID: PMCPMC3514663.

73. Tubio JMC, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, et al. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science. 2014;345(6196):1251343. doi: 10.1126/science.1251343. PubMed PMID: 25082706; PubMed Central PMCID: PMCPMC4380235.

74. Helman E, Lawrence MS, Stewart C, Sougnez C, Getz G, Meyerson M. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. Genome research. 2014;24(7):1053-63. doi: 10.1101/gr.163659.113. PubMed PMID: 24823667; PubMed Central PMCID: PMCPMC4079962.

75. Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, et al. Natural mutagenesis of human genomes by endogenous retrotransposons. Cell. 2010;141(7):1253-61. Epub 2010/07/07. doi: 10.1016/j.cell.2010.05.020. PubMed PMID: 20603005; PubMed Central PMCID: PMCPMC2943760.

76. Kemp JR, Longworth MS. Crossing the LINE Toward Genomic Instability: LINE-1 Retrotransposition in Cancer. Front Chem. 2015;3:68. Epub 2016/01/07. doi: 10.3389/fchem.2015.00068. PubMed PMID: 26734601; PubMed Central PMCID: PMCPMC4679865.

77. Rodic N, Burns KH. Long interspersed element-1 (LINE-1): passenger or driver in human neoplasms? PLoS genetics. 2013;9(3):e1003402. Epub 2013/04/05. doi: 10.1371/journal.pgen.1003402. PubMed PMID: 23555307; PubMed Central PMCID: PMCPMC3610623.

78. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007;8(4):272-85. Epub 2007/03/17. doi: 10.1038/nrg2072. PubMed PMID: 17363976.

79. Bourc'his D, Bestor TH. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature. 2004;431(7004):96-9. Epub 2004/08/20. doi: 10.1038/nature02886. PubMed PMID: 15318244.

80. Carmell MA, Girard A, van de Kant HJ, Bourc'his D, Bestor TH, de Rooij DG, et al. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. Dev Cell. 2007;12(4):503-14. Epub 2007/03/31. doi: 10.1016/j.devcel.2007.03.001. PubMed PMID: 17395546.

81. Jones BC, Wood JG, Chang C, Tam AD, Franklin MJ, Siegel ER, et al. A somatic piRNA pathway in the Drosophila fat body ensures metabolic homeostasis and normal lifespan. Nature communications. 2016;7:13856. Epub 2016/12/22. doi: 10.1038/ncomms13856. PubMed PMID: 28000665; PubMed Central PMCID: PMCPMC5187580.

82. Perrat PN, DasGupta S, Wang J, Theurkauf W, Weng Z, Rosbash M, et al. Transposition-driven genomic heterogeneity in the Drosophila brain. Science. 2013;340(6128):91-5. Epub 2013/04/06. doi: 10.1126/science.1231965. PubMed PMID: 23559253; PubMed Central PMCID: PMCPMC3887341.

83. Liang G, Chan MF, Tomigahara Y, Tsai YC, Gonzales FA, Li E, et al. Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. Mol Cell Biol. 2002;22(2):480-91. PubMed PMID: 11756544; PubMed Central PMCID: PMCPMC139739.

84. Van Meter M, Kashyap M, Rezazadeh S, Geneva AJ, Morello TD, Seluanov A, et al. SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. Nature communications. 2014;5:5011. Epub 2014/09/24. doi: 10.1038/ncomms6011. PubMed PMID: 25247314; PubMed Central PMCID: PMCPMC4185372.

85. Muotri AR, Marchetto MC, Coufal NG, Oefner R, Yeo G, Nakashima K, et al. L1 retrotransposition in neurons is modulated by MeCP2. Nature. 2010;468(7322):443-6. Epub 2010/11/19. doi: 10.1038/nature09544. PubMed PMID: 21085180; PubMed Central PMCID: PMCPMC3059197.

86. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature. 2005;435(7044):903-10. Epub 2005/06/17. doi: 10.1038/nature03663. PubMed PMID: 15959507.

87. Arjan-Odedra S, Swanson CM, Sherer NM, Wolinsky SM, Malim MH. Endogenous MOV10 inhibits the retrotransposition of endogenous retroelements but not the replication of exogenous retroviruses. Retrovirology. 2012;9:53. Epub 2012/06/26. doi: 10.1186/1742-4690-9-53. PubMed PMID: 22727223; PubMed Central PMCID: PMCPMC3408377.

88. Li X, Zhang J, Jia R, Cheng V, Xu X, Qiao W, et al. The MOV10 helicase inhibits LINE-1 mobility. J Biol Chem. 2013;288(29):21148-60. Epub 2013/06/12. doi: 10.1074/jbc.M113.465856. PubMed PMID: 23754279; PubMed Central PMCID: PMCPMC3774381.

89. Stetson DB, Ko JS, Heidmann T, Medzhitov R. Trex1 prevents cell-intrinsic initiation of autoimmunity. Cell. 2008;134(4):587-98. Epub 2008/08/30. doi: 10.1016/j.cell.2008.06.032. PubMed PMID: 18724932; PubMed Central PMCID: PMCPMC2626626.

90. Gasior SL, Roy-Engel AM, Deininger PL. ERCC1/XPF limits L1 retrotransposition. DNA Repair (Amst). 2008;7(6):983-9. Epub 2008/04/09. doi: 10.1016/j.dnarep.2008.02.006. PubMed PMID: 18396111; PubMed Central PMCID: PMCPMC2483505.

91. Ostertag EM, Prak ET, DeBerardinis RJ, Moran JV, Kazazian HH, Jr. Determination of L1 retrotransposition kinetics in cultured cells. Nucleic Acids Res. 2000;28(6):1418-23. Epub 2000/02/24. PubMed PMID: 10684937; PubMed Central PMCID: PMCPMC111040.

92. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74. Epub 2015/10/04. doi: 10.1038/nature15393. PubMed PMID: 26432245; PubMed Central PMCID: PMCPMC4750478.

93. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501(7468):506-11. doi: 10.1038/nature12531. PubMed PMID: 24037378; PubMed Central PMCID: PMCPMC3918453.

94. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012;7(3):500-7. Epub 2012/02/22. doi: 10.1038/nprot.2011.457. PubMed PMID: 22343431; PubMed Central PMCID: PMCPMC3398141.

95. Penzkofer T, Jager M, Figlerowicz M, Badge R, Mundlos S, Robinson PN, et al. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. Nucleic Acids Res. 2017;45(D1):D68-D73. doi: 10.1093/nar/gkw925. PubMed PMID: 27924012; PubMed Central PMCID: PMCPMC5210629.

96. Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics. 2015;31(22):3593-9. Epub 2015/07/25. doi: 10.1093/bioinformatics/btv422. PubMed PMID: 26206304; PubMed Central PMCID: PMCPMC4757950.

97. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40(10):4288-97. doi: 10.1093/nar/gks042. PubMed PMID: 22287627; PubMed Central PMCID: PMCPMC3378882.

98. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016;48(1):22-9. doi: 10.1038/ng.3461. PubMed PMID: 26642241; PubMed Central PMCID: PMCPMC4909355.

99. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74. doi: 10.1038/nature15393. PubMed PMID: WOS:000362095100036.

100. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28(10):1353-8. doi: 10.1093/bioinformatics/bts163. PubMed PMID: 22492648; PubMed Central PMCID: PMCPMC3348564.

101. Lindtner S, Felber BK, Kjems J. An element in the 3' untranslated region of human LINE-1 retrotransposon mRNA binds NXF1(TAP) and can function as a nuclear export element. RNA. 2002;8(3):345-56. PubMed PMID: 12003494; PubMed Central PMCID: PMCPMC1370256.

102. Garcia-Perez JL, Morell M, Scheys JO, Kulpa DA, Morell S, Carter CC, et al. Epigenetic silencing of engineered L1 retrotransposition events in human embryonic

carcinoma cells. Nature. 2010;466(7307):769-73. doi: 10.1038/nature09209. PubMed PMID: 20686575; PubMed Central PMCID: PMCPMC3034402.

103. Pavlicek A, Jabbari K, Paces J, Paces V, Hejnar JV, Bernardi G. Similar integration but different stability of Alus and LINEs in the human genome. Gene. 2001;276(1-2):39-45. PubMed PMID: 11591470.

104. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome research. 2014;24(12):1963-76. doi: 10.1101/gr.168872.113. PubMed PMID: 25319995; PubMed Central PMCID: PMC4248313.

105. Jacques PE, Jeyakani J, Bourque G. The majority of primate-specific regulatory sequences are derived from transposable elements. Plos Genet. 2013;9(5):e1003504. doi: 10.1371/journal.pgen.1003504. PubMed PMID: 23675311; PubMed Central PMCID: PMC3649963.

106. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell. 2012;148(1-2):335-48. doi: 10.1016/j.cell.2011.11.058. PubMed PMID: 22244452; PubMed Central PMCID: PMC3368268.

107. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966;53(3-4):325-38.

108. Ihaka R, Gentleman R. R: a language for data analysis and graphics. Journal of computational and graphical statistics. 1996;5(3):299-314.

109. Lappalainen T, Sammeth M, Friedlander MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501(7468):506-11. doi: 10.1038/nature12531. PubMed PMID: WOS:000324826300049.

110. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, et al. Ensembl 2013. Nucleic Acids Res. 2013;41:D48-55. doi: 10.1093/nar/gks1236. PubMed PMID: 23203987; PubMed Central PMCID: PMCPMC3531136.

111. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012;7(3):500-7. doi: 10.1038/nprot.2011.457. PubMed PMID: WOS:000300948500007.

112. t Hoen PA, Friedlander MR, Almlof J, Sammeth M, Pulyakhina I, Anvar SY, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol. 2013;31(11):1015-22. doi: 10.1038/nbt.2702. PubMed PMID: 24037425.

113. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology. 2004;5(10):R80.

114. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28(10):1353-8. doi: 10.1093/bioinformatics/bts163. PubMed PMID: WOS:000304053300009.

115. Yang JA, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011;88(1):76-82. doi: 10.1016/j.ajhg.2010.11.011. PubMed PMID: WOS:000286501500007.

116. Diabetes Genetics Initiative of Broad Institute of H, Mit LU, Novartis Institutes of BioMedical R, Saxena R, Voight BF, Lyssenko V, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science. 2007;316(5829):1331-6. doi: 10.1126/science.1142358. PubMed PMID: 17463246.

117. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. P Natl Acad Sci USA. 2005;102(43):15545-50. doi: 10.1073/pnas.0506580102. PubMed PMID: WOS:000232929400051.

118. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006;34(Database issue):D108-10. doi: 10.1093/nar/gkj143. PubMed PMID: 16381825; PubMed Central PMCID: PMC1347505.

119. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome research. 2009;19(9):1639-45. doi: 10.1101/gr.092759.109. PubMed PMID: 19541911; PubMed Central PMCID: PMC2752132.

120. Ewing AD. Transposable element detection from whole genome sequence data. Mobile DNA. 2015;6:24. doi: 10.1186/s13100-015-0055-3. PubMed PMID: 26719777; PubMed Central PMCID: PMC4696183.

121. Rishishwar L, Tellez Villa CE, Jordan IK. Transposable element polymorphisms recapitulate human evolution. Mobile DNA. 2015;6:21. doi: 10.1186/s13100-015-0052-6. PubMed PMID: 26579215; PubMed Central PMCID: PMC4647816.

122. Hayden MS, Ghosh S. NF-kappa B, the first quarter-century: remarkable progress and outstanding questions. Gene Dev. 2012;26(3):203-34. doi: 10.1101/gad.183434.111. PubMed PMID: WOS:000300125700001.

123. Martin-Subero JI, Gesk S, Harder L, Sonoki T, Tucker PW, Schlegelberger B, et al. Recurrent involvement of the REL and BCL11A loci in classical Hodgkin lymphoma. Blood. 2002;99(4):1474-7. PubMed PMID: 11830502.

124. McGovern DP, Gardet A, Torkvist L, Goyette P, Essers J, Taylor KD, et al. Genome-wide association identifies multiple ulcerative colitis susceptibility loci. Nat Genet. 2010;42(4):332-7. doi: 10.1038/ng.549. PubMed PMID: 20228799; PubMed Central PMCID: PMCPMC3087600.

125. Gregersen PK, Amos CI, Lee AT, Lu Y, Remmers EF, Kastner DL, et al. REL, encoding a member of the NF-kappa B family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. Nat Genet. 2009;41(7):820-U77. doi: 10.1038/ng.395. PubMed PMID: WOS:000267786200013.

126. Cobaleda C, Schebesta A, Delogu A, Busslinger M. Pax5: the guardian of B cell identity and function. Nature immunology. 2007;8(5):463-70. doi: 10.1038/ni1454. PubMed PMID: 17440452.

127. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648-60. doi: 10.1126/science.1262110. PubMed PMID: 25954001; PubMed Central PMCID: PMC4547484.

128. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017;18(2):71-86. doi: 10.1038/nrg.2016.139. PubMed PMID: 27867194; PubMed Central PMCID: PMCPMC5498291.

129. Clayton EA, Wang L, Rishishwar L, Wang J, McDonald JF, Jordan IK. Patterns of Transposable Element Expression and Insertion in Cancer. Front Mol Biosci. 2016;3:76. doi: 10.3389/fmolb.2016.00076. PubMed PMID: 27900322; PubMed Central PMCID: PMCPMC5110550.

130. Wang L, Rishishwar L, Marino-Ramirez L, Jordan IK. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. Nucleic Acids Res. 2016. doi: 10.1093/nar/gkw1286. PubMed PMID: 27998931.

131. Rishishwar L, Wang L, Clayton EA, Marino-Ramirez L, McDonald JF, Jordan IK. Population and clinical genetics of human transposable elements in the (post) genomic era. Mob Genet Elements. 2017;7(1):1-20. doi: 10.1080/2159256X.2017.1280116. PubMed PMID: 28228978.

132. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896-D901. doi: 10.1093/nar/gkw1133. PubMed PMID: 27899670; PubMed Central PMCID: PMCPMC5210590.

133. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9(3):215-6. doi: 10.1038/nmeth.1906. PubMed PMID: 22373907; PubMed Central PMCID: PMCPMC3577932.

134. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature.

2015;518(7539):317-30. doi: 10.1038/nature14248. PubMed PMID: 25693563; PubMed Central PMCID: PMCPMC4530010.

135. Xie X, Ma W, Songyang Z, Luo Z, Huang J, Dai Z, et al. CCSI: a database providing chromatin-chromatin spatial interaction information. Database (Oxford). 2016;2016. doi: 10.1093/database/bav124. PubMed PMID: 26868054; PubMed Central PMCID: PMCPMC4750547.

136. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348(6235):660-5. doi: 10.1126/science.aaa0355. PubMed PMID: 25954002; PubMed Central PMCID: PMCPMC4547472.

137. Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, et al. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science. 2015;348(6235):666-9. doi: 10.1126/science.1261877. PubMed PMID: 25954003; PubMed Central PMCID: PMCPMC4537935.

138. Anthony RM, Ravetch JV. A novel role for the IgG Fc glycan: the anti-inflammatory activity of sialylated IgG Fcs. J Clin Immunol. 2010;30 Suppl 1:S9-14. doi: 10.1007/s10875-010-9405-6. PubMed PMID: 20480216.

139. Kaneko Y, Nimmerjahn F, Ravetch JV. Anti-inflammatory activity of immunoglobulin G resulting from Fc sialylation. Science. 2006;313(5787):670-3. doi: 10.1126/science.1129594. PubMed PMID: 16888140.

140. Lauc G, Huffman JE, Pucic M, Zgaga L, Adamczyk B, Muzinic A, et al. Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. Plos Genet. 2013;9(1):e1003225. doi: 10.1371/journal.pgen.1003225. PubMed PMID: 23382691; PubMed Central PMCID: PMCPMC3561084.

141. Ackermann AM, Wang Z, Schug J, Naji A, Kaestner KH. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. Mol Metab. 2016;5(3):233-44. doi: 10.1016/j.molmet.2016.01.002. PubMed PMID: 26977395; PubMed Central PMCID: PMCPMC4770267.

142. Wang YJ, Schug J, Won KJ, Liu C, Naji A, Avrahami D, et al. Single-Cell Transcriptomics of the Human Endocrine Pancreas. Diabetes. 2016;65(10):3028-38. doi: 10.2337/db16-0405. PubMed PMID: 27364731; PubMed Central PMCID: PMCPMC5033269.

143. Quesada I, Tuduri E, Ripoll C, Nadal A. Physiology of the pancreatic alpha-cell and glucagon secretion: role in glucose homeostasis and diabetes. J Endocrinol. 2008;199(1):5-19. doi: 10.1677/JOE-08-0290. PubMed PMID: 18669612.

144. Palmer ND, Goodarzi MO, Langefeld CD, Wang N, Guo X, Taylor KD, et al. Genetic Variants Associated With Quantitative Glucose Homeostasis Traits Translate to

Type 2 Diabetes in Mexican Americans: The GUARDIAN (Genetics Underlying Diabetes in Hispanics) Consortium. Diabetes. 2015;64(5):1853-66. doi: 10.2337/db14-0732. PubMed PMID: 25524916; PubMed Central PMCID: PMCPMC4407862.

145. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. Nature. 1980;284(5757):601-3. PubMed PMID: 6245369.

146. Orgel LE, Crick FH. Selfish DNA: the ultimate parasite. Nature. 1980;284(5757):604-7. PubMed PMID: 7366731.

147. Gould SJ, Vrba ES. Exaptation—a missing term in the science of form. Paleobiology. 1982;8(1):4-15.

148. Miller WJ, Hagemann S, Reiter E, Pinsker W. P-element homologous sequences are tandemly repeated in the genome of Drosophila guanche. Proceedings of the National Academy of Sciences of the United States of America. 1992;89(9):4018-22. PubMed PMID: 1315047; PubMed Central PMCID: PMCPMC525623.

149. Jiang C, Chen C, Huang Z, Liu R, Verdier J. ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. BMC Bioinformatics. 2015;16:72. Epub 2015/04/19. doi: 10.1186/s12859-015-0507-2. PubMed PMID: 25887332; PubMed Central PMCID: PMCPMC4351942.

150. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. Genome research. 2017. Epub 2017/09/01. doi: 10.1101/gr.218032.116. PubMed PMID: 28855259.

151. Thung DT, de Ligt J, Vissers LE, Steehouwer M, Kroon M, de Vries P, et al. Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol. 2014;15(10):488. Epub 2014/10/29. doi: 10.1186/s13059-014-0488-x. PubMed PMID: 25348035; PubMed Central PMCID: PMCPMC4228151.

152. Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics. 2013;29(3):389-90. Epub 2012/12/13. doi: 10.1093/bioinformatics/bts697. PubMed PMID: 23233656; PubMed Central PMCID: PMCPMC3562067.

153. Wu J, Lee WP, Ward A, Walker JA, Konkel MK, Batzer MA, et al. Tangram: a comprehensive toolbox for mobile element insertion detection. BMC Genomics. 2014;15:795. Epub 2014/09/18. doi: 10.1186/1471-2164-15-795. PubMed PMID: 25228379; PubMed Central PMCID: PMCPMC4180832.

154. Zhuang J, Wang J, Theurkauf W, Weng Z. TEMP: a computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Res. 2014;42(11):6826-38. Epub 2014/04/23. doi: 10.1093/nar/gku323. PubMed PMID: 24753423; PubMed Central PMCID: PMCPMC4066757.

155. Fiston-Lavier AS, Barron MG, Petrov DA, Gonzalez J. T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. Nucleic Acids Res. 2015;43(4):e22. Epub 2014/12/17. doi: 10.1093/nar/gku1250. PubMed PMID: 25510498; PubMed Central PMCID: PMCPMC4344482.

156. Santander CG, Gambron P, Marchi E, Karamitros T, Katzourakis A, Magiorkinis G. STEAK: A specific tool for transposable elements and retrovirus detection in high-throughput sequencing data. Virus Evol. 2017;3(2):vex023. Epub 2017/09/28. doi: 10.1093/ve/vex023. PubMed PMID: 28948042; PubMed Central PMCID: PMCPMC5597868.

157. Disdero E, Filee J. LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. Mobile DNA. 2017;8:5. Epub 2017/04/14. doi: 10.1186/s13100-017-0088-x. PubMed PMID: 28405230; PubMed Central PMCID: PMCPMC5385071.

158. Witherspoon DJ, Xing J, Zhang Y, Watkins WS, Batzer MA, Jorde LB. Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. BMC Genomics. 2010;11:410. Epub 2010/07/02. doi: 10.1186/1471-2164-11-410. PubMed PMID: 20591181; PubMed Central PMCID: PMCPMC2996938.

159. Ewing AD, Kazazian HH, Jr. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome research. 2010;20(9):1262-70. Epub 2010/05/22. doi: 10.1101/gr.106419.110. PubMed PMID: 20488934; PubMed Central PMCID: PMCPMC2928504.

160. Sanchez-Luque FJ, Richardson SR, Faulkner GJ. Retrotransposon Capture Sequencing (RC-Seq): A Targeted, High-Throughput Approach to Resolve Somatic L1 Retrotransposition in Humans. Methods Mol Biol. 2016;1400:47-77. Epub 2016/02/20. doi: 10.1007/978-1-4939-3372-3_4. PubMed PMID: 26895046.

161. Tang Z, Steranka JP, Ma S, Grivainis M, Rodic N, Huang CR, et al. Human transposon insertion profiling: Analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer. Proceedings of the National Academy of Sciences of the United States of America. 2017;114(5):E733-E40. Epub 2017/01/18. doi: 10.1073/pnas.1619797114. PubMed PMID: 28096347; PubMed Central PMCID: PMCPMC5293032.

162. Burns KH. Transposable elements in cancer. Nat Rev Cancer. 2017;17(7):415-24. doi: 10.1038/nrc.2017.35. PubMed PMID: 28642606.

163. Wang L, Rishishwar L, Marino-Ramirez L, Jordan IK. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. Nucleic Acids Res. 2017;45(5):2318-28. Epub 2016/12/22. doi: 10.1093/nar/gkw1286. PubMed PMID: 27998931; PubMed Central PMCID: PMCPMC5389732.

164. Gibson G, Powell JE, Marigorta UM. Expression quantitative trait locus analysis for translational medicine. Genome Med. 2015;7(1):60. doi: 10.1186/s13073-015-0186-7. PubMed PMID: 26110023; PubMed Central PMCID: PMCPMC4479075.

165. Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. Elife. 2016;5. Epub 2016/12/03. doi: 10.7554/eLife.20777. PubMed PMID: 27911260; PubMed Central PMCID: PMCPMC5167521.

166. Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, et al. Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS genetics. 2015;11(1):e1004915. Epub 2015/01/09. doi: 10.1371/journal.pgen.1004915. PubMed PMID: 25569788; PubMed Central PMCID: PMCPMC4287451.

167. Payer LM, Steranka JP, Yang WR, Kryatova M, Medabalimi S, Ardeljan D, et al. Structural variants caused by Alu insertions are associated with risks for many human diseases. Proceedings of the National Academy of Sciences of the United States of America. 2017;114(20):E3984-E92. Epub 2017/05/04. doi: 10.1073/pnas.1704117114. PubMed PMID: 28465436; PubMed Central PMCID: PMCPMC5441760.

168. Wang L, Norris ET, Jordan IK. Human Retrotransposon Insertion Polymorphisms Are Associated with Health and Disease via Gene Regulatory Phenotypes. Front Microbiol. 2017;8:1418. doi: 10.3389/fmicb.2017.01418. PubMed PMID: 28824558; PubMed Central PMCID: PMCPMC5539088.