The non-coding genome and its impact on gene expression

LYKOSKOUFIS, Nikolaos Michail Rafail

Abstract
In this thesis, we were interested in better understanding the relationship between the coding and non-coding part of the genome, specifically between non-coding regulatory variants, transposable elements (TEs) and genes. In the first part, we investigated the causal relationship between expression Quantitative Trait loci (eQTL) variants, TEs and genes in colorectal cancer samples and discovered 1,758 TEs impacting the expression of 1,626 genes, specifically in cancer. In the second part of this thesis, we were interested in better characterising the role of transposable elements in the regulation of gene expression in adult somatic tissues. For this, we used the Genotype-Tissue Expression (GTEx) project to understand how TEs mediate genetic effects, altering the expression of nearby genes in 43 different adult tissues. We discovered 28,888 TEs affecting the expression of 12,599 nearby genes and showed that the effects of these TEs on nearby genes is shared among closely related tissues.

Reference

DOI : 10.13097/archive-ouverte/unige:164507
URN : urn:nbn:ch:unige-1645073
La partie non-codante du génome et ses effets sur l'expression des gènes

The non-coding genome and its impact on gene expression

THÈSE

présentée aux Facultés de médecine et des sciences de l’Université de Genève pour obtenir le grade de Docteur ès sciences en sciences de la vie mention Genomics and Digital Health

par

Nikolaos Michail Rafail LYKOSKOUFIS

de
Athènes, Grèce

Thèse N° 186

GENÈVE
Septembre 2022
DOCTORAT ÈS SCIENCES EN SCIENCES DE LA VIE DES FACULTÉS DE MÉDECINE ET DES SCIENCES MENTION GÉNOMIQUE ET SANTÉ DIGITALE

Thèse de M. Nikolaos Michail Rafael LYKOSKOUFIS

intitulée :

« The non-coding genome and its impact on gene expression »

Les Facultés de médecine et des sciences, sur le préavis de Monsieur Simon BRAUN, Professeur assistant et directeur de thèse (Département de Médecine Génétique et Développement), Monsieur Emmanouil DERMITZAKIS, Professeur ordinaire et co-directeur de thèse (Département de Médecine Génétique et Développement), Monsieur Evgeny ZDOBNOV, Professeur ordinaire et Président du jury (Département de Médecine Génétique et Développement), Monsieur Ioannis XENARIOS, Professeur (Centre de Génomique Intégrative, Université de Lausanne (UNIL) Suisse) autorisent l’impression de la présente thèse, sans exprimer d’opinion sur les propositions qui y sont énoncées.

Genève, le 23 septembre 2022

Thèse - 186 -

Le Doyen
Faculté de médecine

Le Doyen
Faculté des sciences

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

FACULTÉ DE MÉDECINE
FACULTÉ DES SCIENCES
Abstract

Since the completion of the sequencing of the human genome in the beginning of the 21st century, technologies have drastically developed. These advancements gave a unique opportunity to the scientific community to understand the relationship between genetic variation and complex human traits. Genome-wide association studies (GWAS) and expression Quantitative Trait Loci (eQTL) studies showed that genetic variants associated with complex traits or with gene expression are both falling in non-coding regulatory regions of the genome which make them very difficult to study. The non-coding genome represents 98% of the human genome and contains a myriad of different molecular elements. Strikingly, more than half of the human genome is made up of repeated DNA sequences called transposable elements (TEs). Although still poorly understood, these interspersed repeat sequences have been shown in multiple studies to play crucial roles in the regulation of gene expression during human development, immune responses and have been linked to many diseases such as cancer. Thus, better understanding their role in the regulation of gene expression in both physiological and disease conditions is important in order to obtain a better understanding how genetic effects impact human traits and diseases, and better comprehend the interplay between the coding and non-coding genome. This thesis is composed of two projects where we tackled on the regulation of transposable elements (TEs) and gene expression by non-coding regulatory variants (eQTLs) and the interplay that exist between these three in colorectal cancer, as well as in healthy adult human tissues.

In the first project we aimed to determine whether transposable elements are under genetic control and if so, how do they mediate genetic effects onto nearby genes in cancer. We analysed 275 normal colon and 276 colorectal cancer (CRC) samples from the System Biology of Colorectal Cancer (SYSCOL) cohort. We discovered 10,111 and 5,152 TE expression Quantitative Trait loci (eQTLs) in normal and tumor samples, respectively. Amongst these, 429 were tumor-specific and driven most likely by changes in levels of DNA methylation. We discovered that transcription factors show higher enrichment for tumor-specific TE-eQTLs compared to shared TE-eQTLs, highlighting that these transcription factors (TFs) could regulate TE expression more specifically than gene expression in cancer. Then, we inferred the causal relationship between eQTL variants, TEs and genes and discovered 1,758 TEs acting as mediators of genetic effects, altering the expression of 1,626 genes in cis significantly more in tumor than normal. Finally, we showed that in some cases, tumor-specific TE-eQTLs are the trigger for TEs to impact the expression of nearby genes. Altogether, the results of this project highlight a new class of cancer drivers, thereby enhancing our understanding of tumorigenesis and revealing new targets for cancer treatment.

In the second project, we wanted to extend our previous work and study the effect TEs have on nearby genes in healthy adult human tissues. We used the Genotype-Tissue
Expression (GTEx) project to study the genetic control of TEs and the extent to which they impact gene expression in 43 different adult tissues. We discovered TE-eQTLs in all tissues, ranging from 1,086 in the hippocampus of the brain, up to 22,634 in the thyroid, indicating that TEs are under strong genetic control. We observed many TFs with stronger enrichments for TE-eQTLs compared to gene-eQTLs, indicating that these TFs are more specifically regulating TE expression than gene expression. Then, we assessed the causal relationship between eQTL variants, TEs and genes and pooling results from all 43 tissues together, we identified 28,888 TEs that mediate genetic effects onto 12,599 nearby genes, impacting their expression. Finally, we showed that the causal effect of TEs on gene expression is shared among closely related tissues and we discovered several cases where TEs impact nearby genes by acting as alternative promoters. Collectively, these results shed further light on the transcriptional activity of TEs in adult human tissues and highlight an important functional role for TEs in the transcriptional regulation of nearby genes. These results serve as a comprehensive atlas of TE-eQTLs that can be used to further study the functional role of TEs in the regulation of gene expression but also to understand the role TEs play in complex traits and diseases.

Together these studies reveal the importance of transposable elements in the regulation of gene expression and further challenge the assumption TEs are inactive in adult tissues. Our findings illustrate how by integrating additional information from different molecular phenotypes (genes, TEs and genetic variants) across individuals, we can better understand the interplay between the coding and non-coding genome.
Résumé

L'aboutissement du séquençage du génome humain et l'avancé technologique ont permis de mieux comprendre la relation entre la variation génétique et les maladies complexes. En effet, de nombreuses études ont démontré que les variants génétiques impactant les maladies complexes résident dans la partie non codante du génome ce qui rend leur interprétation très difficile. La fraction non-codante du génome représente environ 98% du génome humain. Mis à part les variants génétiques, beaucoup d'autres éléments non-codants y résident suscitant un fort intérêt. Un exemple sont les éléments transposables (TEs). Ce sont des séquences DNA répétées représentant plus de la moitié du génome humain dont plusieurs études ont démontré leur implication dans la régulation des gènes durant le développement, les réponses immunitaires mais aussi dans certaines maladies, comme le cancer. Par conséquent, il est d’une importance majeure de comprendre les relations existantes entre la partie non-codante et codante du génome, afin d’obtenir une meilleure compréhension de l’étiologie de nombreuses maladies. Cette thèse est composée de deux projets dans lesquels l’effet de variants génétiques sur l’expression des TEs et des gènes ainsi que la relation entre ces trois éléments a été investiguée.

Le premier projet a comme but de déterminer le contrôle génétique des éléments transposables et de comprendre comment les effets génétiques sont transmis à travers les TEs sur les gènes avoisinants. Ceci a été possible en comparant des 275 tissus de colon sains et 276 tissus tumoraux provenant de patient atteint de cancer du côlon de la cohorte System Biology of Colorectal Cancer (SYSCOL). Nous avons découvert 10 111 et 5 152 loci quantitatifs d’expression TE (eQTL) dans les échantillons normaux et tumoraux, respectivement. Parmi ces loci, 429 étaient spécifiques à la tumeur et très probablement liés à des changements dans les niveaux de méthylation de l’ADN. Nous avons découvert que les facteurs de transcription montrent un enrichissement plus élevé pour les TE-eQTLs spécifiques aux tumeurs par rapport aux TE-eQTLs partagés, soulignant que ces facteurs de transcriptions pourraient réguler l’expression des TE plus spécifiquement que l’expression des gènes dans le cancer. Ensuite, nous avons déduit la relation de cause à effet entre les variants eQTL, les TE et les gènes et découvert 1 758 TE agissant comme médiateurs d’effets génétiques, modifiant l’expression de 1 626 gènes significativement plus dans les échantillons cancéreux que normaux. Enfin, nous avons montré que dans certains cas, les TE-eQTLs spécifiques à la tumeur déclenchent l’impact des TEs sur les gènes avoisinants. Dans l’ensemble, les résultats de ce projet mettent en évidence une nouvelle classe de facteurs de cancer, améliorant ainsi notre compréhension de la tumorigénèse et révélant de nouvelles cibles pour le traitement du cancer.

Dans le second projet, nous avons voulu étendre nos précédents résultats et étudier l’effet des TEs sur les gènes avoisinants au sein de tissus humains adultes sains. Nous avons
utilisé la cohorte Genotype-Tissue Expression (GTEx) pour étudier le contrôle génétique des TEs et mesurer leur impact sur l'expression des gènes dans 43 tissus adultes différents. Nous avons découvert des TE-eQTL dans tous les tissus, allant de 1 086 dans l'hippocampe du cerveau à 22 634 dans la thyroïde, ce qui indique que les TE sont sous un contrôle génétique fort. Nous avons observé de nombreux facteurs de transcription enrichis dans les TE-eQTLs par rapport aux gène-eQTLs, ce qui indique que ces facteurs de transcriptions ont une plus grande affinité pour les eQTLs régulant l'expression des TE que ceux régulant l'expression des gènes. Nous avons ensuite évalué la relation de cause à effet entre les variants eQTL, les TE et les gènes et, en regroupant les résultats des 43 tissus, nous avons identifié 28 888 TE qui transmettent des effets génétiques à 12 599 gènes voisins, influençant leur expression. Enfin, nous avons montré que l'effet causal des TE sur l'expression des gènes est partagé entre des tissus étroitement liés et nous avons découvert plusieurs cas où les TEs ont un impact sur des gènes proches en agissant comme promoteurs alternatifs. Collectivement, ces résultats apportent d'avantage d'informations sur l'activité transcriptionnelle des TE dans les tissus humains adultes et soulignent un rôle fonctionnel important des TE dans la régulation transcriptionnelle des gènes avoisinants. Ces résultats constituent un atlas complet de TE-eQTLs qui peut être utilisé pour étudier d'avantage le rôle fonctionnel des TE dans la régulation de l'expression des gènes, mais aussi pour comprendre le rôle des TE dans les traits et les maladies complexes.

En conclusion, ces études révèlent l'importance des éléments transposables dans la régulation de l'expression des gènes et remettent en question l'hypothèse selon laquelle les TE sont inactifs dans les tissus adultes. Nos résultats illustrent comment, en intégrant des informations supplémentaires provenant de différents phénotypes moléculaires (gènes, éléments transposables et variants génétiques) chez les individus nous permet de mieux comprendre l'interaction entre la partie codante et non codante du génome.
Acknowledgements

I would like to start by thanking my supervisor Manolis Dermitzakis for trusting me and giving me a great opportunity to be a member of his laboratory since 2016. Thank you for giving me the chance to meet and collaborate with great scientists and work on very interesting projects. I will like to thank him further for the scientific guidance and support that he shown me during my PhD.

Secondly, I would like to thank my current supervisor, Prof. Simon Braun for giving me a new "home" for the last year of my PhD. It was a pleasure working with him and I am very happy to be the first PhD student to graduate from his lab.

I would also like to thank Prof. Evgeny Zdovnov and Prof. Ioannis Xenarios for having accepted to evaluate my thesis and be part of my jury.

A special thanks to all the people I collaborated during my PhD, specifically to Professor Didier Trono for giving me the opportunity to be part of his lab this past year and giving me the chance to learn more about the fascinating word of transposable elements.

A lot of people have helped me during this journey. I’ll name just a few but I sincerely remember everyone and thank you all! First I would like to thank Ana Viñuela and Andrew Brown who were the first two to take me in charge during my internship in the FunPopGen lab back in 2016 and helped me understand not only what eQTLs are but allowed me to learn so much during the years we were colleagues. Thank you for showing great amount of patience when I was bugging you constantly with questions. Andrew, thank you very much for introducing me to the magnificent word of ggplot2, I can never go back now. I would also like to thank Halit Ongen who has been supervising me during my masters and PhD and with whom it was a pleasure working with for the past 3 years. I would also like to thank Hadrien Soldati for the many discussions, laughs and beers we had together over this past year. Also, I would like to thank Christelle Borel for the great scientific and everyday life discussions we had over the past years.

I am also very grateful to all my former and current colleagues for the friendly and pleasant environment they created in the lab, for all the laughs we had together and for the scientific advices and discussions. For their valuable help and support I would like to thank Olivier Delaneau, Cedric Howald, Niko Panousis, Marco Garieri, Ambra Sartori, Aline Real, Anna Ramisch, Maris Alver, Pauline Sararols, Chloe Mayère, Maria Josefina Lascano Maillard, Hanna Schwämmle, Younju SO, Tatjana Stevanovic Stanic, Diana Avalos, Evarist Planet and Séverine Reynard.
A special thank you to our lovely Ancilla Stefani for all the great discussions and laughs we had over the years. It was truly a pleasure working with you and discuss each and every morning about random things. You are greatly missed.

A big thank you goes to my second family, Carmen Mury, Stewart Jackson and Sophia Reyes for all the lovely discussions and support you gave me during all these years. You have been part of this PhD journey since the beginning providing me with lots of positive attitude.

A Huge thanks goes to my family for their unparalleled love, support and help during these past years. ’Ότι και να πω δεν θα είναι αρκετό για να περιγράψω την ευγνωμοσύνη μου για όλα όσα κάνατε για έμενα, για τη συνεχή ενθάρρυνση, πίστη και αγάπη που μου δείχνετε. Ευχαριστώ πολύ τον ιπτάμενο αδελφό μου για τη μεγάλη συμπαράσταση που μου δείχνει. Στους γονείς μου, σας αφιερώνω αυτό το διδακτορικό γιατί χωρίς εσάς, τίποτε δεν θα ήταν δυνατό.

Finally, I would like to thank the most important person in my life, my colochita, Marisol Reyes. I cannot express my feelings and gratitude in a few words. I will only say a huge thank you for being part of my life, for helping me during the tough times to stay positive and for encouraging me to become better. Thank you for the patience you showed during my PhD years. We finally made it and the future can only be brighter.
## Contents

Abstract iii  
Résumé v  
Acknowledgements vi  

1 **Introduction**  
1.1 Regulation of gene expression  
1.2 Transposable elements in the human genome  
1.2.1 Overview of transposable elements  
1.2.2 Host mechanisms controlling transposable elements  
1.2.3 Transposable elements and their impact on the human genome  
1.2.4 Transposable elements and cancer  
Cancer in a nutshell  
Widespread epigenetic dysregulation of transposable elements  
Transcriptional deregulation of TEs  
Genomic instability and chromosomal rearrangements  
1.3 Genetics of gene expression  
1.3.1 A short history of genetics  
1.3.2 Complex traits and Diseases  
Genome-wide association studies  
expression Quantitative Trait Loci (eQTL) studies  
1.4 Thesis aims  
Study 1: Transposable elements mediate genetic effects altering the expression of nearby genes in colorectal cancer  
Study 2: The interplay between genetic variants, transposable elements and gene expression in adult human tissues  

2 Transposable elements mediate genetics effects onto genes altering their expression in colorectal cancer 19  

3 The interplay between regulatory variants, transposable elements and gene expression in adult human tissues 70  

4 Discussion 111  
4.1 Transposable elements as mediators of genetic effects onto genes in cancer 112  
4.2 Transposable elements play an important role in the regulation of gene expression in human adult tissues 113  
4.3 Limitations 114
CHAPTER 1

Introduction

In Brno, Austrian Empire, from 1857 to 1864, Gregor Mendel studied inheritance patterns of thousands of common peas and tracked distinct traits from parents to offspring. He explained his findings in terms of discrete inherited units that give rise to observable physical characteristics [1]. These discrete inherited units are what we call genes, a term introduced around twenty years later by Wilhelm Johannsen [2].

Our genome is constituted of a sequence of nucleotides (A, T, C, G) and is divided in two main categories. The coding-genome which contains the genes, the basic physical and functional unit of heredity and the non-coding genome which does not encode for genes. Many genes encode for proteins, the building blocks of cells, but some do not. In humans, genes vary in size from a few hundreds nucleotides to more than hundreds of thousands of nucleotides. Every person has two copies of each gene, one inherited from each parent. There is only 1% of the human genome that is different amongst individuals and this 1% is responsible for all the different human traits (height, eye color, skin color and many more).

Gene expression is the combinatorial effect of a myriad of complex processes in which genetic information encoded in a gene gives rise to a functional product, necessary for cells to operate and function properly. Gene products can be either proteins or non-coding RNA molecules and represent the phenotypic manifestations of genes. All these complexes are under strong regulation encoded in the genome of every living organism and summarised in the "central dogma of molecular biology" postulated by Francis Crick in 1958 [3, 4]. The information encoded in genes is transferred from DNA to RNA in a process called transcription. Then, the RNA is used either as a template to produce proteins (mRNA) through translation or it can have a direct functional role (e.g. rRNAs, microRNAs or tRNAs).

Gene expression is a very complex process that requires multiple steps in order to obtain the final functional product. As proteins are crucial for cell function and structure, all steps going from DNA to protein production are under very strong regulation in cells.
1.1 Regulation of gene expression

One of the major mechanisms of gene regulation is transcriptional control [5] since the initiation of transcription is a general phenomenon found in all human cells despite their type and function. Transcriptional control takes place with the support of cis regulatory elements located in the vicinity of genes. These regulatory elements are recognised by transcription factors (TFs) which are recruited to start and control gene expression. TFs and their cofactors (e.g. non-coding RNAs and chromatin remodellers) [6] are the major elements of gene transcriptional control. TFs possess DNA binding domains and can act as activators, by activating transcription or repressors, by repressing transcription [7–10]. Cofactors lack DNA binding domains and are generally recruited by TFs to stabilise the RNA pol II complex or to modify the chromatin structure which is essential for the initiation and elongation of transcription.

Over the last years, the scientific community has made tremendous efforts in identifying and characterising the regulatory elements of the human genome. Consortia such as ENCODE (Encyclopedia of DNA elements) [11] have made major progress by assigning a function to almost 80% of the genome. The two most studied regulatory elements in the genome are promoters and enhancers. Promoters are sequences of DNA approximately 200 base pairs long right upstream or right next to a gene that is transcribed. They are either characterised as core promoters where the transcriptional machinery bind or as proximal promoters where specific TFs such as activators are recruited [12]. Enhancers on the other hand, can be found upstream or downstream, close or far away from the Transcription Start Site (TSS) of the genes they regulate and in various genomic locations like inside exons, introns or in intergenic regions [13]. They have the ability to control transcription of genes in a spatio-temporal way, that is within specific tissues and within specific times like during development [14]. Enhancers are binding sequences for transcription factors, specifically activators, increasing the probability that a gene they regulate gets transcribed. As enhancers can be far away from genes, formation of DNA loops are necessary to bring them close to promoter and activate transcription.

As previously stated, enhancers and promoters are the two most studies regulatory elements but other types exists as well. Insulators prevent undesirable gene regulation by either preventing enhancer - promoter interactions [15,16] or by blocking silencing of genes through repressive chromatin spreading [17]. Silencers are very similar to enhancers but bind repressors reducing the expression levels of the genes they regulate. Locus control regions (LCRs) are groups of different regulatory elements (enhancers, promoters, insulators) acting together to regulate gene clusters or loci in a spatio-temporal manner [12]. Overall, the interactions between cis regulatory elements, TFs and chromatin accessibility act synergistically to regulate gene expression. However, genetic variation also plays a crucial role in transcriptional regulation and will be discussed later in this thesis. Strikingly, many of these regulatory elements and various regulatory RNAs are present in the DNA sequence of repetitive elements, commonly called transposable elements (TEs). TEs are becoming more and more appreciated for their impact on the regulation of gene expression. In this next section, we are going to dive into the fascinating word of TEs, discuss about the different ways the host organisms are able to control their activity and finally highlight how TEs can impact the human genome in a positive but also detrimental way giving rise to various diseases such as cancer.
1.2 Transposable elements in the human genome

In 1950, Barbara McClintock discovered transposition (the ability of transposable elements to change positions in the genome) through cytogenetic analyses of maize chromosomes and proposed that mobile genetic elements are essential controllers of gene expression [18]. However, at that time, her claims were tremendously criticised by the scientific community which were convinced that DNA was the static carrier of genetic information, transmitting unchanged genes from parents to offspring. Later on, the rediscovery of transposons in prokaryotes and the description of the Lac operon [19] led to the realization that McClintock’s proposal was correct.

A haploid copy of the human genome comprises of approximately three billion base pairs where only 1% to 2% of the DNA encodes for exomes (protein coding genes) [20–24]. Strikingly, more than half of the human genome is derived from transposable elements [21, 25]. Additionally, the contribution of transposable elements to the genome is likely underestimated by genetic analyses, because old TEs are unrecognisable and because the repetitive nature of TEs complicates efforts to assign their sequences to specific genomic loci. All these findings suggest that TEs account for more than two thirds of the genome with 3 to 4 million individual integrants [26].

TEs were initially thought to be purely parasitic and coined junk DNA. However, TEs are more and more recognised as crucial contributors to evolution through processes such as somatic mutations, gene shuffling, transcriptional modulation, dispersion of regulatory sequences and many other types of chromosomal rearrangements. TEs and their hosts co-evolved by achieving a fine balance between potentially deleterious consequences of specific mutations for the individual and adaptive benefits of genetic diversity for the species [27, 28]. In the next sections, I will describe briefly the different classes and families of transposable elements and outline the different ways TEs can impact the human genome.

1.2.1 Overview of transposable elements

The simplest way to characterise TEs is through their transposition mechanism [27, 29]. There are two main classes of transposable elements, the DNA transposons representing approximate 3% of the human genome with around 500,000 DNA transposon insertion [30]. DNA transposons move through a non-amplifying mechanism, called cut-and-paste process mediated by a TE-encoded recombinase (Figure 4.2). The different members of DNA transposons are solely still active in zebrafish and Drosophila whereas in primate lineages, they have lost their transposition capabilities around 37 million years ago (Mya) [31]. The other main class of transposable elements are the retrotransposons (Figure 4.2). They represent around 90% of TEs in the human genome and account for approximately one-half to two-thirds of genomic DNA. Retrotransposons, also known as endogenous retroelements (ERE)s are phylogenetically related to retroviruses [32]. EREs spread through a copy-and-paste mechanism where the EREs get transcribed, then reverse transcribed to cDNA and finally integrated into a new genomic location, amplifying EREs.

Retrotransposons can further be subdivided into autonomous and nonautonomous EREs. Autonomous EREs encode all the necessary proteins for their reverse transcription and integration, whereas nonautonomous rely on autonomous EREs for reverse transcrip-
1.2. TRANSPOSABLE ELEMENTS IN THE HUMAN GENOME

Autonomous EREs can be subdivided further into long terminal repeat (LTR) and non-LTR retrotransposons (Figure 4.2). LTR retrotransposons are also named Endogenous Retroviruses (ERVs) or HERVs for human ERVs because they are relics of exogenous retrovirus that integrated into the germline of the host species or ancestors [32]. Non-LTR retrotransposons are more abundant than ERVs with an estimated 3 million integrants representing around 40% of the human genome. Among them, we can find the long interspersed nuclear elements (LINES), the short interspersed nuclear elements (SINEs) which in humans comprise of Alu elements, and SVAs (Figure 4.2). SINEs and SVAs do not code for proteins and require proteins encoded by LINES to spread [32]. Collectively, non-LTR retrotransposons are responsible for approximately one new integration every 20 to 50 births in humans [30,33].

TEs come in different shapes and forms and the number of different types of TEs is remarkable. Active TEs in the genome of an organism can be very dangerous as they can very quickly start integrating in various locations, damaging the host genome. Thus, in the next paragraphs, we are going to focus on the various mechanisms that host organisms employ to regulate the activity of TEs.

**Figure 1.1** – Structure of transposable elements. An example of each of the main classes of TEs are depicted with the average genome length indicated underneath. Modified from [32].
1.2.2 Host mechanisms controlling transposable element

Mammalian species have developed a broad variety of ways to prevent TEs inducing damage to their genomes. Two main mechanisms are used, cells can either prevent the novel integration of TEs or prevent the expression of TEs by regulating their transcription. [32].

Integration in the genome of EREs is often blocked by a variety of pre-integration restriction factors which recognise ERE’s RNA transcripts or proteins. For example, members of the APOBEC family of cytidine deaminases act by editing reverse transcripts [34, 35]. They are able to inhibit a broad range of retroelements from ERVs, exogenous retrovirus, L1 and Alu [35–42].

Restricting pre-integration of TEs limits the load of TEs in higher organisms, however, targeting TE transcription by host factors is of greater evolutionary relevance as they not only control the spread of TEs but also allow the domestication of TEs. This is mainly achieved by locus-specific deposition of repressive chromatin marks, primarily of DNA methylation and histone H3 lysine methylation [32]. DNA methylation at CpG dinucleotides plays an important role in silencing TEs in both somatic and germ cell [43, 44]. Interestingly, demethylation does not affect all TE classes similarly where LINEs lose significantly more 5-methyl-cytosine phosphate guanine (5-mCpG) compared to other TE families [45–47]. DNA methylation is not the sole repressor of TEs, histone repressive marks play a major role as well, notably H3K9me3, a histone modification found at many TEs [48–54].

Thus, the fundamental roles of DNA and histone methylation in silencing TEs is firmly established. However, sequence-specific protein repressors and TE-driven small RNAs seem to play a role in repressing the expression of TEs. Indeed, the tetrapod-specific Krüppell-associated box (KRAB)-containing zing finger proteins (KRAB-ZNFs) recognize TEs and are able to repress their activity via histone and DNA methylation. In parallel to being targeted by protein repressors, TEs are also regulated by RNA-based mechanisms. RNA interference is accomplished thanks to small RNA molecules that guide repressor protein complexes to targets through a sequence-specific manner and act both at the transcriptional and post-transcriptional level. Many of these small RNAs (e.g piRNAs, siRNAs) are produced from TEs, which they silence in return. This negative feedback loop likely helps reestablish TE repression during development after their activation by the genome [32].

1.2.3 Transposable elements and their impact on the human genome

Transposable elements are a powerful genetic force and major driver of evolution. They have the ability to overcrowd the genome with thousands of identical sequences promoting recombination events that trigger deletions or duplications [55, 56]. They are also able to induce a wide range of transcriptional changes either by disrupting exons, acting as promoters and initiating gene transcription, inducing premature polyadenylation and acting as enhancers for nearby genes (Figure 1.2) [32].
Transposable elements have a vast positive effect on mammalian species from millions of years of purifying selection. For example, the recombination-activating genes RAG1 and RAG2, two genes critical for V(D)J recombination (mechanism of somatic recombination occurring in early stages of B and T cell maturation giving rise to the highly diverse repertoire of antibodies and T cell receptors) and development of the immune system are likely to have originated from a DNA transposon around 500 Million years ago (Mya) [57, 58].

However, the major influence of transposable elements on higher species is there vast reservoir of transcriptional regulators. TEs contain cis-acting regulatory sequences with the power to modify gene expression through enhancer, promoters or insulator effects (Figure 1.2) [59–67]. Also, many TEs produce small regulatory RNAs (e.g. microRNAs or PIWI-interacting RNAs) several of which are able to alter transcription in trans. Interestingly, the host mechanisms are often evolved in the control of TE spreading, subsequently altering the host gene expression [32]. There is a growing list of cases where vertebrates co-opt ERE derived cis-regulatory elements such as the exaptation of ERV enhancers which appears to have contributed to the diversification of the placenta [68] or ERC sequences acting as promoters in mice and humans enabling ubiquitous or tissue-specific expression of various genes like the mouse sex-limited protein (Slp) gene which is dictated by an upstream ERE [69].

Figure 1.2 – Some of the genomics effects transposable elements can cause on the genome
Modified from [32]
Transposable elements are an excellent source of trans-acting regulatory factors. Indeed, EREs overlap with approximately 5,000 to 10,000 long non-coding RNAs (lncRNAs) in the human genome [70–74]. lncRNAs are non-coding molecules of more than 200 nucleotides in length. Most lncRNAs are transcribed by RNA polymerase II and similarly to mRNAs, are spliced and polyadenylated. They can also be regulated by tissue-specific promoters and enhancers [71,75–77]. The proportion of lncRNAs with functional significance remains unclear but it appears that more than two-thirds of mature lncRNAs contain ERE transcripts accounting for 30% of their total sequence [78]. Even though there is a wide variation between species, in humans, HERVs are significantly overrepresented in lncRNAs [78,79].

Retrotransposons have the ability of being incorporated into exons of coding and non-coding genes [80,81]. Retrotransposon derived exons have been identified in untranslated regions [82–85], in miRNAs [86,87] and in alternatively spliced exons of protein coding genes [83,85,88–90].

All in all, the activity of transposable elements need to be very finely regulated by the host genome in order to avoid their abnormal expression which can be quickly very detrimental for the organism and can lead to severe diseases like cancer. In the next part, we will describe some of the role TEs play in cancer.

1.2.4 Transposable elements and cancer

Cancer in a nutshell

Cancer formation, also called tumorigenesis, is the process where normal cells become cancer cells. This transformation occurs through changes at the cellular, genetic and epigenetic level causing abnormal cell division. Cell division is a physiological phenomenon occurring in almost all tissues. In normal conditions, proliferation and apoptosis is maintained to ensure the integrity of organs and tissues. However, somatic mutations in the DNA and epimutations (changes in the chemical structure of DNA that does not change the DNA sequence) alter the afore-mentioned processes through interference with the regulatory programs controlling cell division and cell death [91–93]. Most mutations are called passenger mutations. These are mutations that do not cause cancer. Only a small subset of somatic mutations, called driver mutations will lead to cancer.

Inherited mutations can predispose to cancer. In addition, environmental factors like radiation or carcinogens can contribute to the formation of cancer. Random mistakes during DNA replication may also result in cancer causing mutations [94]. Generally, one mutation is not enough to develop cancer but a succession of multiple ones is required for the development of cancer. Recent comprehensive studies done by The Cancer Genome Atlas (TCGA) program revealed that on average 12 driver events are needed [95]. Cancer is fundamentally a disease of tissue growth dys-regulation. Mutations in genes that regulate apoptosis, DNA repair and cell division may result in uncontrolled cell proliferation and subsequently lead to cancer. Genetic and epigenetic changes can occur at various levels, from gain of entire chromosomes, to single point mutations affecting just one DNA nucleotide, or even to silencing or activating microRNAs that control the expression of hundreds of genes [96,97]. There are two broad categories of affected genes by the mechanisms described above, oncogenes and tumor-suppressor genes. When proto-oncogenes
are mutated they become oncogenes and can cause normal cells to become cancer cells. Often proto-oncogenes encode for proteins functioning in the stimulation of cell division, inhibition of cell differentiation and halting of cell death [98]. Tumor suppressor genes, on the other hand, regulate cell division and replication and need to be silenced in order tumorigenesis to occur [99].

Another type of mutation that can cause cancer is the integration of genomic material from DNA viruses or retrovirus in the genome. If these contain oncogenes, they can trigger the growth of tumor in the host through a process called viral transformation. Interestingly, there is a growing body of evidence pointing towards an essential role of TEs in human tumorigenesis. We will discuss about the various roles TEs play in tumorigenesis in the following paragraphs.

**Widespread epigenetic dysregulation of transposable elements**

TEs are suggested to play a crucial role in the genome integrity maintenance, thus deregulation of TE activity can lead to significant genomic instability and carcinogenesis [5,30]. TEs and host regulatory factors work in synergy to control and prevent aberrant TE activity through a myriad of epigenetic mechanisms [100]. TEs independently recruit silencing signals to repress single genes or large chromosomal regions [100, 101] and deploy self-control mechanisms repressing their retrotransposition. The primary factors regulating TE transcription are epigenetic, mainly DNA methylation and histone modifications, that successfully suppress TEs [100,101]. In the context of cancer, there is a global hypomethylation leading to reactivation of L1 elements as has been observed in colorectal cancer [102], breast cancer [103] or hepatocellular carcinoma [104]. The hypomethylation of L1 elements and subsequent increase in expression is reported in both solid tumors and leukemia and is frequently correlated with poor clinical outcomes [102,105]. Additionally, the hypomethylation of L1 element and Alus in circulating blood can aid in cancer diagnoses [106]. Moreover, the human endogenous retroviruses (HERVs) are also related to carcinogenesis. Indeed, it was reported that HERV RNA and reverse transcriptase enzymes see their levels increase in lymphomas and breast cancer [107]. HERV-like-viruses have also been associated with metastatic melanomas [107] and ovarian, colon and testicular cancers express higher levels of HERV [107].

Heterochromatin formation, in cooperation with DNA methylation and small RNAs, is also important in regulating TE silencing [100]. Histone tail modifications impact positively or negatively the binding of proteins and transcription factors. Nucleosomes associated with TE genomic sequences are commonly methylated at the histone 3 lysine 9 (H3K9) indicating transcriptional repression and inactive chromatin [108]. Various studies have shown that mutations of H3K9 methyltransferases cause TE upregulation and deregulation of DNA methyltransferases during embryonic development lead to upregulation of TEs, associated with developmental disorders [100,109]. Additionally, many non-coding RNAs and specific DNA sequences are implicated in maintenance of DNA methylation at specific genomic loci, including TEs [100,110]. Many proteins involved in chromatin remodeling are also involved in TE silencing [100,111]. It has been observed that many TEs, particularly SINEs, contain binding sites for CTCF that function as insulators and regulate chromatin condensation [66,112]. A study showed that the transcriptional repressor CTCFL (BORIS) binds to DNA repeats, preferentially SVA repeats (SINES, VNTRs, Alus) suggesting a role of BORIS in the regulation of active TEs in cancer [113]. Even though there is more and more evidence highlighting that mutations and deregulation of chro-
matin remodeling genes induce an increase in TE activity in human cancers, their direct effects on the reactivation of repetitive elements to initiate genomic instability and oncogenic activation needs further clarification [114,115].

**Transcriptional deregulation of TEs**

TEs can alter gene transcription through transposition into transcriptional regulatory regions and induce functional defects of various regulatory regions like promoters, silencers or enhancers or even create new exons that interfere with the biological functions of genes [101]. TE insertions in exons may affect the open reading frame (ORF) and induce missense or nonsense mutations that subsequently can have deleterious consequences [101]. Additionally, integration of TEs in genes can create new splice sites, create polyadenylated signals or perturb the canonical splice sites [115–117]. The integration of TEs in intronic regions can also create new alternative spliced exons that are biologically functional [118,119]. A study showed that more than 80% of Alu insertions in exons caused frame-shift or premature stop codons [90]. Furthermore, TE insertions in 3'UTRs and introns can impact mRNA stability, localization, and translation [120] leading to a downregulation of the impacted genes.

Many transcription factors that are involved in carcinogenesis like FOXA1, GATA, P53 have binding sites in Alu element [121]. In addition, P53 has been found to interact with multiple LTR and non-LTR elements. Approximately, 400,000 p53 binding sites in the human genome are predicted to fall within Alu elements [121]. A study showed that mutations on TP53 cause an increase in the expression levels of Alu RNA in many cancers [122]. Another study showed that P53 is able to repress retrotransposon activity through direct interactions with the PIWI-interacting RNA pathway [124].

**Genomic instability and chromosomal rearrangements**

The majority of TE insertions target particular genomic loci that are specifically recognised by endonuclease domains leading to alterations of nearby gene expression and significant genomic deletions [125]. In certain conditions, random TE integrations can lead to insertional mutagenesis leading to loss of heterozygosity, deletions, duplications, translocations, and inversions accounting for around 0.3% of total mutations observed in cancer [126,127]. Multiple mechanisms leading to TE-mediated genomic instability in cancer have been reported such as the insertion of inverted TEs in the human genome leading to DNA damage [128]. Alus that are inversely inserted during DNA replication tend to form secondary hairpin structures that can lead to double-strand breaks (DSBs) [128, 129] even though this phenomenon is quite rare in the human genome. Additionally, TEs induce genomic instability through induction of unstable microsatellite seeding after TE insertions [126].

Microsatellite instability is an important feature of multiple cancers like colorectal cancer or brain cancer [131]. The incorporation of TEs into introns can produce unstable long repeats in the microsatellites making these regions prone to DNA damage [126,131]. TE-mediated genomic instability also alters overall transcription in cancer. Introduction of new TE-mediated promoters can upregulate stress-associated genes that subsequently increase the chance of DNA damage [127,132].

TE-mediated genomic instability can lead to DNA repair signalling from the host organism followed by recruitment of the DNA repair machinery to the TE insertion loci. This is then followed by cell cycle arrest subsequently altering the efficiency of TE mobility.
However, TE insertions target genes that are involved in the DNA repair pathway. Many studies have shown that the breast cancer susceptibility gene \textit{BRCA2} has been affected by multiple TE insertions in cancer [133,134] and the same for the \textit{APC} gene, suggested to play a role in colorectal cancer [135,136]. Additionally, L1 elements transposition is also associated with DNA breakage and even after repair, deletions occur [137,138]. Transposition events can also cause chromosomal rearrangements. Alu elements can be found in several breakage point of chromosomal translocations and are closely related to cancer [139]. TE-mediated insertions induce deletions ranging from a few bases to thousand nucleotides [140].

Collectively, transposable elements are key players in the regulation of gene expression. However, their activity needs to be finely orchestrated as they can deploy a plethora of different ways for causing serious damage to an organisms' genome.

### 1.3 Genetics of gene expression

Genetic variation plays a crucial role in the regulation of gene expression. In this section, we will focus on human genetics and how recent advancements in technology have allowed scientists to answer important biological and medical questions shedding light on the information that is stored in our DNA.

#### 1.3.1 A short history of genetics

The basic concepts of heredity were probably already known thousands of years ago as humans started domesticating animals [141] and plants [142,143]. Our ancestors had probably realised that offsprings had more similarities to their parents than to unrelated individuals, allowing them to better select crops and to tame animals. Thus, domestication was a key step which allowed societies to switch from a hunter status to an agricultural one, granting them the ability to thrive.

More than 2000 years ago in ancient Greece, two philosophers, Hypocrates and Aristotle suggested that offsprings inherit traits from their mother and father through a blend and until the beginning of the 20th, this concept of heredity was still prevalent until later, Jean-Baptiste Lamarck proposed his own theory of evolution [144]. Modern genetics began around 1900 with the laws of inheritance postulated by Gregor Mendel. He discovered that the first generation (F1) of a cross between pure breed peas (homozygotes) had an uniform phenotype that corresponded to the dominant phenotype in the F0 generation suggesting that there is no mixing of parental traits. He later proposed that traits are inherited independently, which we know today that this is only true if the underlying genetics are not in close proximity to each other.

It was only until 1944, that the physical support of heredity, molecules of DNA that contain genetic information were discovered by Avery, MacLeod and McCarty [145]. Their findings contradicted the popular opinion that proteins were the ones transmitting genetic information. In 1953, Jim Watson and Francis Crick with the precious help from Rosalind Franklin and Raymon Gosling discovered that DNA is a double helix (Figure 1.3) of complementary strands of nucleic acids.
In 1972, Walter Fiers and his team at the University of Ghent, Belgium, were the first to determine the sequence of a gene encoding the bacteriophage MS2 coat protein [146]. A few years later, Richard J. Roberts and Phillip Sharp discovered in 1977 that genes can be split into segments leading to the idea that one gene can make several proteins. Several year later in 1995, the genome of bacterium *Haemophilus influenzae* is the first genome of a free living organism to be sequenced [147] and in 1996, *Saccharomyces cerevisiae* was the first eukaryotic genome sequenced to be released. Since then, the successful sequencing of many organisms’ genomes has complicated the molecular definition of the gene. In particular genes do not always sit side by side on DNA like discrete beads. Instead, regions of the DNA producing distinct proteins may overlap and the idea that “genes are one long continuum” emerged [148, 149]. In 1998 the first genome sequence of a multicellular eukaryote, *Caenorhabditis elegans* was released. Two years later, the full genome sequence of *Drosophila melanogaster* was completed. The biggest step was achieved in 2001, when the sequencing of the human genome was completed yielding unexpected findings. For example, the human genome is composed of only around 20,000 to 25,000 protein coding genes [150], a much smaller number than previously anticipated and closer to the number of genes of less complex organisms like *C. elegans* [151]. Naturally, many questions started puzzling scientists, like to what extent genomes differ between individuals and how these differences could impact health, diseases or responses to external factors.

A year later the HapMap consortium [152–154] started with the aim to provide answers to the above-mentioned questions. Initially, the project genotyped one single nucleotide polymorphism (SNP) every 5kb across 269 individuals of different geographical origin [152]. Interestingly, they discovered that a non-random association between alleles at specific loci within a population, known as linkage disequilibrium (LD) [153]. After
studying solely common genetic variants, in the final phase of the HapMap project, they incorporated rare genetic variants as well [154]. In (Figure 1.4) we can appreciate how well genetic variation between population reflect the geographical origins [155].

The cost of exome and genome sequencing dropped sufficiently to allow studying not just SNPs but the entire exome (whole exome sequencing, WES) or even the entire genome (whole genome sequencing, WGS). From this another collaborative effort took place named the 1000 Genomes Project Consortium [156,157] and later the UK10K project [158] with the ultimate goal to sequence 10,000 individuals. This tremendous gain in statistical power has given scientists the ability to study variants with very low frequencies in the population (rare variants) and understand their effects on health and disease.

![PCA plot depicting genetic variation between European populations from 1,387 individuals](image)

**Figure 1.4** – PCA plot depicting genetic variation between European populations from 1,387 individuals. We can see that genetic variation highly reflects geographical origins. Each individual is represented with different colours depending on their geographical location, as marked in the subfigure on the top right. Modified from [155]

The sequencing of the human genome has started a new era of research where scientists can now perform whole genome association studies between genetic variation and complex traits and phenotypes. Before diving into these, let us first briefly introduce complex traits and the different disease types that occur in humans.
1.3.2 Complex traits and Diseases

Disease progression and susceptibility varies among individuals and often certain individuals react better to treatments for a disease than others. Understanding the causes of these variable susceptibilities and responses and their synergistic effects is crucial for improving prevention and treatment of diseases.

Human diseases have multiple causes, some are caused by environmental factors and lifestyle choices like asthma or chronic obstructed pulmonary disease (COPD), which are mainly due to smoking and air pollutants [159]. Diseases can be caused by microorganisms like bacteria, viruses or fungi. Many diseases have a genetic cause and are classified into three main categories (i) caused by chromosomal abnormalities, (ii) Mendelian diseases and (iii) polygenic or complex diseases.

Diseases caused by chromosomal abnormalities are characterized by inversions, translocations, duplications or deletions (Figure 1.5). Trisomy 21 or Down's syndrome is probably the most known genetic disorder caused by chromosomal segregation where an extra copy of chromosome 21 is present.

Mendelian diseases, also called monogenic, as only one gene is involved in the disease, follow Mendel’s law of inheritance. They can either be recessive, meaning that both copy of the abnormal allele is required for disease manifestation or dominant where only a copy of the abnormal allele is necessary for the disease to develop. Cystic fibrosis and Huntington's disease are excellent examples of recessive and dominant disorders, respectively. In general, monogenic disorders are rare and under purifying selection [160]. Sickle cell anemia is one of the most well characterized monogenic recessive disorders. This disease alters the structure of hemoglobin. Hemoglobin, then, clusters with other hemoglobin molecules causing blood clots all over the body. Interestingly, this change in the molecular structure of hemoglobin prevent the replication of *Plasmodium falciparum*, preventing
malaria [161]. Monogenic diseases are remarkably "easy" to understand by scientist. They just need to find the genetic mutation at a specific gene and the biological mechanism of the disease is discovered. In that regard, linkage mapping studies were very successful in the discovery of the molecular causes of monogenic disorders [162].

The third category of human diseases are the polygenic diseases where multiple genetic factors are in play following an intricate and unclear pattern of inheritance [163]. Polygenic disorders, also named complex disorders, are linked to genetic and environmental factors. Even though these diseases are more prevalent in certain families, the complexity of their inheritance make predicting them using genetics very difficult. The vast majority of human diseases are polygenic and many of them are common in the population like Alzheimer's disease, Parkinson's disease, autism, diabetes and cancer [164]. Genes associated with complex diseases are inherited but the genetic factor is not always the only factor required; environmental and lifestyle factors are key players too. Even though there could be genetic predisposition to a particular complex disease, it does not necessarily mean that individuals will develop the disease. The development of complex diseases depends on lifestyle, environmental factors, genetic factors and their synergistic effects between the three which makes their comprehension and the discovery of treatments very challenging.

Aside from complex diseases, lots of non-pathogenic human traits are polygenic and influenced by environmental and lifestyle factors such as height, body mass, and many more. Some of these traits can be risk factors for some complex diseases but assessing the causality between the two remains a very difficult task.

**Genome-wide association studies**

Recent advances have allowed us to uncover the link between genetic variation and diseases or traits within a genome wide framework. These analyses are called Genome-Wide Association Studies (GWAS) and have been a tremendous help to go beyond Mendelian disorders, which previously tackled using linkage mapping and positional cloning [165]. GWAS are able to quantify the contribution of small effect size SNPs in complex disease or traits where it compared the frequency of alleles of common variants between two populations that are either non-affected and affected by a specific disease or that have or do not have a particular trait (e.g. high versus low BMI) [166,167].

Using the information already available from reference panels such as Haplotype Reference Consortium (HRC) [168] or 1000 genomes , imputation methods were develop in order to infer all the missing SNPs from genotype arrays [169,170]. Notably, it is always a good practice to replicate your findings with an independent dataset in another population to avoid false positive results, thus validating your associations between genetic variants and complex traits or diseases. Due to the increase and popularity of these studies, the GWAS catalog was created in 2008 [171] to concentrate all reports of genetic variants associated with particular traits or diseases. To date (accessed on 28th of January 2022), the GWAS catalog contains 5,553 publications with 195,109 unique SNPs associated with a particular trait or disease for a total of 326,148 associations. With the increase of sample sizes for GWAS over the last decades and with new methods arising such as meta-analysis [172], GWAS boomed, identifying genetic variants even with smaller effect sizes. However, GWAS has its limitations. Typically, GWAS identifies clusters of SNPs in LD rather than the causal variant. To address that, fine-mapping methods were proposed both
statistical and with functional prioritization [173].

The biggest limitation to GWAS is the fact they do not assess the causality between genetic variation and the downstream disease or trait making the understanding of the phenotype very difficult. Additionally, most GWAS hits are located on intergenic regions [174, 175] making their interpretation strenuous. Henceforth, there was an urgent need in understanding the functional role of these genetic variants, allowing for a better characterization of the aetiology of the traits or diseases that these genetic variants were associated with. This gave rise to expression Quantitative Trait Loci (eQTL) studies which we are going to dive more into below.

expression Quantitative Trait Loci (eQTL) studies

There are many intermediate steps between a genetic variant and its impact on a complex trait like height. In order to understand this association we need to understand all the intermediate steps that occur. In this context, the first step and one of the most important ones is the effect of genetic variants on gene expression. As previously discussed, gene expression is also regulated by genetics too. A genomic variant affecting gene expression is called an expression Quantitative Trait Locus, or eQTL [176–178]. eQTLs affect gene expression either in cis, in the case when the eQTL is located near to a gene (usually around 2Mb around the TSS) or in trans, in the case where the eQTL is distant from the gene in the same chromosome or located in another chromosome. In order to discover eQTLs, we need to perform population studies, where the gene expression profiles of hundreds of individuals are grouped based on their combination of alleles (Figure 1.6).

![Figure 1.6](image)

**Figure 1.6** – Detection of cis- and trans-eQTLs in a population of individuals. The density plot shows that homozygote individuals for the A allele are less frequent in the population but have a higher gene expression whereas the homozygote individuals for the G alleles have a higher frequency in the population but a lower gene expression. Modified from [179]

From many studies, we discovered that eQTLs primarily fall in non-coding regulatory regions of the genome. Typically, cis-eQTLs are located near the Transcription start site (TSS) or on promoters of the associated gene but can also fall on enhancer regions
when they are more distant. cis-eQTLs falling on enhancer regions display higher tissue-specificity than cis-eQTLs present near the TSS of genes \([180, 181]\). Furthermore, it is crucial to study the genetics of gene expression in a tissue relevant for a particular trait or disease. To this end, the Genotype Tissue Expression consortium (GTEx) was created where the end goal was to perform genotyping and profiling of gene expression in 838 individuals in 49 adult somatic tissues \([182]\). To date, this represents the most comprehensive study on how genetic variants affect gene expression and regulation across human tissues. From the results, we learnt that almost all genes have a cis-eQTL and confirmed that they are enriched in promoter and enhancer regions. Additionally they, showed that cis-eQTLs present allele-specific expression. Regarding trans-eQTLs, their numbers were quite low, probably due to lack of statistical power. A very important results discovered was that not all eQTLs are causative, thus many different methods were developed in order to assess the probability of causation of the various eQTLs discovered \([183]\).

Additionally, many methods were developed to colocalize eQTLs with known GWAS variants bridging the gap between a particular trait or disease and the molecular phenotypes associated with it \([184–186]\). All in all, eQTL studies helped the scientific community to better understand gene regulation and transcriptional control but also raised many new questions. All this will eventually serve in the context of personalised medicine \([187]\), allowing for the creation of better treatments for patients.

In this introduction, we briefly discussed about the regulation of gene expression, about various non-coding regulatory elements, in particular transposable elements and we finished by highlighting some key aspects of the genetics of gene expression with the end goal to introduce the various subjects tackled during this thesis. We are now going to dive into the different aims of this thesis.

1.4 Thesis aims

Over the last decades, we have gradually begun to understand the high complexity and variability of the mechanism of gene expression regulation. We identified players that until a few years ago were not even known or observed or even linked to the regulation of gene expression. Combined with these thrilling discoveries, novel technologies have emerged, allowing us to better identify and properly quantify genes and other molecular phenotypes. The current era of genomics is both exciting due to the new discoveries that can be made and challenging because of the continuous efforts needed in order to properly analyse and differentiate true biological signals from a variety of possible false positives.

Thanks to the development of high-throughput methods to quantify omics data, we have unprecedented possibilities to obtain much broader and complete knowledge about different phenotypes such as diseases or complex traits. In this thesis, we were interested in better characterising the relationship between non-coding and coding genome. More specifically, we focused on the interplay between non-coding regulatory variants (eQTL variants), transposable elements and genes in healthy and cancer patients in two distinct studies. The main aims of the studies are presented below along with the main findings.
Study 1: Transposable elements mediate genetic effects altering the expression of nearby genes in colorectal cancer

The goal of this study was to understand whether transposable elements mediate genetic effects onto genes in colorectal cancer. To do this, we used the SYSCOL cohort comprising of 275 normal colon and 276 colorectal cancer samples. We performed eQTL analyses to discover whether TEs are under genetic control and assess any differences to gene-eQTLs. Then, we were interested in understanding the causal relationship between eQTL variants, TEs and genes and see whether TEs are drivers of gene expression by mediating genetic effects onto nearby genes and whether this phenomenon is more prevalent in cancer compared to the normal state. In this project, we discovered that TE expression is under strong genetic control and that certain TEs are drivers of gene expression, specifically in cancer, allowing for genetic effects to be transferred onto nearby genes altering their expression.

Study 2: The interplay between genetic variants, transposable elements and gene expression in adult human tissues

The goal of this second study was to extend the findings made from the first project (study 1) to all human tissues and understand to which extent TEs impact the expression of nearby genes in healthy individuals. For this, we used available genotype data from whole genome sequencing (WGS) and transcriptomic data from RNA-seq from the Genotype-Tissue Expression (GTEx) project to discover. We started by performing eQTL analyses to uncover whether TEs are under genetic control through eQTLs. We then assessed the causal relationship between eQTL variants, transposable elements and genes to determine the extent of TEs impact on the expression of nearby genes. In this study, we challenged further the assumption that TEs are transcriptionally inactive in human adult tissues, showing as in the previous project (study 1) that TEs are under strong genetic control. We found that 28,828 expressed TEs across all 43 tested tissues impact the expression of nearby genes. Interestingly, we identify a subset TE-eQTLs that are specifically enriched in brain tissues, suggesting a potential role in regulating brain function.

Overall, this thesis yielded novel insights into the regulation of gene expression. We shed additional light into the transcriptional activity of TEs, showed that they are under tight genetic control and demonstrated that TEs are important for the regulation of genes in adult human tissues and are key drivers in the alteration of gene expression in cancer.
CHAPTER 2

Transposable elements mediate genetics effects onto genes altering their expression in colorectal cancer

Contribution:

Nikolaos M.R. Lykoskoufis performed all the analyses, prepared all the figures and wrote the manuscript with contributions from all authors. Nikolaos M.R Lykoskoufis is the first and corresponding author of the manuscript.

This chapter is a preprint under review in Nature Communications.

Abstract

Transposable elements (TEs) make up more than half of the human genome and are increasingly recognised as key players in the regulation of gene expression. TEs have a variety of ways that they can impact gene expression and need to be very tightly regulated by the host genome as changes in their activity can lead tumorigenesis. However, the impact of TEs on gene expression in cancer remains to be fully elucidated. Here, we used the Systems Biology of Colorectal Cancer (SYSCOL) cohort comprising of 275 normal colon and 276 colorectal cancer (CRC) samples to discover the genetic control of TEs and the extent to which they impact nearby genes in CRC. By combining genotype arrays and mRNA-sequencing we discovered 10,231 and 5,199 TE eQTLs in normal and tumor, respectively, of which 429 were tumor-specific and 525 shared between both states. We used Bayesian networks to assess the causal relationship between eQTL variants, transposable elements and gene expression and discovered that 1,766 TEs are acting as mediators of genetic effects, altering the expression of 1,558 nearby genes significantly more in tumor compared to normal samples. Of these, 1,558 genes, 55 are cancer driver genes. Additionally, we show that tumor-specific TE-eQTLs trigger the driver capabilities of TEs. These findings collectively shed light on a new class of cancer drivers, advancing our knowledge of tumorigenesis and presenting us with possible new targets for drug development.
Transposable elements mediate genetic effects altering the expression of nearby genes in colorectal cancer

Nikolaos M. R. Lykoskoufis\textsuperscript{1,2,3}, Evarist Planet\textsuperscript{4}, Halit Ongen\textsuperscript{1,2,3,*}, Didier Trono\textsuperscript{4,**}, Emmanouil T. Dermitzakis\textsuperscript{1,2,3,*,**}

\textsuperscript{1} Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland
\textsuperscript{2} Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, 1211 Geneva, Switzerland
\textsuperscript{3} Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland
\textsuperscript{4} School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015, Lausanne, Switzerland

* Corresponding authors: N.M.R Lykoskoufis: nikolaos.lykoskoufis@unige.ch and E.T.D: emmanouil.dermitzakis@unige.ch

** Co-last authors
ABSTRACT

Transposable elements (TEs) are interspersed repeats that contribute to more than half of the human genome, and TE-embedded regulatory sequences are increasingly recognized as major components of the human regulome. Perturbations of this system can contribute to tumorigenesis, but the impact of TEs on gene expression in cancer cells remains to be fully assessed. Here, we analyzed 275 normal colon and 276 colorectal cancer (CRC) samples from the SYSCOL colorectal cancer cohort and discovered 10,231 and 5,199 TE expression quantitative trait loci (eQTLs) in normal and tumor tissues, respectively. Amongst the latter, 429 were exclusive to CRC, likely driven by changes in methylation patterns. We identified that transcription factors are more enriched in tumor-specific TE-eQTLs than shared TE-eQTLs, indicating that TEs are more specifically regulated in tumor than normal. Using Bayesian Networks to assess the causal relationship between eQTL variants, TEs and genes, we identified that 1,766 TEs are mediators of genetic effect, altering the expression of 1,558 nearby genes significantly more in tumor compared to normal, of which 55 are cancer driver genes. We show that tumor-specific TE-eQTLs trigger the driver capability of TEs subsequently impacting expression of nearby genes. Collectively, our results highlight a global profile of a new class of cancer drivers, thereby enhancing our understanding of tumorigenesis and providing potential new candidate mechanisms for therapeutic target development.
INTRODUCTION

Understanding the mechanisms underlying tumorigenesis has been one of the main research questions in cancer biology. While somatic mutations, chromosomal rearrangements and gene amplification are the three main hallmarks driving cancer progression, they are unable to provide a complete explanation of tumorigenesis. Recent discoveries have demonstrated that transposable elements (TEs) have contributed to the evolution of gene regulation and can alter the landscape of gene expression in development and disease [1-5]. Transposable elements (TEs) are interspersed repeats that contribute more than half of the human genome. TEs, more specifically TE-embedded regulatory sequences (TEeRS) are broadly active during the phases of genome reprogramming that occur in the germline and the early embryo, and then controlled by epigenetic mechanisms that still allow their finely orchestrated participation in biological events as diverse as brain development, immune responses, and metabolic control. The aberrant re-activation of TEeRSs is observed under certain conditions and disease states, notably cancer [6-8]. Transcription is defined by the coordinated activity of regulatory elements which are modulated by genetic variation. Thus, we speculate that transposable element expression is influenced by regulatory non-coding variants, also called expression Quantitative Trait Loci (eQTLs), known to contribute to the onset and progression of complex diseases like cancer [9, 10]. To build on this concept, we set out to analyze the interplay between regulatory variants (eQTLs), transposable elements and gene expression to characterize the genetic perturbation of TE and gene expression in cancer. To this end, we integrated genome-wide genotyping data (genotype array) and transcriptomic profiles (bulk RNA-sequencing) from the Systems Biology of Colorectal Cancer (SYSCOL) cohort comprising of 275 and 276 normal and tumor samples, respectively.

RESULTS

Quantifying transposable elements (TEs) and gene expression

To measure the expression of TEs in CRC, we examined transcriptomes obtained by RNA-seq from 275 normal and 276 CRC samples from the SYSCOL cohort[11]. We quantified TE and gene expression using an in-house curated TE annotation list originating from the RepBase database [12] that contains approximately 4.6 million individual TE loci. These annotations were merged with gene annotation from ensembl (v19). Filtering for uniquely mapped reads (Methods) to obtain robust estimates of TE expression resulted in 50,921 TEs and 17,430 genes (protein coding and lincRNAs). We observed that the majority of expressed TEs present in our dataset are SINEs (Alu and MIR), LINEs (L1 and L2) as well as different subfamilies of Long Terminal Repeats (LTRs) and DNA transposons. However, when we looked at the proportion of expressed TEs per subfamily, SVA and ERVK were most prominent (Figure 1A, Supplementary figure 1). Additionally, we used available data from Encode [13] and miRbase [14] to generate a list of regulatory regions and discovered that 13,590 expressed TEs overlapped with at least one previously identified regulatory element. We also discovered that expressed TEs are significantly enriched for most regulatory regions, except for enhancers, compared to non-expressed TEs (Supplementary table 1; Figure 1B, Supplementary figure 2) highlighting their potential role in gene expression regulation.
Transposable elements are under strong genetic control

Using TE expression quantifications and genotype data we first sought to assess the impact of inter-individual genetic variation on TE expression. We conducted cis-eQTL analysis followed by a forward backward stepwise conditional analysis (Methods) and discovered a total of 10,231 and 5,199 TE-eQTLs as well as 6,955 and 1,552 gene eQTLs in normal and tumoral tissue, respectively (Supplementary Figure 3A; Supplementary Tables 2,3). Similarly to gene-eQTLs, TE-eQTLs displayed stronger effects and density closer to the transcription start site (TSS) in both normal and tumor samples (spearman rho = -0.33, P<2.2e-16 in normal, spearman rho = -0.25, P<2.2e-16 in tumor) (Figure 2A-B), yet were more proximal to the TSS compared to gene-eQTLs (Wilcoxon P=2.4e-12 in normal; Wilcoxon P=4.8e-07 in tumor; Supplementary Figure 5). We observed that TEs displayed fewer independent eQTLs per TE than genes (Figure 2C-D) while the minor allele frequencies of TE- and gene-eQTL variants were similar (Supplementary Figure 6). Proximal distance of TE-eQTLs to TSS and the smaller number of independent signals per TE could be due to smaller evolutionary time of TE regulatory landscapes in the human genome compared to genes, making proximal effects much more likely.

To corroborate our findings, we used external datasets to replicate our eQTL discoveries. We downloaded available data from GTEx for colon transverse (N=174) and TCGA for colon adenocarcinoma (TCGA-COAD, N=251). We processed both datasets in a similar way as we did with the SYSCOL dataset (methods 2-3). Not all variant-feature pairs were present in the GTEx colon transverse dataset after all filtering steps. Out of the 10,231 TE-eQTLs and 6,955 gene-eQTLs discovered in normal, 8,380 (82%) and 5,930 (85%) TE- and gene-eQTLs, respectively were present in the dataset and could be replicated. From the 5,199 TE- and 1,552 gene-eQTLs discovered in SYSCOL tumor, only 3,221 (62%) TE- and 1,164 (75%) gene-eQTLs were present in the TCGA-COAD dataset. We observe a high replication of our original (Supplementary Figure 7A-C; Supplementary Table 4) in normal (pi1 TE-eQTLs = 0.831 pi1 gene-eQTLs = 0.686) and tumor (pi1 TE-eQTLs = 0.884; pi1 gene-eQTLs = 0.783) (Supplementary Figure 7D-F; Supplementary Table 5) corroborating our findings.

Given previously established roles of tumor-specific gene-eQTLs in tumorigenesis [11], we aimed next at investigating whether tumor-specific TE-eQTLs could similarly contribute as cancer driving factors. To this end, we used linear mix models with an interaction term between variant and tissue (normal/tumor). We discovered that 429 (8%) of the tumor TE-eQTLs are tumor-specific and 1,697 (24%) of the normal TE-eQTLs are normal-specific, with 525 TE-eQTLs shared between both settings (Figure 3A; Supplementary Table 6). For genes, we found 117 (%) tumor gene-eQTLs to be tumor-specific and 902 (%) normal gene-eQTLs to be normal-specific, of which 175 were shared (Supplementary Figure 8A; Supplementary Table 7). Shared TE- and gene-eQTLs were closer to the TSS of TEs/genes compared to tissue-specific eQTLs (Wilcoxon P<2.2e-16) (Figure 3B, Supplementary figure 8B). Additionally, we observed that shared eQTLs conserved their effect in both normal and tumor (Figure 3C, Supplementary figure 8C). These results indicate that TE expression is under strong genetic control and that non-coding germline variants act as drivers of TE expression in cancer as similarly observed for gene expression [11].
Transcription factors regulate TE expression more specifically in tumor

To corroborate the biological relevance of the discovered TE-eQTL variants we performed functional enrichment analysis of TE and gene eQTLs in normal and tumor using available ChIP-seq data from the Ensembl Regulatory Build [15] for 202 TFs and 29 histone marks. We then proceeded with multiple test correction with a given FDR of 5% (methods section 1.6.4). We found significant enrichment for many TF binding sites overlapping the eQTL loci highlighting the functional relevance of the variants discovered (Figure 4A-B; Supplementary Figure 9,10; Supplementary Tables 8,9). At 5% FDR, we discovered 5 significant hits (4 TFs and 1 histone marks) and 16 significant hits (12 TFs and 4 histone marks) that displayed stronger enrichment for TE eQTLs compared to gene eQTLs in normal and tumor, respectively. The TF most enriched over TE-eQTLs in normal tissues was ZNF274, a Krüppel-associated box (KRAB) domain-containing zinc-finger protein (KZFP), whereas the most enriched over tumor TE-eQTLs was TRIM28, the master corepressor that is recruited by the KRAB domain of many TE-binding KZFPs and serves as a scaffold for a heterochromatin-inducing complex capable of repressing TEs via histone H3 Lys9 trimethylation (H3K9me3), histone deacetylation and DNA methylation [16, 17]. Additionally, BDP1 and BRF1, two subunits of the RNA polymerase III transcription initiation factor, were more enriched over TE-eQTLs compared to gene eQTLs highlighting potential transcription of Alu or MIR TEs of the SINE family [18]. These results corroborate the biological relevance of TE eQTLs and point to possible transcription and repression of certain TEs.

To assess the differential effects of tumor-specific versus shared eQTLs, we performed functional enrichment analyses using available ChIP-seq data from LoVo colorectal cancer cells for 220 TFs and 2 histone marks [19] (methods section 1.6.4). We observed that in the case of genes, all tested TFs had a stronger enrichment for shared compared to tumor-specific eQTLs, indicating that these TFs are regulating gene expression in both the normal and tumor state. (Supplementary Figure 11, Supplementary Table 10). In contrast, we found at 5% FDR, 60 significant hits (58 TFs and 2 histone marks) displaying enrichment for tumor-specific versus shared TE-eQTLs, pointing to tumor-specific TE regulation (Figure 4C; Supplementary Figure 12; Supplementary Table 11). Of these, 23 were upregulated and 25 downregulated in tumors (12 were missing from our expression data and could not be tested for differential expression analysis), but we did not observe any significant correlation between the tumor-specific TE-eQTL enrichment to shared TE-eQTL enrichment ratio and fold change in the expression of the corresponding transcription factors (Pearson R = 0.19, p-value = 0.86; Supplementary Figure 13). Thus, differential expression of these TFs is not driving the tumor-specific TE-eQTL effects. However, 61 of the 88 tumor-specific TE-eQTLs overlapping the binding sites of the 60 aforementioned TFs are not significantly associated (FDR = 5%) with any nearby (±1 Mb from TSS) TE or gene in normal, indicating that these regulatory regions are inactive in the normal state (Supplementary Figure 14). Additionally, we compared methylation levels between normal and tumor samples for the tumor-specific and shared eQTLs and observed significantly increased (Wilcoxon rank sum test p-value = 2.5e-11 for TEs and p-value = 0.0037 for genes) methylation over tumor-specific compared to shared eQTLs for both gene and TEs (Figure 3D; Supplementary figure 8D).

Altogether these results suggest that many TFs are regulating TE expression. The inactivity of some of the TE eQTLs in normal and the significant changes in methylation between tumorspecific and shared TE-eQTLs indicate that regulatory switches involving the recruitment of these TFs might underlie the effects of tumor-specific TE eQTLs.
Transposable elements as mediators of genetic effects onto genes

Having established that TEs are under genetic control, we next sought to assess the causal relationship between eQTL variants, TEs and genes and discover the extent to which TEs act as drivers of gene expression in tumor. To this end, we focused on regulatory variants affecting both TEs and genes and detected these in an unbiased manner by first associating TEs with genes using a similar approach to QTL mapping. Next, we quantified the identified 20,083 TE-gene pairs found in normal samples and 140,274 TE-gene pairs found in tumor at 1% FDR and used this newly quantified TE-gene pairs to find all eQTL-TE-gene triplets by performing a standard eQTL analysis (Methods; Supplementary figure 15-17). At 5% FDR, we discovered 11,937 and 9,528 triplets in normal and tumor, respectively, for which we inferred the most likely causal relationship using Bayesian networks (Methods) [20-22]. We tested three models, (i) the causal model where the eQTL variant affects TE expression and then gene expression, (ii) the reactive model where the eQTL variant affects gene expression and then TE expression and (iii) the independent model where the eQTL variant affects independently TE and gene expression (Supplementary figure 18). Bayesian Networks were shown to be an adequate method for testing these three models [23]. We observed significantly more causal models in tumor (47%) compared to normal (23%) (Fisher p-value <2e^-16) indicating that TEs are causal for gene expression predominantly in tumor and to a lesser extent in normal (Figure 5A, B; Supplementary figure 19; Supplementary Table 12,13). We also show that the proportion of causal models correlated with the genomic position of the TE with respect to the gene; intronic TEs tend to be reacting to gene expression. whereas TEs outside the gene body tended to be causal. We believe that the predominance of reactive model from intronic TEs and downstream of genes is a consequence of the transcription of the gene and not the transcription of the TE via the TE promoter. Interestingly, there were significantly more causal scenarios when the eQTL variant lied within the TE, rather than outside (Fisher p-value < 2e^-16) pinpointing to direct regulatory effects of the TE onto gene expression (Supplementary figure 20).

We then proceeded with replicating the causal inference of the eQTL – TE – gene triplets to corroborate our findings. We tested the triplets where all three molecular phenotypes were present in either GTEx colon transverse for the SYSCOL normal colon triplets or in TCGA-COAD for the SYSCOL tumor triplets, yielding 9,577 (80%) triplets and 5,893 (62%) triplets in common, respectively. We performed BNs similarly to the original discoveries. We observe a high replication of both normal (62% similarity) and tumor (74% similarity) results (Supplementary figure 21; Supplementary Table 14,15). We believe that the reason the replication of our normal colon eQTLs is lower than for the tumor eQTLs is because of sample size differences between SYSCOL normal colon (N=275) and GTEx colon transverse (N=174) decreasing our statistical power. These results corroborate our findings and highlight that our discoveries are valid.

Altogether, these results show that TEs are significantly more causal for changes in gene expression in tumor than in normal tissue.
Transposable elements are drivers of gene expression during tumorigenesis

These results suggested that genetic variations in TE expression might drive tumorigenesis. To test this hypothesis, we considered the union of all triplets, i.e. the eQTL variant, TE and gene expression, discovered across tumor and normal tissue and using the same BN approach as previously mentioned, we inferred the causal relationship between the triplets in both states (methods). We similarly looked for shared triplets across the 11,937 normal and 9,528 tumor triplets (eQTL-TE-gene triplets are the same in both states or the eQTL for TE-gene pair is in high LD ($R^2$ > 0.9)). In both shared and union triplets, we observed a significant increase in the causal model in tumor (Fisher Exact Test p-value < 2.2e^{-16} for shared and union triplets) mainly due to independent models and to a lesser extent reactive model shifting to causal. (Supplementary figure 22; Supplementary Table 16,17). Focusing on the 9,528 tumor triplets, we discovered 2,584 (27%) triplets that switched to a causal model in tumor compared to normal, highlighting regulatory changes whereby TEs impacted the expression of nearby genes (Figure 5C). These 2,584 triplets constituted of 1,766 unique TEs impacting 1,558 unique genes. Interestingly, we observed that TEs switching to causal were significantly up-regulated compared to TEs that did not switch models between normal and tumor or that switched but not to causal (Wilcoxon p-value 2.2e^{-14}; Supplementary Figure 23). These results suggest that upregulation of TEs could give rise to their gene expression driver capability.

While expression of most TEs was positively correlated with the expression of the associated gene in tumor (n=2,575) (Figure 5D), only a few showed negative correlation (n=9). Of the significant tumor TE-gene pairs tested in normal colon, we observed that 930 maintained the same effect (in terms of size and direction) whereas 36 showed an opposite effect in tumor samples. Interestingly, of the 1,558 genes, 55 were cancer driver genes (CDG) (3 CRC specific; based on Cancer Gene Census [24]) but we did not find a significant enrichment of CDGs in triplets switching to causal compared to all other tumor triplets (Fisher exact test p-value = 0.2185; odds-ratio = 1.276). For 41 out of the 55 CDGs, we did not find a significant correlation between their expression and the expression of the corresponding TEs in normal samples pinpointing that these TEs have no impact on these genes in the normal state. Taken together, these results suggest an important role of TEs as drivers of gene expression during tumorigenesis.
Non-coding germline variants activate driver TEs during tumorigenesis

We investigated whether any of the 9,528 tumor triplets were constituted of any previously identified tumor-specific or shared TE-eQTLs and assess how the model likelihood changed between normal and tumor. We identified 320 and 133 tumor triplets constituted of a shared or a tumor-specific TE-eQTL, respectively (Figure 6A-B) and observed that the 133 tumor triplets constituted with a tumor-specific TE-eQTL are significantly enriched for triplets switching to causal compared to the 320 tumor triplets constituted with a shared TE-eQTL (Fisher Exact test p-value = 6.6e-4; Odds-ratio=2.04) (Figure 6B). Additionally, we observed that for 120 triplets with tumor-specific TE-eQTLs, the eQTL variant was not a significant eQTL for the corresponding gene in the triplet (Figure 6C), highlighting that the eQTLs get activated in the tumor state influencing TE expression that subsequently impact gene expression. Altogether, these results suggest that tumor-specific TE-eQTLs contribute to tumorigenesis by impacting genes through TEs, adding additional proof that germline variants can be contributing to tumorigenesis.

Driver TEs act as alternative promoters for genes in cancer

It has been shown that TEs could impact gene expression by acting as alternative promoters for nearby genes and creating chimeric transcripts (transposichimeric transcripts (tcGTs)) [25, 26]. To assess whether any of the tumor triplets with causal TEs were affected by tcGT events, we looked for cases where transcripts started from a TE and spliced into a single or multiple nearby genes (methods). We only kept tcGTs made up of the same TE and gene as in the 9,528 tumor triplets and that were significantly more abundant in tumor samples compared to normal samples using a Fisher exact test. At 5% FDR, we discovered 126 tcGTs present in 147 tumor triplets. Of these, 78 triplets (66 TcGTs) were causal and 46 triplets (39 TcGTs) switched to causal from normal to tumor. Interestingly, we detected tcGT events with a known tumor suppressor gene RNF43 and two oncogenes ETS2 and SLCO1B3 supporting the extensive contribution of TEs during tumorigenesis.

DISCUSSION

Transposable elements are important contributors to tumorigenesis and provide supplementary means by which gene expression can be altered in cancer. While many studies have used a hypothesis-driven approach and focused at specific TEs or their subfamilies for discovering TEs that alter the expression of nearby genes in cancer [27-29], applying a genome-wide scan could allow to obtain a better picture of the effects of TEs on gene expression during tumorigenesis.

Here, we present a global profile of tumor drivers and show that TEs are highly prevalent mediators of genetic effects on genes altering their expression, specifically in tumor. By combining genome and transcriptome data together we, show that TEs are under tight genetic control and discover that transcription factors regulate TE expression much more in tumor than in normal. By looking at the interplay between eQTL variants, transposable elements, and gene expression, we are able to dissect eQTL effects and show that for several genes, the genetic effect of an eQTL is passed on genes through TEs which act as mediators and drive gene expression. We observe this to occur significantly more in cancer than in normal and show that the majority of TEs increase the expression of affected nearby genes. Interestingly, we discover that TEs affecting known cancer driver genes in cancer have for most part no significant effect
on these genes in normal suggesting a tumor-specific effect of these TEs. Additionally, in our study we show that alongside predisposing alleles and somatic mutations, germline variants are crucial contributors to tumorigenesis as these allow for transcriptional changes to occur at the level of TEs that in turn result in altered expression of nearby genes in cancer as shown previously [11].

It is known that TEs are much more active in tumor than in normal, primarily due to a global hypomethylation in cancer driving their expression [26]. However, in our analysis we observe a higher number of eQTLs in normal than in tumor which may sound contradictory. This has to do with the nature of the tumor tissue being much more heterogeneous increasing the variance in the expression data subsequently affecting statistical power, leading to fewer eQTLs being discovered. By increasing sample size we could minimize this problem and increase the eQTL discovery in tumor.

To assess the function of these eQTLs, we used functional enrichment analysis. Even though we discover more eQTLs in normal, we observe that tumor TE-eQTLs are significantly enriched for more transcription factors compared to normal TE-eQTLs and the higher number of TFs significantly associated with tumor-specific TE-eQTLs indicates that TEs are more active in cancer compared to normal. Interestingly, we observed that for 61 tumor-specific TE-eQTLs, the eQTL loci is not an eQTL for any nearby gene or TE (±1Mb) in the normal state. This indicates that these loci are probably inactive in normal and get activated during tumorigenesis, driving the expression of nearby TEs, specifically in cancer. Interestingly, we observed significant difference between DNA methylation changes at tumor-specific TE-eQTLs than shared TE-eQTL loci (eQTL active in both normal and cancer) which pinpoints that DNA methylation changes at these specific loci to be one of the causes of the activation of these eQTLs in cancer.

While we focused on TEs impacting the expression of nearby genes in an independent manner, it is highly plausible that synergistic effects occur from both cis- and trans-acting TEs. Performing such an analysis could give a fuller picture of the regulatory network behind the regulation of gene expression through TE effects, requiring, however, a high sample size for sufficient statistical power. Nevertheless, because of the highly repetitive nature of transposable element sequences and their evolutionary relatedness among TE families, mapping short reads originating from TEs is a real challenge [18, 30]. Our RNA-seq dataset having a read length of 49bp, it is highly possible that we did not map all expressed TEs subsequently leading to missing information, as shown previously [18, 30]. Future studies where RNA-sequencing is performed with longer read lengths could allow for better mapping of expressed TEs and give us a fuller picture of the number of these driver TEs in cancer.

Altogether, we have outlined that TEs are important mediators of genetic effects onto genes that could potentially be used as risk factors or new therapeutic targets for future drug development and aid in cancer treatment.
METHODS

1.1. SYSCOL dataset

The Systems Biology of Colorectal cancer (SYSCOL) dataset (accession code: EGAS00001000854) contains data from genotypes and RNA-sequencing for matched normal-tumor samples (i.e., both tumor and normal samples originate from the same patient). Samples that had genotype data and molecular phenotype quantifications from tumor and normal (normal adjacent to tumor) tissue were analyzed, yielding 275 normal samples and 276 tumor samples. In case of multiple tumor samples from the same patient, only samples with quantifications from the most advanced tumor were kept.

1.2. Genotypes

We used imputed genotypes and only kept variants with a minor allele frequency (MAF) >=5%, yielding a total of 6,132,240 variants that were used for all downstream analyses.

1.3. Transcriptome quantifications

1.3.1. Read mapping

SYSCOL samples were sequenced using 49bp, 75bp and 100bp read lengths using paired-end non-stranded mRNA-sequencing. We first started by trimming all samples with 75bp (N=73) and 100bp (N=4) reads down to 49bp to reduce any bias in downstream analysis stemming from read length. For this we used cutadapt [31] with the following command “cutadapt -u -Nreads -o <output_file> <input_file>”. All trimmed samples were mapped to the human reference genome (hg37) using hisat2 [32].

1.3.2. Transposable elements (TE) and genes quantifications

Gene and transposable element counts were generated using the featureCounts software [33]. We provided a single annotation file in gtf format to featureCounts containing both genes and transposable elements. This prevents any read assignation ambiguity to occur. For transposable elements, we used an in-house curated version of the Repbase database [12] where we merged fragmented LTR and internal segments belonging to a single integrant. We only used uniquely mapped reads for gene and TE counts. Molecular phenotypes that did not have at least one sample with 20 reads and for which the sum of reads across all samples was lower than the number of samples, were discarded. Furthermore, we normalized molecular phenotypes (TEs and genes) for sequencing depth using the TMM methodology as implemented in the limma package of Bioconductor [34] and used gene counts as library size for both TEs and genes. Finally, we removed any molecular phenotype that had more than 50% of missing data (zeros) in tumor and normal samples separately and took the union of molecular phenotypes, yielding 17,430 genes and 50,921 TEs for a total of 68,351 molecular phenotypes.
1.3.3. Normalization of molecular phenotypes

The observed variability in molecular phenotypes from RNA-sequencing data can be of biological or technical origin. To correct for technical variability, while retaining biological variability, we residualised the molecular phenotype data for the covariates as described below:

1. To correct for population stratification that is observed between the SYSCOL samples, we used Principal Component analysis (PCA) results obtained from genotypes of SYSCOL patients. We only retained the first three principal components (PCs) as covariates.

2. In order to capture experimental/technical variability in the expression data, we performed PCA, centering and scaling, using pca mode from QTLtools software package [35]. To ascertain the number of PCs that capture technical variability, we used QTL mapping (see method 3.4.1 for the description of QTL mapping) to identify the best eQTL discovery power in both tumor and normal samples. To this end, we carried out multiple rounds of eQTL mapping for tumor and normal samples separately, each time using the 3 PCs from genotypes and incrementally adding 0, 1, 2, 5, 10, 20, 30, 40, 50, 60 and 70 PCs as covariates. This approach resulted in identifying 30 PCs in tumor and normal samples for maximizing eQTL discovery.

In total, 33 covariates were regressed out from tumor and normal sample expression data using QTLtools correct mode [35]. We additionally rank-normalized on a per phenotype basis across all samples such that quantifications followed normal distribution with mean 0 and standard deviation 1 N(0,1) using QTLtools --normal option [35].

1.4. DNA Methylation data and differential methylation of eQTLs

We used microarray based DNA methylation data from the SYSCOL project and a similar approach to a previous study to find differential methylation of eQTLs [11]. In brief, we calculated the absolute value difference of the medians of normalized methylation probe betas in normal and tumor that we call median differential methylation. We then compared the distribution of their medians in tumor-specific TE and genes eQTLs vs. the shared TE and gene eQTLs and calculated a P-value using the Mann Whitney U test. P-values were corrected for multiple testing using the R/qvalue package with a given FDR threshold of 5%.

1.5. Differential TE/gene expression analysis

The DESeq2 R package [36] was used in calculating differentially expressed genes and TEs. We normalized the raw TE/gene counts within the DESeq2 package using the sequencing date, GC mean and insert size as covariates. The differential expression P-values were corrected for multiple testing using an FDR threshold of 5%.

1.6. Transcriptome QTL analysis

All analyses were performed separately for normal and tumor samples. We used imputed genotypes with MAF >=5 %, gene expression data with normalized counts per million (CPMs) (as described above) for both eQTL and conditional eQTL mapping.
1.6.1. Expression Quantitative Trait Loci (eQTL) mapping

For eQTL mapping, we used cis mode of the QTLtools software package [35]. For each molecular phenotype:

1. We counted all genetic variants in a 1 Mb window (+/- 1 Mb) around the transcription start site (TSS) of the phenotype and tested all variants within this window for association with the phenotype. We only retained the best hits which are defined as the ones with the smallest nominal p-value.

2. Next, we used permutations to adjust the nominal p-values for the number of variants tested. More specifically, we randomly shuffled the quantifications of the phenotypes 1’000 times and retained only the most significant associations. This created a null distribution of 1’000 null p-values. Then, we fitted a beta distribution on the null distribution and used the resulting beta distribution to adjust the nominal p-value. Principally, this strategy allows to quantify the chance of getting a smaller p-value than the nominal one in random datasets.

This effectively gave us the best variant in cis together with the corresponding adjusted p-value of association for each molecular phenotype. Finally, to correct for the number of phenotypes being tested we used False Discovery Rate (FDR) correction approach. More specifically, we used the R/qvalue package [37] to perform genome-wide FDR correction which ultimately facilitated to extract all phenotype-variant pairs that are significant at a pre-determined FDR threshold, 5% FDR in this case.

1.6.2. Conditional analysis for eQTL mapping

The cis mode informs us on the best phenotype-variant pair only. Given that the expression of molecular phenotypes can be affected by multiple cis eQTLs, we proceeded with conditional analysis to discover all eQTLs with independent functional effects on a given phenotype. Principally, new discoveries are made after conditioning on previous ones. Again, cis mode in the QTLtools software package was used [35]. In brief, after running permutations (method 1.4.1) for each phenotype, we calculated a nominal p-value threshold of being significant. We first determined the adjusted p-value threshold that corresponds to the targeted FDR level and then used the beta quantile function to go from adjusted p-value to a specific nominal p-value threshold. For conditional analysis, forward-backward methodology is used to discover all independent QTLs and to identify the most likely candidate variants, while at the same time controlling for a given FDR (5% FDR in this case). We only kept the top variant for each signal.

1.6.3. Tissue-specific and shared eQTL analysis

To discover tissue specific and shared eQTLs, we used the eQTL results obtained after running the conditional pass. In total, we tested 17,186 eQTLs to discover normal-specific eQTLs and 6,751 to discover tumor-specific eQTLs. To do that, we used linear mix models using an interaction term between dosage and tissue (i.e tumor or normal) to test whether the slopes in normal and tumor are significantly different. Linear mix models are needed here because normal and tumor samples are originating from the same patient thus genotypes will be identical. We did this for tumor and normal eQTLs separately. Then we performed multiple test correction using the R/qvalue package [37] with a given FDR threshold of 5%. Additionally, for all significant results at 5% FDR, if eQTL slopes (slopes given from conditional QTL
mapping using QTLtools) in normal and tumor had the same direction, then we only kept the ones where the SNP-phenotype association in the opposite tissue was not nominally significant (P>0.05) as given by the cis nominal pass mode in the QTLtools package [35].

Shared eQTLs are defined as the ones where the P-value for the interaction term is not significant but need to be significant eQTLs (5% FDR) in both normal and tumor as assessed by the conditional QTL mapping.

1.6.4. Functional enrichment analysis

To compare the QTL variants to a null distribution of similar variants without regulatory association, we sampled for each eQTL variant 100 random regulatory genetic variants matching for relative distance to TSS (within 2.5kb) and minor allele frequency (within 2%) and only kept variants that are not eQTLs for any other TE or gene (nominal p-value > 0.05). The enrichment for a given category was calculated as the proportion between the number of regulatory associations in a given category and all regulatory variants over the same proportion in the null distribution of variants. The p-value for this enrichment is calculated with the Fisher exact test. Finally, we corrected for multiple testing using an FDR threshold of 5% using the “p.adjust” function in the R programming language. The code for performing the functional enrichment analysis can be accessed here: https://github.com/NLykoskoufis/fenrichcpp.

1.6.4.1. Ensembl Regulatory Build ChIP-seq dataset

ChIP-seq data was downloaded from the FTP site (http://ftp.ensembl.org/pub/grch37/current/regulation/homo_sapiens/). The dataset contains ChIP-seq data from 88 human cell types for a total of 209 transcription factors and 29 histone marks (build hg19). For each of the TFs and histone marks, we took the union of all peaks together from all 88 cell types. Overlapping peaks were merged together using the “merge” options in the BEDtools software [38]. This allowed us to create an extensive annotation of peaks for 209 TFs and 29 histones genome-wide.

1.6.4.2. Colorectal cancer LoVo cell line ChIP-seq dataset

We used publicly available ChIP-seq data from colorectal cancer LoVo cell line with accession code GSE49402. The dataset comprises of 202 TFs and 2 histone marks (build hg19). We used BED files containing the coordinates of the peaks for each TF and histone mark for functional enrichment of our eQTLs.

For gene and TE eQTLs in normal and tumor, we used the peak annotation generated from the Ensembl Regulatory Build data to get an extensive comprehension of which TFs regulate the expression of TEs. Regarding the tumor-specific vs. shared TE and gene eQTLs, we used available ChIP-seq data from the colorectal cancer LoVo cell line [19]. We used a cancer specific dataset as we were interested in discovering cancer specific effects.
Testing for associations between TEs and genes

To discover associations between TEs and genes, we proceeded in a similar way to what we did for QTL mapping (method 1.4.1). Effectively, we used TE expression as our “genotypes” and genes as our phenotype. Then, we corrected for multiple testing using the \textit{R/qvalue} package with a given FDR of 1%. We then estimated the nominal p-value thresholds for each phenotype being tested as described in (method 1.4.2) with a given FDR of 1%. Given the nominal threshold we get for each gene, we then extracted all TEs with an association P-value below this threshold which could give multiple TEs for a gene in some cases.

1.7. Quantifying TE-gene pairs

To quantify each of TE-gene pairs that have been found to be significant, we used a dimensionality reduction approach based on PCA as previously described [22]. Specifically, for each TE-gene pair, we aggregated gene expression together with TE expression by using the coordinates on the first PC. This effectively built a quantification matrix with rows and columns corresponding to the number of TE-Gene pairs and individuals, respectively. All quantifications have been rank-normalized on a per phenotype basis so that the values match a normal distribution $N(0,1)$. This prevents outlier effects in downstream association testing. This is all implemented in the \textit{clomics} software package [22].

1.8. Causal inference by Bayesian networks for QTL-TE-Gene triplets

Bayesian networks (BNs) are a type of probabilistic graphical model that uses Bayesian inference to compute probabilities. BNs aim to model conditional dependencies and therefore causation by representing conditional dependencies as edges and random variables as nodes in a directed acyclic graph. The flow of information between two nodes is reflected by the direction of the edges, giving an idea of their causal relationship. BNs have been previously used in a genetic framework [20] to get insight into the most likely network from which the observed data originates.

In BNs, the joint probability density can be divided into marginal probability functions and conditional probability functions for the nodes and edges, respectively. Additionally, BNs satisfy the local Markov property where each variable is conditionally independent of its non-descendants given its parent variables. In the context of this study, we used BNs to learn the causal relationships between triplets of variables, each one containing a genetic variant, a transposable element and a gene. In practice, only three distinct network topologies where relevant to the hypotheses we wanted to test (Supplementary figure 12). More specifically, we looked at:

1. The causal scenario where the genetic variant affects first the TE and then the gene.
2. The reactive scenario where the genetic variant affects the gene first and then the TE.
3. The independent scenario in which the variant affects the gene and the TE independently.

Of note, we only retained network topologies that assume that the signal systematically originates from the genetic variant. In practice, we applied BNs on data that was obtained from running an QTL mapping using the TE-gene pairs using a similar approach to QTL mapping described above (Method 1.4.1) and only kept significant results at 5% FDR which corresponds to 11,937 QTL-TE-gene triplets in normal and 9,528 QTL-TE-gene triplets in tumor.
For each triplet, we build a 275 x 3 matrix in normal and 276 x 3 matrix in tumor containing normalized quantifications and used it to compute the likelihood of the 3 BN topologies using the R/bnlearn package (Version 4.5) [39]. As a last step, we went from likelihoods to posterior probabilities by assuming a uniform prior probability on the three possible topologies. Posterior probabilities where used for all BN-related analyses.

1.9. Transposchimeric transcripts analysis

First, a per sample transcriptome was computed from the RNA-seq bam file using StringTie [40] with parameters –j 1 –c 1. Each transcriptome was then crossed using BEDTools [38] to both the ensembl hg19 coding exons and curated RepBase [12] to extract TcGTs for each sample. Second, a custom python program was used to annotate and aggregate the sample level TcGTs into counts per groups (normal, tumor). In brief, for each dataset, a GTF containing all annotated TcGTs was created and TcGTs having their first exon overlapping an annotated gene or TSS not overlapping a TE were discarded. From this filtered file, TcGTs associated with the same gene and having a TSS 100bp within each other were aggregated. Finally, for each aggregate, its occurrence per group was computed.

2. GTEx dataset

We downloaded available data for colon transverse (N=174) and germline genotypes from dbGAP (accession code: phs000424.v8.p2).

2.1. Germline genotypes

We used the already filtered VCF file provided by GTEx. The following filters were applied and kept all variants with a MAF ≥ 5%, yielding a total of 6,494,417 variants.

2.2. RNA-seq dataset

The RNA-seq dataset was treated similarly to SYSCOL RNA-seq data. We first trimmed the reads down to 49bp using cutadapt [31]. Then we mapped and quantified the samples using the exact same approach as for SYSCOL (methods section 1.3.2). Finally, we combined all samples together into a multi-sample bed file and kept all features (TEs, genes) that had less than 50% of missing expression data across all samples, yielding a total cd of 167,429 TEs and 18,472 genes. Then, we corrected our expression data using the first 3 principal components (PCs) obtained from genotypes, the sex of the samples, the platform they were sequenced and the first 20 PCs obtained from the expression data, for a total of 25 covariates used.

3. TCGA dataset

We downloaded available germline genotypes and RNA-seq data for colon adenocarcinoma (N=251) from The Cancer Genome Atlas (TCGA) database, accession code phs000178.v11.p8.

3.1. Germline genotypes

Germline genotypes were downloaded from the legacy archive GDC portal. We downloaded all germline genotypes for TCGA-COAD in birdseed format. We used birdseed2vcf python
tool (https://github.com/ding-lab/birdseed2vcf) to convert birdseed to VCF format. We then combined all samples together creating a multi-sample VCF file that we spitted per chromosome and uploaded to the Michigan Imputation Server [41] for imputation and phasing using the Haplotype Reference Consortium (HRC) as reference panel, Eagle v2.4 software [42] for phasing and European (EUR) population. Finally, we merged all chromosome VCFs into a single VCF file and kept variants with a MAF ≥ 5%, HWE > 1e-06 and $R^2$ > 0.3, yielding a total of 5,511,779 variants.

#### 3.2. RNA-seq data

As the read length of TCGA-COAD samples is the same as SYSCOL, we did not need to trim the reads. We mapped, quantified, and filtered our RNA-seq data in a similar way as for SYSCOL and GTEx colon transverse samples yielding a total of 19,376 genes and 75,815 TEs. Expression data was corrected using the same approach as for SYSCOL (methods section 1.3.3) using the first 3 principal components (PC) obtained from genotypes and the first PC obtained from expression data for a total of 4 covariates used.

#### 4. Replication of eQTL findings

For the replication of our normal and tumor eQTL discoveries, we used the “rep” mode in the QTLtools software [43]. We then used the pi1 metric to estimate the proportion of significance of our eQTLs in GTEx colon transverse. The pi1 is equal to 1 – pi0 where pi0 is the proportion of true null p-values obtained using pi0est from the Qvalue R package [44].

#### 5. Replication of the eQTL – TE – gene triplets

We used the same eQTL – TE – gene triplets discovered in normal and tumor and replicated them in GTEx or TCGA-COAD, respectively. We used the exact same approach as previously (methods section 1.8). We then calculate the mean probability of the causal, reactive and independent model. Finally, we compared the percentage of triplets with the same model predicted in both SYSCOL and the replication dataset.
AUTHOR CONTRIBUTIONS

N.M.R.L, H.O and E.T.D designed the study. N.M.R.L analyzed the data and wrote the manuscript and N.M.R.L, H.O, D.T and E.T.D interpreted the results. E.P shared the quantifications data.

COMPETING INTERESTS

Emmanouil T. Dermitzakis is currently an employee of GSK. The work presented in this manuscript was performed before he joined GSK. All other authors declare no competing interests.

ACKNOWLEDGMENTS

The computations were performed at the University of Geneva on the Baobab cluster. This work was supported by grants from Louis-Jeantet Foundation support (to E.T.D) and SNSF grant (to E.T.D), by grants from the European Research Council (KRABnKAP, No. 268721; Transpos-X, No. 694658), the Personalized Health and Related Technologies (PHRT-508) program (to D.T). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

DATA AVAILABILITY

The RNA-sequencing and genotype data are deposited in the European Genome-phenome Archive (EGA, https://www.ebi.ac.uk/ega/) for controlled accesses; the study accession number is EGAS00001000854.

CODE AVAILABILITY

All custom scripts used can be accessed here: https://github.com/NLykoskoufis/te_project. The code for the functional enrichment analysis can be accessed here: https://github.com/NLykoskoufis/fenrichcpp.
REFERENCES


Figure 1: Description of quantified TEs. (A) Barplot showing the proportion of uniquely mapped and quantified TE subfamilies in our dataset. (B) Pie chart showing the proportion of TEs with different types of regulatory elements within their sequence. We uniquely mapped and quantified 50,921 TEs. The majority of them are SINEs from the Alu and MIR family, L1 and L2 TEs from the LINE family and different subfamilies of LTRs as well as some DNA transposons. When we looked at the proportion of expressed TE per subfamily, we observed that SVA and ERVK are most prominent. Additionally, 13,590 out of the 50,921 TEs contain regulatory elements within sequence.
Figure 2: cis- eQTL discovery. eQTL variant distance to TSS in (A) normal and (B) in tumor. We observe stronger eQTL effect close to the transcription start site of TE and genes in both normal and tumor. Number of secondary eQTLs for TEs and genes in (C) normal and (D) tumor. Gene eQTLs have more functionally independent eQTLs per gene than TEs do.
Figure 3: Tissue specificity of TE-eQTLs. (A) Mosaic plot of tissue specificity of TE-eQTLs. (B) Tissue specificity and distance of TE-eQTL to transcription start site (TSS). The shared TE-eQTLs (black) are closer to the TSS than are the tissue specific TE-eQTLs (red) (Wilcoxon P<2.2e-16). (C) TE-eQTL slopes for the normal specific TE-eQTLs in blue, the tumor specific in red and shared in black. (D) Boxplot of the absolute value difference of median methylation betas between normal and tumor samples for shared and tumor-specific TE-eQTLs.
Figure 4: Functional enrichment of eQTLs. (A) The ratio between TE-eQTL enrichment and gene-eQTL enrichment in log2 scale discovered in normal. 5 TFs show stronger enrichment for TE-eQTLs in normal compared to gene-eQTLs. (B) The ratio between TE-eQTL enrichment and gene-eQTL enrichment in log2 scale discovered in tumor. We observed 16 TFs to have a stronger enrichment for TE-eQTLs than gene-eQTLs in normal. (C) log2 ratio between tumor-specific TE-eQTL enrichment and shared TE-eQTL enrichment. We observe 60 TFs with a stronger enrichment for the tumor-specific TE-eQTLs than the shared eQTLs indicating that these TFs regulate TE expression specifically in tumor.
Figure 5: Causal relationship between eQTL variants, TEs and genes. (A) Barplot representing the mean probability for each of the three models in normal and tumor. We observe significantly more causal cases in tumor compared to normal (Wilcoxon P-value < 2e-16) (B) Barplot representing the model substitutions for the 9,528 tumor triplets from normal to tumor. Independent models tend to shift to a causal in tumor. This is true also for the reactive models in normal but to a much smaller extent. (C) Barplot representing the number of triplets that do not switch models, that switch to a causal model or that switch to reactive/independent from normal to tumor. The majority of triplets do not switch models between normal and tumor. However, 2,584 triplets are switching to a causal model making the corresponding TEs potential drivers of gene expression (D) Each point represents a TE-gene for each of the 2,584 tumor triplets. All points are significant in tumor but not in normal (grey points). We observe that in most cases, TEs are positively correlated with genes except for a few cases. Most cancer driver genes have no significant correlation with any TE in normal indicating that for most part, TEs impact them specifically in tumor.
Figure 6: Tumor-specific and shared TE-eQTLs effects. (A) The barplot represents the frequency of the causal, reactive and independent model for the triplets with shared or tumor-specific TE-eQTLs. (B) The barplot represents the model changes from normal to tumor for the triplets constituted of shared or tumor-specific TE-eQTLs. (C) Barplot that represents the number of tumor-specific TE-eQTLs that are inactive eQTLs for the triplet associated gene.
Supplementary information

Transposable elements mediate genetic effects altering the expression of nearby genes in colorectal cancer

Nikolaos M. R. Lykoskoufis¹,²,³*, Evarist Planet⁴, Halit Ongen¹,²,³,∗, Didier Trono⁴,∗ and Emmanouil T. Dermitzakis¹,²,³,*

¹ Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland
² Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, 1211 Geneva, Switzerland
³ Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland
⁴ School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015, Lausanne, Switzerland

* Corresponding authors: N.M.R Lykoskoufis: nikolaos.lykoskoufis@unige.ch and E.T.D: emmanouil.dermitzakis@unige.ch
Department of Genetic Medicine and Development, University of Geneva, 1 rue Michel-Servet, 1211 Geneva, Switzerland

** Co-last authors
Supplementary figures

Supplementary figure 1 | Proportion of TE subfamilies in the human genome. The majority of transposable elements in the human genome are Alu and tRNA from the SINE family, followed by L1 and L2 elements from the LINE family and ERV1, ERVL and ERVL-MaLR from the LTR family.
Supplementary Figure 2 | TEs overlapping regulatory regions in the human genome. Pie plot representing the proportion of transposable elements overlapping with each regulatory regions. Of the ~4.6 million TEs in the human genome, 820,981 TEs are overlapping with at least one regulatory region.
Supplementary figure 3 | cis eQTL discovery p-value distribution. Histograms of p-value distribution of cis- eQTL discovery in (A) normal and (B) tumor.
Supplementary figure 4 | P-value distributions of significant SNP-gene or SNP-TE pairs tested in the other tissue. The $\pi_1$ statistic estimates the tissue sharing of eQTLs. We observe that SNP-TE and SNP-gene pairs discovered in tumor are replicated better in normal than normal eQTLs tested in tumor.
Supplementary figure 5 | TE and gene eQTL distance to TSS in normal and tumor. We observe that in both normal and tumor, TE-eQTLs are closer to the TSS of TEs than gene-eQTLs are to the TSS of genes. This could be because of the smaller evolutionary time TEs have in the human genome compared to genes, making local effects much more likely to occur.
Supplementary figure 6 | eQTL variant allele frequencies. We observe that the allele frequencies for the TE- and gene-eQTLs are very similar in normal and tumor.
Supplementary figure 7: replication of our eQTL findings in external datasets. Figure representing the p-value distribution of the replication of SYSCOL eQTLs. Panels A to C represent the replications of SYSCOL normal eQTLs in GTEx colon transverse, where (A) represents the p-value distribution of gene- and TE-eQTLs taken together, (B) the p-value distribution of SYSCOL normal TE-eQTLs and (C) the p-value distribution of SYSCOL normal gene-eQTLs. Panels D to F represent the replication of the SYSCOL tumor eQTLs in colon adenocarcinoma (TCGA-COAD), where (D) is the p-value distribution of all gene- and TE-eQTLs taken together, (E) the p-value distribution of SYSCOL tumor TE-eQTLs and (F) the p-value distribution of SYSCOL tumor gene-eQTLs. We observe that in both normal and tumor TE-eQTLs (p1 normal = 83.1%; p1 tumor = 88.4%) and gene-eQTLs (p1 normal = 68.5%; p1 tumor = 78.3%) have a very high replication in GTEx colon transverse and TCGA-COAD, respectively.
Supplementary figure 8 | Tissue specificity of Gene-eQTLs. (A) Mosaic plot of tissue specificity of Gene-eQTLs. (B) Tissue specificity and distance of Gene-eQTL to transcription start site (TSS). The shared Gene-eQTLs (black) are closer to the TSS than are the tissue specific Gene-eQTLs (red) (Wilcoxon P<2.2e-16). (C) Gene-eQTL slopes for the normal specific Gene-eQTLs in blue, the tumor specific in red and shared in black. (D) Boxplot of the absolute value difference of median methylation betas between normal and tumor samples for shared and tumor-specific gene-eQTLs.
Supplementary Figure 9 | Functional enrichment for gene- and TE-eQTLs in normal. Enrichment over the null for gene- and TE-eQTLs in normal. Only plotting cases where the enrichment was significant at 5% FDR for either gene- or TE-eQTLs.
Supplementary Figure 10 | Functional enrichment for gene- and TE-eQTLs in tumor. Enrichment over the null for gene- and TE-eQTLs in tumor. Only plotting cases where the enrichment was significant at 5% FDR for either gene- or TE-eQTLs.
Supplementary figure 11 | Functional enrichment for Tumor-specific and shared gene-eQTLs. (A) Enrichment over the null for tumor-specific and shared gene-eQTLs. (B) log2 ratio between tumor-specific enrichment and shared enrichment. Only plotting cases where the enrichment was significant at 5% FDR for either tumor-specific or shared gene-eQTLs. None of the TFs have a stronger enrichment for tumor-specific gene-eQTLs indicating that these TFs regulate gene expression in both the normal and tumor state.
Supplementary figure 12 | Functional enrichment for tumor-specific and shared TE-eQTLs. (A) Enrichment over the null for tumor-specific and shared TE-eQTLs. Only plotting cases where the enrichment was significant at 5% FDR for either tumor-specific or shared TE-eQTLs.
Supplementary figure 13 | Correlation of differential enrichment of functional binding sites and differential expression of the corresponding transcription factors. We compared the ratio of tumor expression over the normal expression of differentially expressed transcription factors to the ratio of enrichment for the binding sites of the same transcription factors in tumor specific TE-eQTLs over the shared TE-eQTLs. We find no significant correlation indicating that differential expression of the corresponding TFs do not drive the tumor-specific TE-eQTLs.
Supplementary figure 14 | Tumor-specific TE-eQTLs overlapping with the enriched transcription factors and their activity in normal regarding other genes/TEs. We checked whether any of the tumor-specific TE-eQTLs overlapping with the transcription factors we found to have a stronger enrichment for tumor-specific TE-eQTLs are active or inactive eQTLs for other TEs or genes in cis. We discovered that 61 of them are not significant eQTLs for any TE or gene in normal indicating that these regions are probably inactive and get activated in tumorigenesis.
Supplementary figure 15 | Mean number of genes associated per TE in (A) normal and (B) tumor. We observe that in normal there are less genes associated per TE (mean number of associated genes per TE = 1.4) compared to tumor (mean number of associated genes per TE = 3.4)
Supplementary figure 1 | TE-gene effect sizes (regression slope) in (A) normal and (B) tumor. We observe that most TEs are positively associated with a gene.
Supplementary figure 17 | P-value distribution of eQTL TE-gene discovery in (A) normal and (B) tumor. We observe that in normal (A) we discovered 11,937 triplets and in (B) tumor 9,528 triplets at 5% FDR.
Supplementary figure 18 | Causal relationship of eQTLs, TEs and genes approach. To infer the most likely causal relationship between eQTL variants, TEs and genes, we tested three models using Bayesian Networks (BNs). The causal model where the eQTL variant affects the TE and then the gene, the Reactive model where the eQTL variant affects the gene and then the TE and the Independent model where the eQTL variant affects the TE and the gene independently. For each triplet we obtained log likelihoods and we calculated posterior probabilities using a uniform prior probability for each of the three models.
Supplementary figure 19 | Probability of the most likely model in (A) normal and (B) tumor. We can see that the probabilities of the most likely models are for most part above 0.8.
Supplementary figure 20 | Causal relationships depending on the genomic position of the TE in respect to the gene. In each case, we worked out the percentages by averaging the posteriors given by the Bayesian networks across all the triplets falling in each of the categories. We observe that TEs inside genes or downstream of genes tend to react to gene expression whereas TEs outside genes or when the eQTL variant is within the TE sequence, TEs are most likely causal for changes in gene expression.
Supplementary figure 21: This figure represents the replication of the causal inference findings in external datasets. Model probabilities of the causal, reactive and independent models are represented (A) for the 9,577 SYSCOL normal triplets tested in GTEx colon transverse, (B) for the 12,379 normal triplets tested in SYSCOL, (C) for the 9,714 tumor triplets tested in SYSCOL, (D) for the 5,893 SYSCOL tumor triplets tested in TCGA-COAD. (E) represents the percentage of SYSCOL normal triplets with the same model predicted in SYSCOL and in GTEx transverse. (F) represents the frequency of substitutions between the normal triplets in SYSCOL and GTeX transverse. (G) represents the percentage of tumor triplets with the same model predicted in SYSCOL and in GTEx colon transverse. (H) represents the frequency of substitutions between the tumor triplets in SYSCOL and TCGA-COAD. We observe a high replication of our causal inference findings in both GTEx colon transverse and TCGA-COAD. Moreover, because the sample size of GTEx colon transverse is smaller (N=174) compared to the one of SYSCOL normal (N=275), we expected a smaller replication of our finding. We observe that in TCGA-COAD, the percentage of replication is higher as the TCGA-COAD and SYSCOL tumor dataset have almost the same sample size (N_{SYSCOL tumor}=276; N_{TCGA-COAD} = 251), thus higher statistical power.
Supplementary figure 22 | Model shifts between normal and tumor. Model shifts for shared triplets between normal and tumor (A-B) where (A) represent the percentage of shared triplets tested that show either a higher posterior probability for the causal, reactive or independent model in tumor and normal. (B) Represents the model shifts from normal to tumor. (C-D) represent the same as A and B but for the union of normal and tumor triplets. In both shared and union, we observe an increase of the causal model in tumor. Independent models in normal and to a smaller extent reactive models are shifting for a causal model in tumor.
Supplementary figure 23 | log2 fold change of TEs switching to causal versus TEs that do not switch or switch but not to causal. We observe that TEs switching to causal in tumor are significantly more upregulated compared to TEs not switching or switching but not to causal, indicating that this upregulation could explain some of the cases where TEs switch to causal in tumor.
CHAPTER 3

The interplay between regulatory variants, transposable elements and gene expression in adult human tissues

Contribution:

Nikolaos M.R. Lykoskoufis designed and performed all the analyses, prepared all the figures and wrote the manuscript with contributions from all authors. Nikolaos M.R Lykoskoufis is the first and corresponding author of the manuscript.

This chapter is a manuscript in preparation for submission.

Abstract

Transposable elements (TEs) are DNA repetitive sequences that have been shown by a number of studies to be transcriptionally active in adult human tissues. Recently, we demonstrated that TEs are under tight genetic control in normal colon and colorectal cancer and that they are crucial mediators of genetic effects on neighbouring genes. Following these findings, we used the Genotype-Tissue Expression (GTEx) initiative to investigate the genetic regulation of TEs and their transcriptional activity in 43 different adult human tissues. We found hundreds to thousands of TE-expression Quantitative Trait Loci (eQTLs), showing that TE expression is genetically regulated. We discovered a number of TFs and histone marks to be significantly more enrichment for TE-eQTLs compared to gene-eQTLs highlighting that these TFs are more specific towards controlling TE expression than gene expression. Using Bayesian Networks, we assessed the causal inference between eQTL variants, TEs and genes and discovered 28,828 TEs responsible for changes in the expression of 12,599 nearby genes, across the 43 studied tissues. We also demonstrated that TEs impacting gene expression tend to share these effects throughout closely related tissues. Together, these findings show that TEs have a useful function in the transcriptional control of neighbouring genes and provide a thorough atlas of TE-eQTLs that may be exploited to investigate the function of TEs in complex traits and diseases.

Supplementary data for this chapter can be accessed by clicking the following link: supplementary data.
The interplay between regulatory variants, transposable elements and gene expression in adult human tissues

Nikolaos M. R. Lykoskoufis¹²³, Evarist Planet⁴, Didier Trono⁴, Halit Ongen¹²³

¹ Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland
² Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, 1211 Geneva, Switzerland
³ Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland
⁴ School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), 1015, Lausanne, Switzerland

* Corresponding authors: N.M.R Lykoskoufis: nikolaos.lykoskoufis@unige.ch, Halit Ongen: halit.ongen@unige.ch

Department of Genetic Medicine and Development, University of Geneva, 1 rue Michel-Servet, 1211 Geneva, Switzerland
Abstract

Mobile repetitive DNA sequences, called transposable elements (TEs) make up more than half of the human genome. Several studies have showed that transposable elements are more transcriptionally active in adult human tissues than previously thought. Furthermore, we recently discovered that TEs are under genetic control in colorectal cancer patients, where they play important roles as mediators of genetic effects on nearby genes in cancer. Here, we wanted to expand our findings and used the Genotype-Tissue Expression (GTEx) project to study the transcriptional activity and genetic control of TEs in 43 different non-diseased human adult tissues. We discovered hundreds to thousands of TE-expression Quantitative Trait Loci (eQTLs) indicating that TE expression is under tight genetic control. We identified several transcription factors and histone modifications with a stronger enrichment for TE-eQTLs compared to gene-eQTLs in several tissues highlighting that they could regulate TE expression more specifically compared to gene-expression. Moreover, to evaluate the causal relationship between eQTL variants, transposable elements, and genes, we used Bayesian Networks and identified across all tissues a total of 28,828 TEs that are causal for changes in the expression of 12,599 nearby genes. We also showed that TE effects on gene expression are shared among closely related tissues, with brain specific TE-eQTLs being the most abundant. Collectively, our results highlight a functional role for TEs in the transcriptional regulation of nearby genes and serve as a comprehensive atlas of TE-eQTLs that can be used to study the role of TEs in complex human traits and disease.
INTRODUCTION

Transposable elements (TEs) are repetitive DNA sequences that contribute to more than half of the human genome. TEs come in all shapes and forms and the easiest way to characterize them is through their mechanism of transposition. There are two main classes of TEs, retrotransposons and DNA transposons. DNA transposons represent approximately 3% of the human genome with a rough estimate of 500,000 DNA transposon integrants [1]. DNA transposons spread via a cut-and-paste mechanism. On the other hand, retrotransposons spread through a copy-and-paste mechanism where they are first transcribed, then reverse transcribed and finally integrated into a new genomic location [1]. TEs harbor transcription factor binding sites and can act as transcriptional enhancers, promoters, and silencers and thereby regulate neighboring gene expression [2]. TEs, more specifically TE-embedded regulatory sequences (TEeRS) are known to be broadly active in the early embryo and the germline. Although they are mainly silenced in adult tissues, epigenetic mechanisms allow for their activation during brain development, immune responses, and metabolic regulation [1].

In 2014, Fort et al., performed a CAP Analysis Gene Expression Sequencing (CAGE-seq) to study TE activity in different human cell types and showed that endogenous retrovirus subfamilies (ERV1, HERVH-int and LTR9) were highly expressed [3]. A later study revealed that the HERVH subfamily is highly expressed in human embryonic and pluripotent stem cells with active histone marks targeting promoters and enhancers overlapping with HERVH TEs [4]. However, these studies solely focused on specific TE subfamilies or specific TE loci in the genome. A more recent study performed a genome-wide transcriptomic analysis to identify all transcriptionally active TEs and identified thousands of TEs expressed in 49 different human tissues [5]. Additionally, in a previous study, we discovered that transposable elements are under genetic control and act as key mediators of genetic effects onto nearby genes in cancer [6]. Thus, following these observations, we speculated that eQTLs could modulate TE expression in most adult human tissues. We therefore set out to analyze the interplay between eQTLs, transposable elements and expression of nearby genes as well as to characterize the causal relationship between these three genomic elements in over 40 adult human tissues. To this end, we integrated whole genome sequencing (WGS) and transcriptomic data (bulk RNA-seq) from the Genotype-Tissue Expression (GTEx) project, version 7, from 43 different human tissues and 7,727 samples in total.

RESULTS

Quantifying Transposable elements (TEs) and genes

To quantify the expression of TEs in adult tissues, we examined transcriptomes obtained by paired-end RNA-sequencing from the Genotype-Tissue Expression (GTEx) consortium. We first excluded any tissue with less than 80 samples (Supplementary Figure 1) as these do not have sufficient statistical power for all downstream analyses. After this filtering step, we ended up with a total of 43 tissues with sample sizes ranging from 80 in Brain Anterior cingulate cortex (BA24), up to 379 for Skeletal Muscle (Supplementary Figure 1). We then, quantified TE and gene expression using a TE annotation originating from the RepBase database [7] that contains approximately 4.6 million individual TE loci. This annotation was merged with gene annotation from ensembl (v19). We quantified TEs with uniquely mapped reads to obtain
robust estimates of TE expression (Methods 1.2). For all 43 tissues, we were able to quantify TE expression, with number of expressed TEs ranging from 46,553 for Whole blood to 328,133 for Testis (Figure 1.A). We observed that L1s and Alus were the most expressed subfamilies of TEs originating from the LINE and SINE family, respectively (Figure 1.B). Additionally, we used available data from Encode [8] and miRBase [9] to generate a list of regulatory regions and looked for expressed TEs from all 43 tissues that overlap with any of the aforementioned regulatory regions. We discovered that 100,703 expressed TEs overlapped with at least one of the tested regulatory regions (Figure 1.C) highlighting that TEs play a role in the regulation of gene expression.

The genetic control of transposable elements in adult human tissues

Using TE expression quantifications and WGS data, we first wanted to address the genetic control of TEs. For this, we performed cis-eQTL analyses followed by conditional analyses on a per tissue basis (Methods 2.1-2). At 5% False Discovery Rate (FDR), we discovered thousands of TE-eQTLs, ranging from 1,086 in brain hippocampus to 22,634 in thyroid and hundreds of gene-eQTLs ranging from 305 in vagina to 5,380 in thyroid. (Figure 1.A; Supplementary Table 1; Supplementary Figure 2,3). We observed that the number of eQTLs discovered per tissue was significantly correlated with sample size (for TE-eQTLs: spearman p-value =6.36e-13, ρ = 0.85; for gene-eQTLs: spearman p-value < 2.2e-16, ρ = 0.95), as previously observed [10]. As expected, TE-eQTLs showed stronger effects closer to the transcription start site (TSS) of TEs, similarly to gene-eQTLs and exhibit less functionally independent eQTLs per TE than genes do (Figure 2.B,C; Supplementary Figure 4). Pooling all eQTL discoveries from all 43 tissues together, we did not observe any significant difference between TE- and gene-eQTLs regarding the distance of the eQTL variant to the TSS of the TE or gene (Figure 2.B; Wilcoxon Rank Sum test p-value = 0.15). Both TE- and gene-eQTLs were significantly enriched for various genome regulatory elements like for example promoter regions or CTCF binding sites but depleted of enhancer regions (Supplementary Figure 5). Additionally, we did not observe any difference in allele frequency between eQTL variants associated with TEs or genes (Supplementary Figure 6). Altogether, these results showed that TEs are under tight genetic control and that non-coding germline variants regulate TE expression in a similar way to genes.

Functional characterization of TE-eQTLs

To determine whether the discovered eQTLs were functionally relevant, we performed functional enrichment analysis of TE- and gene- eQTLs in all 43 tissues using the Ensembl Regulatory Build dataset comprising of 202 Transcription factors (TFs) and 29 histone marks ChIP-seq experiments [11]. To correct for multiple testing, we adjusted p-values for a given FDR of 5% (Methods 2.4). We found significant enrichments for hundreds of TFs and histone marks at many TE-eQTL loci and gene-eQTL loci, highlighting the functional relevance of the variants discovered (Supplementary Figure 7,8; Supplementary Tables 2-3). We then searched for TFs and histone marks that displayed a stronger enrichment for TE-eQTLs compared to gene-eQTLs in the 43 tissues. For this, we calculated the log2 ratio of TE enrichment odds ratio over gene enrichment odds ratio on a per tissue basis. We observed
that most TFs exhibit a stronger enrichment for gene-eQTLs than for TE-eQTLs. However, in a fraction of tissues, we observed 91 TFs and 8 histone marks that showed stronger enrichment for TE-eQTLs compared to gene-eQTLs, indicating that these TFs and histone marks could have a higher affinity for the regulation of TE expression compared to gene expression (Figure 3; Supplementary Figure 9). Interestingly, we found that in several tissues, ZNF274, TRIM28 and across all tissues, H3K9me3 were all three showing stronger enrichment for TE-eQTLs compared to gene-eQTLs. ZNF274 is a Krüppel-associated box (KRAB) domain-containing zinc-finger protein (KZFP), and TRIM28 a corepressor that is recruited by the KRAB domain of many TE-binding KZFPs for the repression of TEs via histone H3 Lys9 trimethylation (H3K9me3), histone deacetylation and DNA methylation [12, 13]. Additionally, in a subset of tissues, BDP1 and BRF1, showed stronger enrichment for TE-eQTLs compared to gene-eQTLs. These two are subunits of the RNA polymerase III transcription initiation factor highlighting potential transcription of Alus of MIR TEs of the SINE family [14]. Altogether, these results highlight the functional relevance of TE-eQTLs. The higher enrichment of TE-eQTLs compared to gene-eQTLs for TRIM28 motifs and H3K9me3 modifications suggests a tissue-specific silencing of expressed TEs.

**Tissue-specificity of TE-eQTLs**

The high diversity and extensive tissue sampling of the GTEx dataset allowed us to develop a global view of the genetic effects of cis-eQTL variants on TE expression and to discover how these effects vary between tissues of the adult human body. We performed linear mix models with an interaction term between variant and tissue and applied multiple testing correction with a given FDR of 5% (Methods 2.3). Not all variant-TE pairs and variant-gene pairs could be tested as not all TEs or genes are expressed on all tissues. We discovered a range of 16 to 7,036 tissue-specific and a range of 42 to 1,498 shared TE-eQTLs among the pairs of tissues tested and 3 to 2,757 tissue-specific and 18 and 557 shared gene-eQTLs. (Figure 4; Supplementary Figure 10; Supplementary Tables 4-10). We observed that shared eQTLs are significantly closer to the transcription start site (TSS) of TEs or genes compared to tissue-specific eQTLs (Wilcoxon rank sum test p-value < 2.2e-16 for both TE- and gene-eQTLs; Figure 4.A; Supplementary Figure 10.A). Shared TE- and gene-eQTLs maintain the same effect between both tissues whereas the tissue-specific TE- and gene-eQTLs were only active in one of the tissues. (Figure 4.B, Supplementary Figure 10.B). Furthermore, we observed patterns that reflected relationships between related tissues. The strongest association we observed was the high sharing of TE-eQTL effects between brain tissues and between non-brain related tissues. Interestingly, we observed a very low percentage of shared TE-eQTL effects between brain and non-brain related tissues (Figure 4.C). Furthermore, the cerebellum (Brain Cerebellar Hemisphere and Brain Cerebellum) appeared to have a lower percentage of sharing of TE-eQTL effects with the other brain tissues indicating that eQTL effects between the cerebellum and other brain regions are quite distinctive. Withing non-brain tissues, we observed strong sharing among closely related tissues, such as arterial tissues, skeletal muscle and heart and skin tissues. The same patterns of tissue-sharing and tissue-specificity as the ones discovered for TE-eQTLs were observed for gene-eQTLs in our study (Supplementary Figure 10.C) and previously in another study [15]. The patterns of sharing and tissue-specificity were also supported by replication between cis-eQTLs estimated by π1 (the proportion of true positives) [16] among cis-eQTLs identified in one tissue and then tested for replication in another tissue (Supplementary Figure 11,12).
Altogether, these results shed light on the sharing of cis- TE-eQTLs between human adult tissues and point towards similar regulatory functions of TE-eQTLs between more closely related tissues.

**Transposable elements are mediators of genetic effects onto genes**

To assess the causal relationship between eQTL variants, transposable elements and genes, we proceeded as previously described [6]. We focused on regulatory variants affecting both a TE and a gene in an unbiased manner by first finding all significant TE-gene associations at 1% FDR, on a per tissue basis, in a similar approach to QTL mapping (Methods 3.1). Next, we quantified the TE-gene pairs using a dimensionality reduction approach based on Principal Component Analysis (PCA) where we aggregated TE and gene expression together by using the coordinates of the first PC (Methods 3.2) and then used these pseudo-quantifications to find all eQTL-TE-gene triplets by performing a standard eQTL analysis (Methods 3.3). We discovered hundreds to thousands of triplets in all 43 tissues ranging from 821 in Brain Hippocampus to 27,679 in Thyroid, for which we inferred the most likely causal relationship using Bayesian networks [17-19]. We tested three possible scenarios: (1) the causal scenario where the eQTL variant affects the expression of the TE and then the expression of the gene, (2) the reactive scenario where the eQTL variant affects the expression of the gene and then the expression of the TE and finally (3) the independent scenario where the eQTL variant affects independently the expression of the TE and gene (Supplementary figure 13; Methods 3.4). Because of the high variability in sample size, we also performed bootstrapping and kept triplets that had a bootstrap coefficient (percentage of obtaining the same model as the originally discovered after 100 runs of bootstrapping) higher than 60% (Methods 3.5; Supplementary Figure 14). We observed that the reactive model is predominant across all tissues with the causal model ranging from 19% to 28% (Figure 5A,B). Pooling all results together, we discovered a total of 28,828 TEs that are causal for changes in the expression of 12,599 genes. The causal TEs ranged from 141 in hippocampus up to 5,528 in Thyroid with 125 and 2,606 associated genes, respectively, indicating that several TEs have a functional role in the transcriptional regulation of nearby genes (Supplementary Figure 15). Interestingly, we observed that the proportion of causal models changed depending on the position of the TE with respect to the gene and the eQTL variant (Figure 5.C) where intronic TEs were significantly less causal for changes in gene expression compared to intergenic TEs (Fisher exact test p-value < 2.e-16). This was also observed for TEs downstream of the gene body compared to TEs upstream of genes (Fisher exact test p-value < 2.e-16) This decrease of causal models and predominance of reactive scenarios in intronic and downstream TEs could be a consequence of gene transcription and not of bona fide TE transcription from the TE promoter. Interestingly, across all 43 tissues, we observed a significant increase of causal scenarios when the eQTL variant lied inside the TE rather than outside (Fisher p-value < 2e-16) highlighting direct regulatory effects of the TEs onto gene expression (Figure 5.C).
High sharing of TE causality on gene expression across adult tissues

Next, we wanted to understand whether causal TEs in one tissue impact gene expression (i.e. causal) in another tissue from the 43 tested. Effectively, this meant performing 1,806 pairwise comparisons. We first started by looking for shared triplets across pairs of tissues; that is eQTL variants – TE – gene triplets that are either the same between two tissues or where the eQTL variants are different but in high Linkage Disequilibrium ($r^2>0.9$) (Methods 3.6). The number of shared triplets between pairs of tissues ranges from 65 shared triplets between Whole blood and Brain Hippocampus and up to 3060 shared triplets between suprapubic skin not exposed to sun and skin exposed to sun from the lower leg (Supplementary Table 11,12). As the number of shared triplets was not high (Supplementary Figure 16) we observed only a low number of shared of causal TEs, specifically for tissues that are closely related like for adipose or colon tissues (Supplementary Figure 16). To get a better understanding of the sharing of TE causality between tissues, we decided to look at the union of triplets between pairs of tissues. These are defined as eQTL-TE-gene triplets that are shared between pairs of tissues or where the eQTL variant is significantly associated at 5% FDR with the TE-gene pair in the other tissue from the one where the triplet was originally discovered (Methods 3.7). The number of triplets increased almost 5-fold compared to the shared triplets, ranging from 322 between Whole Blood and Brain Hippocampus and up to 13,878 between triplets between suprapubic skin not exposed to sun and skin exposed to sun from the lower leg (Figure 6; Supplementary Table 13). We also observed patterns reflecting relationships between closely related tissues. The observed patterns appeared to be similar to the ones observed with the tissue-specific and shared TE-eQTLs, where high sharing of TE causality is observed between brain related tissues and among non-brain tissues with lower sharing between the two groups (Figure 6). As previously observed, two brain tissues appeared to have lower sharing of TE causality with other brain tissues. Within non-brain tissues, we observed strong sharing among closely related tissues, such as arterial tissues, skeletal muscle and heart and skin tissues. Altogether, these results suggest that TE causality is shared between closely related tissues, as highlighted by the majority of causal TEs in brain tissues being specific to this tissue.

Causal TEs act as alternative promoters for nearby genes

TEs are able to impact gene expression in a plethora of ways [1]. One possible molecular mechanism is by acting as promoters for nearby genes. This allows for the creation of chimeric transcripts (tcGTs) [1, 20] subsequently impacting the expression of genes. Thus, we wanted to assess whether triplets with causal TEs showed evidence of tcGT events between the TE-gene pair of the triplet (Methods 4). We only kept tcGTs made up of the same TE and gene as in the eQTL-TE-gene triplets. We discovered a total of 58 tcGTs in 157 triplets across 34 tissues. Adipose subcutaneous tissue has the majority of these tcGTs (n=13), followed by Thyroid (n=11) and Lung (n=10). Of the 157 triplets constituted with a tcGT, 155 of them the TE is causal of changes in gene expression. Altogether, we discovered several of these tcGT event in our dataset indicating that a subset of TEs influence the expression of 2,757 nearby genes by acting as alternative promoters.
DISCUSSION

Here, we present a global profile of transcriptionally active TEs in 43 different human adult tissues, further challenging the assumption that TEs are transcriptionally inactive in adults and show that TEs are prevalent mediators of genetic effects onto nearby genes, impacting their expression. By combining genome and transcriptome data, we showed that TEs are under genetic control and discovered that many transcription factors regulate TE expression. Interestingly, we observed broad patterns of sharing TE-eQTL effects between closely related tissues and a clear distinction between TE-eQTL effects from brain related tissues and non-brain related tissues. Focusing on the causal relationship between eQTL variants, transposable elements, and gene expression, we dissected the effects of eQTLs and showed that for some of the genes, the eQTL effect is passed on to genes through TEs which act as mediators to regulate gene expression. Similarly, for tissue-specific and shared TE-eQTL effects, we observed that closely related tissues had a higher sharing of causal TEs with again a clear distinction between brain and non-brain related tissues indicating that TEs have different functional roles depending on the tissue they are expressed. Additionally, we investigated whether TEs could act as alternative promoters and drive gene expression and showed that this is true is some tissues where a subset TEs create chimeric transcripts of which some are predicted to give rise to functional proteins.

Several limitations of this project need to be addressed. In this study, we solely focused on transposable elements that impact the expression of nearby genes in an independent manner. However, it is highly plausible that the expression of a single gene is impacted by multiple eQTL variants and TEs. On top of that, synergistic effects from cis- and trans- TE-eQTLs could also be of importance. Thus, in order to obtain a better and fuller picture of the interplay between eQTL variants, TEs and genes, we would need to build more complex networks requiring, however, a higher sample size for sufficient statistical power. Furthermore, because of the highly repetitive nature of transposable element DNA sequences and their relatedness among TE families and subfamilies, uniquely mapping TEs is a true challenge [14, 21]. Future studies involving RNA-seq with read lengths higher than paired-end 100 bp or using long read sequencing technologies could allow for a higher quality mapping of expressed TEs and give us a fuller picture of the TE transcriptome. Also, the samples in the GTEx project were collected postmortem. Thus, there is a possibility that the transcriptome of these samples is different from those of living individuals [22]. Thus, monitoring changes in gene expression of living organisms may give less biased results [23, 24].

Altogether, we revealed that TEs are transcriptionally active in human adult tissues and play role in mediating genetic effects, impacting the expression of nearby genes. These results can serve as an atlas of TE-eQTLs, providing a roadmap to study and discover the role of TEs in complex traits and diseases.
METHODS

1. Genotype Tissue Expression Consortium (GTEx) dataset

Whole genome sequencing and RNA-sequencing data were downloaded from dbGAP (accession code: phs000424.v8.p2).

1.1. Genotypes

We used already filtered genotypes in VCF format provided by GTEx. We kept all variants with a Minor Allele Frequency (MAF) ≥ 5% yielding a total of 6,494,417 variants.

1.2. TE and gene quantification

GTEx samples were processed as previously described [6]. In brief, we mapped the samples using the hisat2 software [25] to the hg19 human reference genome. We then used featureCounts [26] to quantify gene and transposable element expression, providing a single GTF annotation file containing both gene and TE information. Then, all samples were normalized for sequencing depth. We excluded any tissue where the sample size was below 80 yielding a total of 43 tissues to analyze. Finally, any molecular phenotype with more than 50% of missing data (zero expression) on a per tissue basis were removed.

1.2.1. Normalization of molecular phenotypes

To correct for experiment/technical variability in the expression data and retain biological variability, we residualised the molecular phenotype data for the covariates described below:

1. Ancestry. We merged GTEx and 1000 genomes [27] samples into a multi-sample VCF file. Then, we performed a Principal Component Analysis (PCA) using QTLtools pca [28] on genotype data that has been trimmed (one variant with MAF >5% every 100kb) to get the coordinates of each individual on the three first principal components (PCs) serving as a proxy for population stratification in the data (Supplementary Figure 17). Thus, to correct for population stratification that we observed among the GTEx samples, we used the first 3 principal components (PCs) as covariates.

2. Gender. We used gender as a binary variable as a covariate.

3. Sequencing. As GTEx samples where sequenced with different sequencing machines, we corrected for this using a binary variable as a covariate.

4. Experimental. To capture technical/experimental variability in our expression data, we performed PCA, centering and scaling using the pca mode in the QTLtools software [28], on a per tissue basis. To find out the number of Principal Components (PCs) that capture technical variability, we carried out multiple rounds of eQTL mapping for all 43 tissues separately, each time using the 3 PCs from genotypes, gender and sequence covariate and incrementally adding 0, 1, 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200 PCs as covariates. This approach results in identifying the number of PCs required for maximizing eQTL discovery (table below).
These covariates were regressed out from the 43 different tissues using QTLtools correct mode [28] using the –normal option in QTLtools to rank-normalize on a per phenotype basis across all samples such that quantifications follow a normal distribution N(0,1).

2. Transcriptome QTL analysis

All QTL analysis were performed separately for each of the 43 tissues. We used genotypes with MAF ≥5% and gene and TE expression data with normalized counts per million (CPMs) for all downstream analyses of QTL mapping.

2.1. Expression Quantitative Trait Loci (eQTL) mapping

eQTL analyses were performed as previously described [6, 28]. In brief, we used the cis mode of the QTLtools software [28]. For each molecular phenotype, we enumerated all variants in a 1 MB window (+/- 1Mb) around the transcription start site (TSS) of genes/TEs, tested them all and retained only the most significant association (the one with the smallest p-value). Next, we used 1000 permutations to adjust the nominal p-values for the different number of variants being tested per phenotype. We then fitted a beta distribution and used the resulted beta distribution to adjust our nominal p-values. This gave us the best variant-phenotype pair. Finally, we corrected for multiple testing using the R/Qvalue package [29] with a given False Discovery Rate (FDR) of 5%.

2.2. Conditional eQTL analysis

Conditional eQTL analysis was performed as previously described [6, 28]. In brief, after running permutations (Methods 2.1), we calculated a nominal p-value threshold of significance at 5% FDR by determining the p-value threshold corresponding to 5% FDR level and then went from adjusted p-value to nominal p-value threshold using the beta quantile function. Finally, the conditional analysis uses a forward-backward regression to discover all the independent QTLs per phenotype (gene/TE) and identify the most likely candidate variants while controlling for a given FDR of 5%. We only kept the most significant variants per phenotypes for all downstream analyses.

2.3. Shared and tissue-specific eQTL analysis

Tissue-specific and shared eQTL analysis was performed as previously described [6]. In brief, we used linear mix models with an interaction term between genotype and tissue to test whether eQTL effect in each tissue pair were significantly different. We only tested variant-phenotype pairs that were present in both tissues after all filtering steps applied onto gene expression (Methods 1.2). Then, we corrected for multiple test using the R/Qvalue package [29] with a given FDR threshold of 5%. Finally, for all significant results at 5% FDR, when eQTL slopes (as given by the conditional QTL analysis; Methods 2.1) in both tissues had the same direction, then we only kept the associations where the variant-phenotype pair was not nominally significant in the opposite tissue. Shared eQTLs were defined as the ones where the p-value for the interaction term was significant at 5% FDR and where the variant-
phenotype pair was significant at 5% FDR in both tissues (as given from the conditional eQTL analysis).

2.4. **Functional enrichment analysis**

Functional enrichment analysis was performed as previously described [6]. In brief, we first sampled for each eQTL variant 100 random regulatory genetic variants that matched for relative distance to TSS (±2.5kb) and minor allele frequency (±2%) and that were not significant eQTLs. For each category, we created a 2x2 contingency table resembling the one below and used a Fisher exact test to calculate the p-value for the enrichment.

<table>
<thead>
<tr>
<th>eQTL variants</th>
<th>Null distribution of variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variants falling in category</td>
<td># eQTL variant falling in category</td>
</tr>
<tr>
<td>Variant not falling in category</td>
<td># eQTL variants not falling in category</td>
</tr>
</tbody>
</table>

Finally, we corrected for multiple testing using the “p.adjust” function using an FDR threshold of 5%.

2.5. **Ensembl Regulatory Build ChIP-seq dataset**

We downloaded ChIP-seq data from the Ensembl Regulatory Build FTP site (http://ftp.ensembl.org/pub/grch37/current/regulation/homo_sapiens/). It contains ChIP-seq data for 202 transcription factors (TFs) and 29 histone marks for a total of 231 experiments originating from 88 human cell types (hg19 build). For each ChIP-seq experiment, we took the union of peaks across all 88 cell types and merged overlapping peaks together using the merge options in the BEDtools software [30] creating a genome-wide annotation of TFs and histone marks.

3. **Causal relationship between eQTL variants, TEs and genes**

3.1. **TE – gene association analysis**

To discover associations between TEs and genes, we proceeded as previously described [6]. In brief, we used a similar approach to QTL mapping (methods 2.1) where we correlated TE expression with gene expression and only kept significant associations at 1% FDR. Then, we estimated the nominal p-value threshold of significance at 1% FDR for each molecular phenotype as previously described (methods 2.2). Finally, we extracted all TE – gene associations with a p-value below the nominal p-value threshold of 1% FDR.
3.2. Quantifying TE-gene pairs

The quantification was performed as previously described [6, 19]. For all the TE – gene pairs discovered in the previous step, we performed Principal Component Analysis (PCA) and for each TE – gene pairs we used the first principal component (PC) as the new pseudo-quantification and ranked-normalized on a per TE – gene pair so that the values follow a normal distribution N(0,1). All this is implemented in the clomics software package accessible here https://github.com/odelaneau/clomics [19].

3.3. QTL – TE – GENE triplet discovery

For each TE-gene pairs discovered, we performed eQTL analysis using the cis mode of QTLtools as described in methods 2.1. We kept all variant – TE – gene associations that were significant at 5% FDR.

3.4. Causal inference by Bayesian networks (BNs) for QTL-TE-GENE triplets

Causal inference for QTL – TE – gene triplets was performed as previously described [6]. We used BNs to learn the causal relationships between triplets of variables, each one containing a genetic variant, a transposable element and a gene. We tested three possible scenarios (i) the causal scenario where the genetic variant (eQTL) affects first the TE and then the gene, (ii) the reactive model where the genetic variant (eQTL) affects first the gene and then the TE and (iii) the independent model where the genetic variant (eQTL) affects the TE and gene independently (Supplementary figure 13).

We used the R/bnlearn package (Version 4.5) [31] to compute the log likelihoods of the 3 BN topologies. Finally, we went from likelihoods to posterior probabilities by assuming a uniform prior probability on the three possible models. We used posterior probabilities for all downstream analyses.

3.5. Bootstrapping for accuracy estimation

Because of the high variability in sample size between the 43 analyzed tissues, we used bootstrapping to estimate the accuracy of the Bayesian networks and provide confidence for the retrieved posterior probabilities. We performed 100 bootstrap runs for each triplet separately using sampling with replacement. Then, we computed the accuracy estimate for each triplet by counting how many times the most probable model across the 100 bootstrap runs was the same as in the original BN result over the total number of bootstrap runs. We filtered out all triplets with a confidence value below 60%.
3.6. **Shared triplets across pairs of tissues**

We looked for shared triplets across all 43 tissues by performing pairwise comparisons, giving rise to 1806 tissue pairs. For each tissue pair, we looked for common triplets between both tissues or triplets where the TE – gene pair was the same and where eQTL variants were in high Linkage Disequilibrium (LD) with an $r^2$ higher than 0.9.

3.7. **Union of triplets across pairs of tissues**

For each of the 1806 tissue pairs, we kept all triplets where the eQTL variant from triplets of tissue 1 were significantly associated with the TE-gene pairs of the associated triplets with a given FDR of 5% in tissue 2 of the tissue pairs. For all the triplets kept, we ran BNs to get the most probable model in tissue 2 as described in methods 2.8.

4. **Transpochimeric transcripts analysis**

For the discovery of transpochimeric transcripts (tcGTs) we first computed the transcriptome from the RNA-seq bam files using StringTie [32] with -j 1 and -c 1 parameter, on a per sample basis. Then, we intersected the transcriptomes with ensembl hg19 coding exons and the curated RepBase [7] using BEDtools [30] to extract all tcGTs for each sample. We then aggregated the sample level tcGTs into counts per tissue and discarded any tcGT where their first exon was overlapping with an annotated gene or where the TSS was not overlapping with a TE. Finally, for each tcGT events, we computed the occurrence per tissue.
AUTHOR CONTRIBUTIONS

N.M.R.L and H.O designed the study. N.M.R.L analyzed the data and wrote the manuscript and N.M.R.L, H.O, and D.T interpreted the results. E.P shared the quantifications data.

COMPETING INTERESTS

All other authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported by grants from the European Research Council (KRABnKAP, No. 268721; Transpos-X, No. 694658), the Personalized Health and Related Technologies (PHRT-508) program, the Swiss National Science Foundation (310030_152879 and 310030B_173337), the Swiss Cancer League and the Aclon Foundation to D.T.

DATA AVAILABILITY

GTEx WGS and RNA-seq data can be accessed on dbGAP with accession code phs000424.v8.p2.e

CODE AVAILABILITY

All custom scripts used can be accessed here: https://github.com/NLykoskoufis/GTExTE.
REFERENCES


**FIGURES**

**Figure 1:** Expressed transposable elements. (A) number of expressed TEs per tissue. (B) heatmap representing the proportion of expressed TEs per subfamilies in all 43 tissues. (C) Pie chart showing the proportion of TEs overlapping different types of regulatory elements. We uniquely mapped and quantified a total of 521,466 TEs across all 43 tissues. The majority of TEs are L1s and Alus, followed by L2s and MIR. Additionally, 100,703 TEs contain regulatory elements within their sequence highlighting a potential regulatory role for genes.
Figure 2: cis- eQTL discovery. (A) represents the number of cis-eQTLs discovered as a function of the tissues sample size. We observe that samples with higher sample sizes have more cis-eQTLs discovered than the ones with lower sample sizes. (B) Represents the distance in base pairs of the eQTL variant to the TSS of the TE/gene in absolute log10 values. (C) the barplot represents the proportion of eQTLs per TE and the points represent the mean number of eQTLs per TE. The majority of TEs have only one eQTL per TE with only a small proportion of them having more.
Figure 3: Functional enrichment of TE- and gene- cis-eQTLs. Heatmap representing the log2 odds-ratio between the TE enrichment versus gene enrichment. Values above 0 indicate a higher enrichment for TE-eQTLs whereas negative values indicate a stronger enrichment for gene-eQTLs. Grey indicates either those TFs or histone marks where not significant for TE and gene-eQTLs or that they did not overlap with any eQTL in the tissues. We excluded cases where a TF or histone mark had a higher enrichment for gene-eQTLs in only one tissue. We observed that the majority of TFs and histone marks have a stronger enrichment for gene-eQTLs. However, some are more enriched for TEs such as Bdp1, Brf1, ZNF274, TRIM28 and H3K9me3 indicating that these are more specific towards the regulation of transposable elements.
Figure 4: Tissue sharing and specificity of TE-eQTLs. (A) Represent the distance of the eQTL variant from TSS of the TE in absolute log2 scaled values from all tissues taken together. We observed that shared TE-eQTLs are significantly closer to the TSS of TEs than tissue-specific TE-eQTLs are. (B) Effect size (regression slope) of the tissue-specific and shared TE-eQTLs from all tissues taken together. We observe that shared TE-eQTLs have the same effect from one tissue to the other whereas tissue-specific TE-eQTLs do not. (C) Heatmap representing the percentage of tissue-specific TE-eQTLs per pairs of tissues tested. The black boxes represent the major closely related groups of tissues where we observed the highest percentage of sharing of eQTL effects. The grey color represents either a comparison between same tissues or absence of common TE between the pair of tissues.
Figure 5: Causal inference between eQTL variants, TE and genes. (A) Represents the number of triplets discovered per tissue after filtering for a bootstrap confidence value of 60%. (B) Represents the mean posterior probabilities for the three tested models in the 43 tissues tested. (C) Barplot representing the causal relationship between the eQTL variant, the TE and gene depending on their genomic position. We combined all discovered triplets and worked out the percentages by averaging the posteriors given by the Bayesian Networks across all the triplets falling in each of the categories. We observed that TEs inside genes or downstream of genes tend to react to gene expression, whereas TEs outside genes or when the eQTL variant is within the TE sequence, TEs are more likely causal for changes in gene expression.
Figure 6: Sharing of TE causality between tissues. (A) Represents the distribution of the number of triplets passing 5% FDR. (B) Heatmap representing the percentage of triplets that are causal in the original and testing tissue. We observed higher sharing of TE causality between closely related tissues. The black boxes represent the major groups of tissues where TE causality had the highest percentage of sharing.
Supplementary material for

The interplay between regulatory variants, transposable elements and gene expression in adult human tissues

Nikolaos M. R. Lykoskoufis, Evarist Planet, Didier Trono, Halit Ongen

1 Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland
2 Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, 1211 Geneva, Switzerland
3 Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland
4 School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), 1015, Lausanne, Switzerland

* Corresponding authors: N.M.R Lykoskoufis: nikolaos.lykoskoufis@unige.ch, Halit Ongen: halit.ongen@unige.ch
Department of Genetic Medicine and Development, University of Geneva, 1 rue Michel-Servet, 1211 Geneva, Switzerland
Supplementary Figures

Supplementary Figure 1: Sample size per tissue. Barplot representing the sample size per tissue. We decided to exclude any tissue with less than 80 samples as these ones will not have sufficient statistical power for the downstream analyses.
Supplementary Figure 2: Number of covarites used per tissue maximizing the eQTL discovery. Barplot representing the number of eGenes/eTEs discovered per tissue and the number of covariates used for detecting them. Tissue abbreviations and their corresponding tissue name are detailed below. ADPSBQ = Adipose Subcutaneous, ADPVSC = Adipose Visceral Omentum, ADRNLG = Adrenal Gland, ARTAORT = Artery Aorta, ARTCRN = Artery Coronary, ARTTBL = Artery Tibial, BRNACC = Brain Anterior cingulate cortex BA24, BRNCDT = Brain Caudate basal ganglia, BRNCHB = Brain Cerebellar Hemisphere, BRNCHA = Brain Cerebellum, BRNCTXA = Brain Cortex, BRNCTXB = Brain Frontal Cortex BA9, BRNHP = Brain Hippocampus, BRNHT = Brain Hypothalamus, BRNCC = Brain Nucleus accumbens basal ganglia, BRNPTM = Brain Putamen basal ganglia, BREAST = Breast Mammary Tissue, FIBRBL = Cells Transformed fibroblasts, CLNSGM = Colon Sigmoid, CLNTRN = Colon Transverse, ESPGEL = Esophagus Gastroesophageal Junction, ESPLCS = Esophagus Mucosa, ESPMEL = Esophagus Muscularis, HRTAA = Heart Atrial Appendage, HRTLV = Heart Left Ventricle, LIVER = Liver, LUNG = Lung, MSCLSK = Muscle Skeletal, NERVET = Nerve Tibial, OVARY = Ovary, PNCRES = Pancreas, PTTARY = Pituitary, PRSTTE = Prostate, SKINNS = Skin Not Sun Exposed Suprarepubic, SKINS = Skin Sun Exposed Lower leg, SNTTRM = Small Intestine Terminal Ileum, SPLEEN = Spleen, STACH = Stomach, TESTIS = Testis, THYROID = Thyroid, VAGINA = Vagina, WHBLBD = Whole Blood
Supplementary figure 3: Proportion of eQTLs discovered. The figure represents the proportion of TE- or gene-eQTLs discovered as a function of the tissue sample size. We can see that higher sample sizes have a higher proportion of eQTLs discovered compared to tissues with lower sample size.
Supplementary Figure 4: functionally independent gene-eQTL discovery. The barplot represents the proportion of eQTLs per TE and the points represent the mean number of eQTLs per TE. The majority of TEs have only one eQTL per TE with only a small proportion of them having more.
Supplementary Figure 5: eQTL variant enrichment for regulatory elements. The figure represents the enrichments for the different regulatory features in the Ensembl Regulatory Build dataset. We combined all TE- and gene-eQTLs discoveries together and performed an enrichment analysis. We observed that gene and TE eQTL variants are significantly enriched for most regulatory regions but depleted for enhancer regions.
Supplementary Figure 6: Allele frequencies of eQTL variants. We combined all TE- and gene-eQTLs from all tissues and observe that allele frequencies of TE- and gene-eQTLs are very similar.
Supplementary Figure 7: Functional enrichment of TE-eQTLs. The barplot represents the number of TFS and histone marks found to be significantly enriched at 5% FDR for TE-eQTLs per tissue.
Supplementary Figure 8: Functional enrichment of gene-eQTLs. The barplot represents the number of TFS and histone marks found to be significantly enriched at 5% FDR for gene-eQTLs per tissue.
Supplementary Figure 9: Functional enrichment of TE- and gene- cis-eQTLs. Heatmap representing the log2 odds-ratio between the TE enrichment versus gene enrichment for all the TFs and histone marks tested. Values above 0 indicate a higher enrichment for TE-eQTLs whereas negative values indicate a stronger enrichment for gene-eQTLs. Grey indicates either those TFs or histone marks where not significant for TE and gene-eQTLs or that they did not overlap with any eQTL in the particular tissues.
Supplementary Figure 10: Tissue sharing and specificity of gene-eQTLs. (A) Represent the distance of the eQTL variant from TSS of the gene in absolute log2 scaled values from all tissues taken together. We observed that shared gene-eQTLs are significantly closer to the TSS of genes than tissue-specific gene-eQTLs are. (B) Effect size (regression slope) of the tissue-specific and shared gene-eQTLs from all tissues taken together. We observe that shared gene eQTLs have the same effect from one tissue to the other whereas tissue-specific gene-eQTLs do not. (C) Heatmap representing the percentage of tissue-specific gene-eQTLs per pairs of tissues tested. The black boxes represent the major closely related groups of tissues where we observed the highest percentage of sharing of eQTL effects. The grey color represents either a comparison between same tissues or absence of common gene between the pair of tissues.
Supplementary Figure 11: replication of TE-eQTLs. Heatmap representing the replication of TE-eQTLs in another tissue. For each of the tissues (original tissue, rows), we computed the $\pi_1$ metric (distribution of true positives) in all other 43 tissues (testing tissue, columns). We observed that closely related tissues have a higher sharing ($\pi_1$ value) of eQTL effects between them. Tissue abbreviations are detailed in Supplementary Table 1.
Supplementary Figure 12: replication of gene-eQTLs. Heatmap representing the replication of gene-eQTLs in another tissue. For each of the tissues (original tissue, rows), we computed the $\pi_1$ metric (distribution of true positives) in all other 43 tissues (testing tissue, columns). We observed that closely related tissues have a higher sharing ($\pi_1$ value) of eQTL effects between them. Tissue abbreviations are detailed in Supplementary Table 1.
Supplementary Figure 13: Causal relationship of eQTLs, TEs and genes approach. To infer the most likely causal relationship between eQTL variants, TEs and genes, we tested three models using Bayesian Networks (BNs). The causal model where the eQTL variant affects the TE and then the gene, the Reactive model where the eQTL variants affects the gene and then the TE and the Independent model where the eQTL variant affects the TE and gene independently. For each triplet we obtained log likelihoods and we calculated posterior probabilities using a uniform prior probability for each of the three models.
Supplementary Figure 14: Bootstrap confidence. For each triplet, we performed 100 bootstrap runs and computed the bootstrap confidence value which is the percentage of the number of times the bootstrap run gave the same model as the originally predicted one divided by the number of bootstrap runs, times 100.
Supplementary Figure 15. Causal TEs and the number of associated genes per tissue. The barplot represents the number of causal TEs and associated genes per tissue.
Supplementary Figure 16: Shared causal triplets. (A) Represents the distribution of number of shared triplets between the pairs of tissues. (B) Heatmap representing the percentage of shared causal triplets between pairs of tissues.
Supplementary Figure 17: Population stratification analysis using PCA. To correct for ancestry, we performed a PCA using GTEx genotypes merged with 1000 genomes. We detected that GTEx patients were from different origins, thus we used the 3 first PCs to correct for the depicted population stratification.
We are living in unprecedented times where cutting edge technologies are drastically advancing medical and biological research. The sequencing of the human genome in the beginning of the 21st century, followed by the development of genotyping arrays and later on of next generation sequencing (NGS) technologies has allowed us to map thousands of genetic variants. In addition, the creation of new statistical tools has enabled us to associate thousands of rare and common genetic variants with complex traits and diseases.

However, as most of these variants fall in the non-coding genome there is an increased need for understanding their functional role. The development of microarray technology and RNA sequencing has allowed for whole genome transcriptomic profiles of cells, revolutionising medical and biological research through the discovery of new biomarkers for disease prediction and development of new treatments [188–194].

Expression Quantitative Trait locus (eQTL) studies, a method integrating both gene expression and genotyping data arose as a crucial tool for better understanding how genetic variation amongst individuals influences disease risk and was an essential step forward regarding the interpretation of GWAS variants. Indeed, GWAS variants are not giving us any information on the genes associated with a particular trait or disease, whereas eQTL variants do. Thus, eQTL studies and GWAS studies have become interconnected as together they allow us to link variants with traits and diseases and subsequently discover which genes are associated with them.

Very quickly, the scientific community understood that what was previously called “junk DNA” is anything but junk and plays a fundamental role in the regulation of gene expression by a plethora of mechanisms. For example, there is increasing evidence that repetitive DNA sequences, called transposable elements, are important regulators of gene expression during development and disease. Thus, getting a better understanding of the non-coding genome and its role in the regulation of gene expression is fundamental.

In this thesis, we aimed to better characterise the relationship between the non-coding and coding genome in healthy individuals and in cancer by studying the interplay between non-coding regulatory variants (eQTLs), transposable elements and genes. In the following paragraphs, we will discuss about the main discoveries, the limitations we faced and the future challenges. Finally, we will describe what are the crucial next steps required to better comprehend the aetiology of complex traits and diseases.
4.1 Transposable elements as mediators of genetic effects onto genes in cancer

In the first project presented in the thesis, we studied the genetic control of transposable elements and how they mediate genetic effects onto nearby genes to impact their expression in cancer. Multiple studies have shed light on the important contribution of TEs to tumorigenesis and how they provide additional means whereby gene expression can be altered [195–197]. However, these studies have focused at specific TE loci or specific sub-families. In our project we applied a genome-wide scan that allowed us to obtain a better picture of the effects of TEs on gene expression during tumorigenesis. We made several key findings which will be discussed below.

By combining genotype and transcriptomic data from normal colon and colorectal
cancer (CRC) samples we, firstly, showed that transposable elements are under strong genetic control and discovered that transcription factors (TFs) regulate TE expression more specifically in tumor than in the normal state. Secondly, we were able to dissect eQTL effects and discover 1,766 TEs that are drivers of the expression of 1,558 genes, specifically in cancer. Interestingly, 55 of these genes were cancer driver genes where the TE had no effect on them in normal colon suggesting a tumor specific effect of these TEs. We validated our findings using external datasets for colon transverse from GTEx v7 and colon adenocarcinoma (TCGA-COAD) from The Cancer Genome Atlas (TCGA) project and showed that our findings were very well replicated, corroborating our results. Finally, in this project, we demonstrated that alongside predisposing alleles and somatic mutation, germline variants are crucial key players to cancer as they allow for changes in TE expression to occur, subsequently altering the expression of nearby genes in cancer.

TEs are under-appreciated for their role as key development regulators and drivers of oncogenesis. Understanding and categorising the abundance and the functional role of TEs in normal development may shed light regarding their role in tumorigenesis. Rather than focusing on group-wise expression and regulation of TE subfamilies, locus-specific regulation of TEs would add clarity to their role in the regulation or dysregulation of gene expression in normal or disease context. The investigation of TEs during early development and in tumorigenesis will be of high clinical relevance due to the known implications of TE expression and re-activation in response to therapies. Thus, digging deeper in the understanding of the functional role of TEs in the human genome will provide novel insights that will likely improve outcomes in the clinic [198].

This study can be extended to all cancer types that are present in The Cancer Genome Atlas (TCGA) project or the International Cancer Genome Consortium (ICGC) as long as the sample size is sufficient and genotype and transcriptomic data is available. We believe that this project has provided valuable insights into the important role of TEs as drivers of gene expression in cancer.

**4.2 Transposable elements play an important role in the regulation of gene expression in human adult tissues**

Following the results of the first project, we were interested in addressing the genetic control of TEs and their potential in mediating genetic effects on nearby genes in human adult tissues. Transposable elements are known to be broadly active during development and regulate the expression of important developmental genes. Some studies have shown that TEs are transcriptionally active in human adult tissues [61,199–201] but they did not focus on the genetic control of TE expression or whether TEs impact nearby genes. To this end, we made several important findings which will be discussed below.

By combining whole genome sequencing (WGS) and RNA-sequencing data from 43 human adult tissues from the GTEx v7 project, firstly, we showed that thousands of transposable elements are transcriptionally active in adult human tissues, as previously shown [?]. Additionally, we discovered that their expression is under strong genetic control by non-coding germline variants (eQTLs). Moreover, we detected various transcription factors that are significantly more enriched for TE-eQTLs compared to gene-eQTLs, indicating that these TFs are regulating TE expression with a higher specificity compared to genes.
Secondly, we discovered strong sharing of TE-eQTL effects between brain related tissues and between non-brain related tissues and a very low sharing between these groups. Finally, we discovered 28,828 of TEs that are causal for changes in the expression of 12,599 genes across all tissues. Interestingly, we observed similar sharing patterns as the ones observed with the tissue-specific and shared TE-eQTLs, where TE causality is shared between closely related tissues.

This study can be extended in several different ways. Firstly, GTEx contains data from short read sequencing technologies and allowed us to discover thousands of transcriptionally active TEs. However, using long read sequencing will facilitate the mapping of repetitive elements, allowing for a better characterization of the landscape of TE expression. Secondly, the samples in the GTEx project were collected postmortem. Therefore, the transcription levels from these samples may be significantly different from those of living individuals [202]. Thus, monitoring changes in the expression of transposable elements in living individuals could give us more relevant results [203, 204].

In summary, this project gave insights into the longstanding hypothesis whether TEs are transcriptionally active in adult human tissues. It is to our knowledge the first catalogue of TE-eQTLs in human adult tissues that can be used by the scientific community in the hope of better characterising TE expression and further understand the role of TEs complex traits.

4.3 Limitations

Even though the biological questions we tackled in the two projects presented above were different, the approaches and methods used were quite similar and both had common limitations which we will discuss below.

Despite the robustness of the numerous analyses we performed for this project, there are still multiple aspects that need to be addressed to elucidate the impact of TEs. In our study, we took a somehow simplistic approach where each eQTL variant, TE and gene triplet were studied in an independent manner. However, we observed many cases where the same genetic variant was affecting multiple TEs, or cases where the same TEs where affecting multiple genes pinpointing that there are more complex interactions occurring between the three molecular phenotypes. We wanted to try to build more complex models for both projects involving multiple eQTLs, TEs and genes but we faced one of the most common problems in statistical analyses which is limited power. Sample size plays a crucial role in statistical analyses as it dictates statistical power. Higher sample sizes give rise to more robust and trustworthy results. In the second project (chapter 2), several of the analysed tissues had a small sample size which could be depicted by the smaller number of eQTL discoveries made compared to the ones with bigger sample sizes. Henceforth, increasing the sample size will allow us to discover more associations between genetic variants and transposable element expression, giving us the opportunity to better understand how the genome regulated TE expression and how these genetic effects are later on transmitted onto gene expression. Additionally, while we only exclusively focused on cis-eQTLs, there is a high probability that trans-eQTLs affect the expression of TEs, as they do for genes [182]. Thus, incorporating, cis- and trans- eQTLs together and building more complex networks could give us a fuller picture of the role of TEs in the regulation of gene expression in cancer.
TEs are highly repetitive DNA sequences and their evolutionary relatedness among TE families makes mapping them using short read technologies a real challenge (18,29). The SYSCOL dataset (chapter 1) and the GTEx dataset (chapter 2) used in our projects have RNA-sequencing data obtained using 49 base pairs (bp) and 76bp read lengths, respectively. With these read lengths, a previous study estimated that only approximately 68% and 85% of TEs can be uniquely mapped, subsequently leading to missing information [205]. Future studies with longer read lengths (>100bp short read sequencing) or using long-read sequencing technologies could aid to obtain a better picture of the transcriptomic landscape of TEs.

4.4 How to improve research in the future

The colossal advance of technology and the tremendous success of genomics and transcriptomics allowed the discovery of thousands of variants associated with complex traits and diseases. On top of that, the success of eQTL studies gave insights into the functional interpretation of non-coding germline variants. Despite all that, we are still far from fully understanding the aetiology of complex diseases like cancer.

High-throughput technologies have revolutionised medical research. The advent of genotyping arrays enabled large-scale genome-wide association studies and methods for examining global transcript levels, which gave rise to the field of "integrative genetics". Other omics technologies like proteomics and metabolomics are now often incorporated into the everyday methodology of biological research, an approach commonly called multi-omics. Compared to single omics, multi-omics can provide researchers with a greater understanding of the flow of information, from the original cause of disease (genetic, environmental) to the functional consequences or relevant interactions [4,5].

Moreover, there is a great need for sufficient analysis, storage and sharing of data to the community to maximise reproducibility and opportunities to validate findings on external datasets. Additionally, sample size is crucial as statistical power can be detrimental. Thanks to the decrease in price of sequencing, sample sizes should be able to increase over the years and allow for a better characterisation.
4.4. HOW TO IMPROVE RESEARCH IN THE FUTURE

The integration of multiple layers of data will improve our knowledge on the impact of genetic variation on disease but it is important that we also assess these effects over time and with different environmental stimuli as gene-environment interactions can contribute to complex disease phenotypes. Indeed, some variants are deleterious only under specific conditions such as in a particular tissue, at a particular time or under specific environmental stimuli. Longitudinal studies could aid the understanding of genetic effects over time. The decrease in sequencing price and storage of data has made such studies more and more popular [207]. However, multiple studies have evaluated how genetic variation modulates the response of cell to different environmental stimuli illustrating that an important number of genetic variants are masking their effects until they are stimulated [208–211].

This thesis tackled the interplay of the non-coding and coding genome by studying how non-coding regulatory variants and transposable elements impact the expression of nearby genes, pushing the boundaries of our knowledge with the ultimate hope that one day we will be able to fully understand the aetiology of complex traits and diseases.

Figure 4.2 – "Multiple omics data types and approaches to disease research. Layers depict different types of omics data (Box 1). Omics data are collected on the entire pool of molecules, represented as circles. Except for the genome, all data layers reflect both genetic regulation and environment, which may affect each individual molecule to a different extent. The thin red arrows represent potential interactions or correlations detected between molecules in different layers—for example, the red transcript can be correlated to multiple proteins. Within layer interactions, although prevalent, are not depicted. Thicker arrows indicate different potential starting points or conceptual frameworks for consolidating multiple omics data to understand disease. The genome first approach implies that one starts from associated locus, while the phenotype first approach implies any other layer as the starting point. The environment first approach (not shown) examines environmental perturbations. Modified from [206]


[57] V. V. Kapitonov and J. Jurka, “Rag1 core and v (d) j recombination signal sequences were derived from transib transposons,” PLoS biology, vol. 3, no. 6, p. e181, 2005.


125


In this part, I present additional work performed during my PhD that I contributed to.

- **Numerous putative non-coding cis-regulatory drivers in chronic lymphocytic leukaemia** In this study, I performed part of the analyses. Here, we investigated the contribution of the non-coding genome in cancer using clustered regions of non-coding regulatory elements of genes within cis regulatory domains (CRDs) and tested for CRDs accumulating an excess of somatic mutations in chronic lymphocytic leukaemia (CLL).

- **Leveraging interindividual variability of regulatory activity refines genetic regulation of gene expression in schizophrenia** In this study, I helped with designing and performing various analyses. We investigated the interplay between regulatory variants, cis regulatory domains (CRDs) and gene expression and characterise genetic perturbations on the regulatory machinery specific to Schizophrenia.
NUMEROUS PUTATIVE NON-CODING CIS-REGULATORY DRIVERS IN CHRONIC LYMPHOCYTIC LEUKAEMIA

Halit Ongen$^{1,2,3}$*, Olivier Delaneau$^{1,2,3}$, Nikolaos M. R. Lykoskoufis$^{1,2,3}$, Michael W. Stevens$^{1,2,3}$, Cedric Howald$^{1,2,3}$, Emmanouil T. Dermitzakis$^{1,2,3}$*

1 Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland
2 Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, 1211 Geneva, Switzerland
3 Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland
* Corresponding authors

ABSTRACT

Perturbations of the coding genome and their role in cancer development have been studied extensively. However, the non-coding genome’s contribution in cancer is poorly understood$^1$, due to limited functional interpretation of non-coding DNA as well as difficulty in defining the regulatory landscape of genes. In this study, we try to resolve these issues by employing clusters of non-coding regulatory regions of genes with coordinated activity (Cis Regulatory Domains or CRDs) defined in a recent study$^2$. We test for CRDs that accumulate an excess of somatic mutations in chronic lymphocytic leukaemia (CLL) accounting for somatic mutational patterns and biases. At 5% FDR, we find 32 CRDs with significant excess somatic of mutations in CLL, 29 of which are assigned to 104 genes. These include some already implicated in tumorigenesis, and are enriched in pathways already implicated in CLL development such as the B-cell receptor signalling pathway. We discover that the somatic mutations in the significant CRDs are hitting bases more likely to be functional than the mutations in non-significant CRDs and 14 of the significant genes identified exhibit extreme allele specific expression compared to non-significant genes, providing independent dimensions of functional significance. We also find that the transcription factor binding sites (TFBSs, identified in silico) that are disrupted by the somatic mutations in significant CRDs are enriched for TFBSs of factors known to be involved in cancer development. Finally, mutational signatures observed in the regulatory regions of significant CRDs deviate significantly from their counterparts in non-significant CRDs. Collectively, these results suggest selection acting on non-coding DNA during tumorigenesis.
and corroborate their driver potential. Our approach is generalizable and is a new powerful way to
discover non-coding regions involved in tumorigenesis in other cancers.

INTRODUCTION

The Cancer Genome Atlas Consortium (TCGA)³ and the International Cancer Genome Consortium
(ICGC)⁴ have so far been tremendous resources to generate and characterize genomic data from a
variety of human cancer types. By investigating somatic mutations in the protein coding parts of the
genome, they have identified many genes as putative driver genes for multiple cancers⁵, as well as
many known and novel putative drivers in specific cancers⁶⁻¹². However, as expected the vast
majority of the somatic mutations observed in whole genome sequencing (WGS) data lie in the non-
coding portion of the human genome¹ and it is still an open question whether these non-coding
somatic mutations are involved in tumorigenesis, or whether they are simply passenger mutations.

Unlike the impact of the coding genome on tumorigenesis, the contribution of the non-coding
genome to cancer development is less well understood, since interpreting the effect of mutations in
it remains challenging¹. A few studies have identified non-coding regulatory mutations that are
involved in tumorigenesis such as the mutation in the TERT promoter in melanoma¹³,¹⁴ and in other
cancers¹⁵, and recurrent non-coding mutations in chronic lymphocytic leukaemia (CLL)¹⁶ and breast
cancer¹⁷. However, these studies have each found these mutations in a small number of regulatory
regions mainly due to statistical power issues, as individually studied regulatory elements tend to be
small in size. Recent studies have also shown that, as a whole, non-coding mutations play an
important role in tumorigenesis¹⁸. We have previously shown that non-coding variants in the
germline genome are likely involved in tumorigenesis, but also identified genes with putative
somatic regulatory drivers in colorectal cancer, using perturbations in allele specific expression (ASE)
during tumour development¹⁹. However, direct identification of non-coding regulatory regions that
accumulate an excess of somatic mutations using WGS data is an unsolved problem due to the
difficulty in defining the bounds of the non-coding regulatory regions, assessing which genes these
regulate, identifying the null genome to compare to, and failure to accumulate signal over multiple regulatory regions thus decreasing power in detecting these regions. In this study, we address this problem by identifying sets of coordinated non-coding regulatory regions, called cis regulatory domains (CRDs), based on the correlation amongst three histone modifications, and by finding which genes these CRDs regulate. We then test whether these CRDs accumulate an excess of somatic mutations by controlling for differential somatic mutation rate across the genome and between open and closed chromatin regions.

RESULTS

Identification of cis regulatory domains

In order to define likely regulatory non-coding regions of the genome we have used the CRD definitions in Delaneau et. al. 2019. In brief, the authors use H3K27ac, H3K4me1, H3K4me3 histone modifications assayed by ChIP-seq, and transcriptomes assayed by RNA-seq from 317 lymphoblastoid cell lines (LCL, immortalized B-cells) to first identify CRDs based on the correlation between the chromatin marks in the population, and subsequently, to discover the genes these CRDs regulate. The authors identify 12584 CRDs, and by using the resulting CRD and gene expression quantifications they were able to discover 15161 genes (FDR = 5%, 1 Mb cis window) that the CRDs regulate in LCLs (Supplementary methods & Supplementary figure 1). These CRD definitions are significant for this study, since they allow the estimation of the non-coding regulatory regions of genes and group multiple regulatory elements into a single unit, which will not only enable identification of non-coding regulatory regions with excess somatic mutations due to increased statistical power, but also link these regions to specific genes.

Identifying CRDs with excess somatic mutations accounting for somatic mutation biases.

Next, we devised a method to detect excess somatic mutations in CRDs, which would signify positive selection in tumorigenesis and hence indicate driving potential. To do so we first identified the
regulatory regions of the CRDs from the SGX project\(^2\) that do not overlap with known exons.

Subsequently, we created a set of regions that did not overlap with any of the ChIP-seq peaks in SGX or with known exons, which make up non-exonic non-regulatory null regions that we call spacers. In order to account for differential somatic mutation rates observed across the genome\(^{20-23}\), we took the spacers in-between or flanking the regulatory regions of a CRD as the null for a given CRD.

Activation induced deaminase (AID) is known to cause hypermutation\(^{24,25}\) and excess mutations observed in these regions may be due to hypermutation rather than positive selection. To account for this in our model, we categorized the somatic mutations seen in CLL to canonical and non-canonical AID signatures and measured the AID mutation rate across the genome. We then segmented the genome into regions with similar AID mutation rates and scaled the AID mutations observed in these regions by the ratio of non-AID mutation rate to AID mutation rate (Supplementary methods, Supplementary figure 9 & 10). This approach allowed us to penalize individual AID mutations in hypermutated regions of the genome, and effectively normalize the AID mutation rate across the genome.

Different cancers differ in their mutational signatures, i.e. in the types of somatic mutations they accumulate\(^{26}\). In order to incorporate this into our method, we considered the local context of the somatic mutations observed by taking into account the immediate 5′ and 3′ reference bases (trinucleotide context) flanking the somatic mutation position. These trinucleotide contexts were summarized on the pyrimidine bases of the observed somatic mutation. Thus, we counted how many of the 32 possible context triplets are present in the spacer regions of a CRD and how many of them harbour a somatic mutation, to calculate an expected mutation rate for each of the 32 classes. We then counted the number of times these triplets are observed in the regulatory regions of a CRD and using the expected mutation rates for each of the 32 classes, we found the total number of expected mutations. Subsequently we counted the observed number of mutations in the regulatory
regions and using a one-tailed Poisson test calculated a nominal p-value for the excess of somatic mutations in a CRD.

We wanted to account for differential mutation rates between open (regulatory regions) and closed chromatin (spacers)\textsuperscript{27}. Thus, we calculated the local ratio of open chromatin mutation rate to closed chromatin mutation rate from each CRDs flanking regions (500 kb either side) excluding the peaks and spacers in the CRD boundary. We then set this ratio to a minimum of one, thereby requiring the CRD to always have a higher mutation rate in its regulatory regions than its spacers. When this ratio is greater than one we scaled our expected mutation rate in the regulatory regions of the CRD accordingly (Supplementary methods).

Lastly, in order to control for short hypermutated regions, we required that at least two of the non-overlapping regulatory regions of a CRD to have at least 90\% of the average regulatory mutation rate of the CRD. Please see supplementary methods section 3.1, 3.2, and 3.3 for the details of these steps and section 3.4 for how we accounted for all the known biases. Using this methodology, we controlled for the differential mutation rate across the genome, local hypermutation (observed in CLL), the local context of the types of somatic mutations, and the different rates of somatic mutation inferences between open and closed chromatin, therefore resulting in a robust estimate of the significance of excess somatic mutations in CRDs.

In order to estimate the occurrence of false positives in our analysis we ran a simulation analysis. In the simulations, we divided the genome into 100 regions, which contained similar number of bases overlapping with the ChIP-seq peaks. Within these regions we randomly distributed the somatic mutations, preserving the mutational signature proportions and open vs. closed chromatin mutation rates. On top of these mutations, we added 20 short open chromatin regions that were hypermutated, 5 large regions where both open and closed chromatin are hypermutated, and 2 regions with AID hypermutations (Supplementary methods, Supplementary figure 2).

Subsequently we re-run our analysis with the simulated somatic mutations. The results showed that
we effectively control for hypermutations, and other biases in the data. The Q-Q plots are markedly
different between the observed and simulated results, with the simulated results showing a near
null distribution with 4 CRDs lying outside of the 95% confidence intervals of the null
(Supplementary figure 3 & 4).

CRDs with excess somatic mutations

We aimed to find CRDs with significant excess of somatic mutations with the aforementioned
methodology using publicly available somatic mutation calls. To this end, we acquired somatic
mutation calls from WGS data for 150 CLL samples, which comprised a total of 359456 somatic
mutations (mean = 2396.4 ± 155.2)\(^1\). CLL was chosen because it affects the cell type analysed in the
SGX project. We found 32 CRDs that accumulate significantly more somatic mutations (FDR = 5%), 29
of which regulate 104 genes (Figure 1, Supplementary table 1). Due to limitations in power, not all
CRDs are assigned to genes.

One possible technical issue that might confound our analysis is coverage in WGS. If the regulatory
regions of the significant CRDs had systematically higher coverage than their spacers, then this
would cause more somatic mutations to be called in regulatory regions. To assess this, we compared
the number of reads overlapping the somatic mutation calls in the regulatory regions of the
significant CRD to the number of reads of around somatic mutations in their spacers. There was no
significant increase in coverage of somatic mutations in regulatory regions compared to their spacers
(Supplementary figure 6), in fact, the coverage in the regulatory regions was lower than in the
spacers, median 22 reads in regulatory vs. 30 in spacer in CLL. We also investigated whether
inclusion of different spacers, i.e. including or excluding flanking spacers, might influence our results;
thus we reran the analysis without the flanking regions. We observed that the p-values were
significantly positively correlated between the two analyses (rho = 0.92, p = 9.71e-14). These results
indicate that our methodology accounts for any heterogeneity of coverage and variant call
confidence and that our results hold with different ranges of spacer sequence inclusion.
We examined the per sample somatic mutation rates of the significant CRDs and observed that similar to the protein coding drivers there are CRDs that are mutated in most samples, and others exhibiting mutations in just few samples (Figure 2). This suggests that the patterns of selection in non-coding cis-regulatory regions are similar to protein coding sequences with some regulatory regions probably important for tumorigenesis in most or all cancers while others being more specific to one or a few tumours.

Among the genes identified are KRAS and IKZF3, putative drivers previously identified due to excess protein altering mutations\textsuperscript{16}, BCL6 and BCL7A, which are involved in the development of lymphomas\textsuperscript{28,29}, and LPP, where the CRD contains a variant that confers predisposition to CLL\textsuperscript{30}. Overall the genes identified in CLL are enriched for B-cell receptor signalling pathway, one of the major pathways involved in CLL development\textsuperscript{31}, and for pathways involved in the tumorigenesis of other types of cancers (Supplementary table 3). The genes identified are significantly enriched for putative CLL protein coding drivers when compared to genes of non-significant CRDs and the null (Fisher’s exact test $p = 0.049$ and permutation $p = 0.03$, respectively, Supplementary methods).

Furthermore, we find that four of the genes with recurrent non-coding mutations identified in Puente et. al.\textsuperscript{16} are also identified as putative non-coding drivers in our study. This also represents a significant enrichment when compared to the non-significant CRD genes and the null (Fisher’s exact test $p = 0.0001$ and permutation $p < 9.9e-6$, respectively, Supplementary methods). Two of the significant CRDs are in the hypermutated IGHV locus; however not only we account for this hypermutation rate in our analysis but also, we show below using the functional assessment of somatic mutations, that these are unlikely to be false positives. We show that by using our methodology we can identify putative non-coding regulatory drivers for genes that are known to be involved in cancer development as well many novel ones.

We also conducted our analysis excluding all AID mutations from the list of somatic mutations, to see if any of the significant CRDs we find are purely driven by AID hypermutations. We found that
after the removal of AID mutations, all significant CRDs remain significant (Supplementary Figure 11), indicating that our method is correctly accounting for this bias in the data. Lastly, we tested the 32 significant CRDs with ActiveDriverWGS32, an independent but theoretically a less powerful method for discovering non-coding regions with excess somatic mutations, and find that 18 of them have a nominally significant p-value with this method (Supplementary Figure 12).

Comparison of significant CRDs vs. non-significant ones

We observed that the number of genes significant CRDs regulate is significantly higher than the number of genes regulated by non-significant CRDs (Supplementary Figure 8f), indicating that the CRDs identified as cancer drivers tend to regulate more genes and therefore the accumulation of likely jointly contributing mutations to tumorigenesis is expected. The significant CRDs are not significantly different from non-significant CRDs when we compare their size, number of chromatin peaks, correlation strength, and the distance between them and the genes they regulate (Supplementary Figure 8a, b, c, d, e).

Functional assessment of somatic mutations in the significant CRDs

We wanted to assess whether the somatic mutations found in the significant CRDs are perturbing likely functional nucleotide bases. In order to do so we acquired the probabilities for purifying selection acting on the bases of the human genome as calculated by the LINSIGHT method33. Subsequently we compared the purifying selection acting on the positions of the somatic mutations in the significant CRDs vs. the positions in the non-significant CRDs (Supplementary methods). We found that somatic mutations in significant CRDs are hitting bases that are significantly (Mann-Whitney p < 2.2e-16) more likely to be under purifying selection than the mutated bases in non-significant CRDs (Figure 3). We further investigated functional impact of the somatic mutations in the 2 significant CRDs overlapping the hypermutated IGHV locus in CLL34. We found that the somatic mutations in the regulatory regions of these 2 CRDs are also hitting bases that are significantly more likely to be functional than both the mutated bases in non-significant CRDs (Mann-Whitney p < 2.2e-
16) and the non-significant CRDs in the IGHV locus (Mann-Whitney p = 0.006, Supplementary figure 5). This result is indicative of selection in tumorigenesis, even in the hypermutated region.

We expect the functional mutations in the regulatory regions of the CRDs to hit only one of the haplotypes, thus they will have an impact on allelic expression of the genes they regulate. Therefore, under selection, we expect the significant CRD genes to show extreme allelic imbalance compared to non-significant CRD genes. In order to test this, we used the ICGC CLL dataset, which had germline genotypes and RNA-seq data from 71 CLL samples, and ran an ASE analysis (Supplementary methods). We calculated allelic imbalance as the absolute value of the difference between the proportion of the RNA-seq reads that had the reference allele overlapping a transcribed heterozygous biallelic single nucleotide polymorphism (SNP) in a CRD gene, and 0.5 (balanced allelic expression). 64 out of the 104 significant CRD genes had the required RNA-seq coverage (minimum 16 reads overlapping the SNP) in a minimum of 5 samples. Of these 64 genes, 14 showed significantly extreme ASE compared to non-significant CRD genes. The genes with extreme ASE include genes involved in tumorigenesis like BCL6, BCL7A, IKZF3, and LPP. By running permutations we showed that the discovery of 14 genes with extreme ASE is unlikely to arise by chance (p = 0.04, Supplementary methods), corroborating the driver potential of the genes identified.

In order to assess the impact of somatic mutations on gene expression, we correlated the mutational burden in the CRDs with the expression of the genes they regulate. We identified 10 genes with significant correlations between CRD mutational burden and gene expression (Supplementary methods, FDR=20%, this was chosen due to the small sample size). We observed that increased mutational burden in the CRDs tend to increase gene expression more often (Figure 4a). Two examples of these significant associations are given in figure 4, namely the BCL7A gene where the increase in mutational burden is correlated with decreased expression, and the LPP gene in which as the mutational burden increases the gene expression is increased. Notably, these genes are significantly downregulated and upregulated in CLL, respectively, when compared to expression
from LCLs (Supplementary methods). Here, we highlight the functional impact the somatic mutations in CRDs are having on gene expression.

To further investigate the selective pressure acting on the significant CRDs during tumorigenesis, we compared the types of mutational signatures observed between the regulatory sequences of significant CRDs vs. regulatory sequences of non-significant CRDs, and the regulatory regions vs. spacers in significant CRDs. To this end, we classified the single nucleotide somatic mutations into 96 classes representing six possible substitutions of the pyrimidine of the mutated Watson–Crick base pair with incorporating information on the immediate flanking bases (Supplementary methods). If there is selective pressure on the significant CRDs, then we would expect the mutational signatures to differ between their regulatory regions and spacers and also between significant and non-significant CRDs, since we expect the latter ones not to be under selection. We observed that there were significant differences in the mutational signatures in both comparisons (Figure 5), furthermore these differences are also present when we normalize for the trinucleotide compositions of the regions (Supplementary Figure 13). These results signify that during tumorigenesis there is selection acting on the CRDs, which we identify independently of the underlying mutation rate, verifying that the signal of excess somatic mutations is not due to hypermutation, thus corroborating their driver potential.

Lastly, we wanted to examine whether there are specific transcription factor binding sites (TFBSs) perturbed by the somatic mutations in the regulatory regions of the significant CRDs. Thus, we used sequences from regions flanking the somatic mutations by 15 bp (i.e 31 bp in total) in the regulatory regions of significant and non-significant CRDs. Then using the HOMER software package we searched for TFBSs enriched around somatic mutations in significant CRDs compared to the non-significant CRDs (Supplementary methods). We found that transcription factors known to be involved in tumorigenesis are being perturbed, e.g. NF-E2, MAFK, and AP-2 (Figure 6). It is important to note that these binding sites are defined in silico, thus are not experimentally verified,
meaning that other TFs than ones identified here may be involved. Therefore, this analysis suggests changes in transcription factor binding sites as a potential mechanistic cause for the driver potential exhibited by these CRDs.

**DISCUSSION**

In this study, we describe a methodology to identify non-coding cis-regulatory drivers in cancer in an unbiased and genome-wide manner using clusters of coordinated regulatory elements to increase power. We show that this method can find putative non-coding drivers in CLL, cancer affecting main cell type analysed in the SGX project. We confirm our and others’ previous findings that non-coding drivers are important players in tumorigenesis\(^\text{18,19}\) and that the number of genes involved is numerous. As datasets like the SGX project or similar datasets to find coordinated regulatory elements are extended to more cell types, we will be able to find these types of non-coding drivers, which are currently under-studied due to the difficulty in interpreting the non-coding genome, in other cancers. We believe there should be a significant shift in focus to identify non-coding drivers in all cancer types, and this study describes a powerful methodology to do so. Our findings, that numerous cis regulatory regions have driver potential, overall support a model under which cancer tumorigenesis and progression is a complex phenomenon, driven by many driver mutations with different effect sizes, coming both from the somatic and the germ-line genome both coding and non-coding, play an important role in the process. Under this model the well-known protein-coding somatic driver mutations in key genes are the necessary components for tumorigenesis and define many of the common properties of tumours, while the larger number of rarer driver mutations define the specific characteristics of each tumour. It is likely that the specific progression characteristics as well as the response of such tumours to specific therapies and our immune system depend on these larger sets of rare mutations in much the same way they do in complex diseases and this hypothesis provides a strong rationale to explore the non-coding cancer genome more deeply.
AUTHOR CONTRIBUTIONS

H.O and E.T.D designed the study. H.O. and N.M.R.L analysed the data and H.O and E.T.D interpreted the results. O.D. analysed and shared the SGX data, M.W.S and C.H. were involved in data analysis. H.O. wrote the manuscript and E.T.D edited it.

ACKNOWLEDGMENTS

This research was supported by grants from NIH-NIHM (GTEx), European Commission FP7, European Research Council, and Swiss National Science Foundation. The computations were performed at the Vital-IT Centre for high performance computing of the SIB Swiss Institute of Bioinformatics. We thank all members of the Sysgenetix for their crucial input in data generation and analyses.

REFERENCES

Figure 1 – (a) An example of a CRD that has significantly excess somatic mutations in CLL. The red boxes are the positions of the CHiP-seq peaks of the CRD, i.e. the regulatory regions of this CRD, and the blue boxes are the spacers. The mutation rate averaged over a sliding window of 101 bp is shown as black bars. The gene structure in the region is represented as blue boxes. (b) Manhattan plot of the $-\log_{10}(p$-values) for excess somatic mutations in CRDs. The blue line indicates the 5% FDR threshold. Alternating chromosomes are coloured differently. The triangles represent truncated points. The significant CRDs are abundant and distributed across the genome.
Figure 2 – The clustering of samples based on presence of mutations in significant CRDs. The rows are the CRDs and the columns represent samples. The hierarchical clustering of samples and CRDs are shown as a dendrograms. The clustering is conducted with hierarchical clustering using complete linkage method on the Euclidian distances. This suggests the patterns of selection in non-coding cis-regulatory regions are similar to protein coding sequences with some regulatory regions probably being important for tumorigenesis in most cancers while others being more specific to a few tumours.

Figure 3 – (a) The distributions of estimates of purifying selecting as calculated by LINSIGHT acting on somatic mutations in CRDs with significant excess of somatic mutations vs. ones that are non-significant. Boxplots are truncated at 0.15 probability of purifying selection. There is a significant increase in the purifying selecting acting on mutations in significant CRDs vs. non-significant ones, indicating that functional bases are being impacted in significant CRDs more so than in non-significant ones, suggesting selecting in tumorigenesis. (b) The 14 genes regulated by significant CRDs (red) that show extreme ASE compared to genes regulated by CRDs but do not have an excess of somatic mutations (grey). This result corroborates the driver potential of these genes.
Figure 4 – The mutational burden analysis. (a) The distribution of the direction of the correlation between mutational burden and gene expression. We observed that as the number mutations in a CRD increases this results in an increased gene expression more often than a decrease in gene expression. Two examples of significant associations, the BCL7A gene in (a) and the LPP gene in (b).

Figure 5 – The mutational signatures observed in the regulatory regions of the significant vs. non-significant CRDs (a) and regulatory regions vs. spacers in significant CRDs (b). Bars are coloured by the nucleotide change observed and stratified by the local context. In (a) the proportion of each mutational signature in the regulatory regions of significant CRDs are presented as red bars, whereas the non-significant CRDs are shown as blue bars. In (b) the mutational signatures in regulatory regions are shown in red and the spacers in blue. The numbers above the bars are the Benjamini-Hochberg corrected –log10(p-value) of the observed enrichment or depletion where the significant classes are shown in cyan. In both cases, there are multiple signatures that are significantly different in the respective comparisons, indicating selection during tumorigenesis.
Figure 6 – The transcription factors that are significantly enriched around somatic mutations in significant CRDs vs. the ones in non-significant CRDs. The horizontal black line indicates the base line, and the numbers above the bars are the Benjamini-Hochberg corrected $-\log_{10}(p\text{-value})$ of the observed enrichment.
1 Calling cis regulatory domains (CRDs)

The details of how the ChIP-seq peaks are defined and quantified, and the methodology for calling the CRDs from the correlations among the ChIP-seq peaks are given in Delaneau O. et. al. 2017. In brief, firstly we perform agglomerative hierarchical clustering of the data on a per chromosome basis. Specifically, we start with each chromatin being assigned to its own cluster and iteratively merge clusters together as we move up in the hierarchy. To decide which pair of clusters has to be merged together, we maintain a correlation matrix all along the procedure that initially corresponds to the absolute correlation matrix between chromatin peaks and we search in this matrix for the pair of clusters exhibiting the highest correlation value. This results in a binary tree of histone modifications. Calling CRDs from this binary tree can be viewed as identifying the minimal set of internal nodes that capture most of the overall correlation mass (i.e. cumulative sum of correlation). To retain an internal node N as a CRD, we required two empirical criteria to be fulfilled:

1. Overall correlation. The mean absolute correlation between all possible pairs of chromatin peaks downstream to N must be at least twice as high as the mean correlation between all chromatin peaks on the chromosome. This criterion ensures that CRDs are formed of highly correlated chromatin peaks.

2. Edge correlation. The mean absolute correlation between all pairs of chromatin peaks involving either the first or the last chromatin peak (on the basis of their genomic location) downstream to N has to be at least twice as high as the same value derived for the first and last peaks on the chromosome. This criterion ensures that CRDs have well defined boundaries.

2 Assignment of CRDs to genes

Each CRD is quantified as the mean of the peak quantifications belonging to the CRD in each of the samples, which are then rank transformed into a normal distribution. Transcriptomes are quantified from RNA-seq using QTLtools¹ quan CRD, and the resulting gene quantifications are normalized using 50 and 10 principle components for LCLs and fibroblasts, respectively. Subsequently the gene quantifications are rank transformed into a normal distributions and genes not quantified in 10% of the samples are filtered out. The associations between CRD and gene quantifications are conducted using the QTLtools cis CRD, with a cis window size of 1 Mb and 1000 permutations. All genes associated with the quantifications of a CRD at genome-wide FDR threshold of 5% are considered.

3 Identifying CRDs with significantly excess somatic mutations

3.1 Segmentation of the genome according to AID mutation rate and scaling of AID mutations

There are two types of AID mutations observed in CLL, canonical whose signature is C to T/G mutations at WRCY motifs (W=A or T, R=purine, Y=pyrimidine), and non-canonical whose signature is A to C mutations at WA motifs. First, we categorized each somatic mutation observed in CLL samples into three categories: canonical-AID, non-canonical-AID, or other types of somatic mutations. We then calculated the mutation rates for these categories across the genome averaging over a 100 kb long sliding window with steps of 5 kb. Canonical-AID and non-canonical-AID mutation rates were separately binned into 10 bins distributed equally between the minimum non-zero mutation rate and the maximum mutation rate. Mutation rate of 0 was considered as an additional bin. These bins were used to discretize the data and the bins that contained no observations were dropped. We segmented the genome using a Hidden-Markov-Model (HMM) using these observations. The model had the same number of states as the number of bins used, which signified the segmentation of the
genome from low to high mutation rates. HMM’s start, transition, and emission probabilities were estimated using 10 rounds of Baum-Welch algorithm and it was decoded using the Viterbi algorithm. We then used the labels generated by the Viterbi algorithm to identify contiguous blocks of 100 kb windows with the same label as distinct regions, and these regions were considered to have a similar mutation rate across its entirety. Subsequently, we scaled all the AID mutations falling into one of these regions using the ratio of the non-AID mutation rate in the block to the AID mutation rate in the block. For example, if the canonical-AID mutation rate was 0.04 and the other mutation rate was 0.01, each canonical-AID mutation in this block counted as 0.25 (0.01 / 0.04) mutations. If the other mutation rate was higher than the AID mutation rate, then the AID mutations were counted as 1 mutation.

3.2 Nominal p-value calculation

For each CRD, we calculate a null mutation rate for that CRD’s from its spacers. To this end we find regions of the genome that do not overlap known exons (GENCODE v19 annotation) and the ChIP-seq peaks found in this study, which we label spacers (non-coding non-regulatory regions). In order to account for differential mutation rates observed across the genome in tumorigenesis, we calculate a background (null) mutation rate for a CRD from its local spacers. The local spacers are the spacers that are in between the regulatory regions of a CRD and the two flanking spacers downstream and upstream of the terminal regulatory regions. Furthermore, we increase the number of spacers so that the total length of the spacers is at least as long as the total length of the regulatory regions of a CRD, by progressively including more flanking spacers. To calculate the background mutation rate from the spacers we take local context of the mutation into account. The local context is the reference allele of the somatic mutation and the immediate 5’ and 3’ bases. The 64 possible trinucleotide contexts are represented as changes in the pyrimidines. We count the number of mutations observed in spacers for each of the 32 triplet contexts, and the total number of these triplets. We then calculate a mutation frequency for each of the triplet contexts. Subsequently we count the number of triplets observed in the regulatory regions and multiply these with the corresponding mutational frequencies calculated from the spacers to come up with the expected number of mutations. Finally, we calculate a nominal p-value with the Poisson distribution using the observed and expected mutation counts (one tailed test).

3.3 Permutation scheme

The previous section describes a method that accounts for differential mutation rate across the genome, however, we also need to control for the differential mutation rates between open and closed chromatin. To this end, we devised a permutation scheme. For each base in the regulatory region, we select a random base in the regulatory region of a random CRD with the same context, and a base with the same context from the downstream spacer of the same random CRD. We then check if these pair of random bases are mutated, and we iterate over all the bases in the regulatory regions. This enables us to create a random pseudo CRD with mutations, which conserves the differential mutation rates between open and closed chromatin but breaks the clustering of mutations. We do this 100000 times for each CRD and at each iteration, we calculate a p-value for the random pseudo CRD created using the same method described in the previous section. Finally, we count the number of times we observed the same or more significant p-value in the 100000 iterations than the nominal p-value of the CRD in question, which is used to calculate the adjusted p-value. The adjusted p-values are corrected for multiple testing using the qvalue R package, and significant CRDs at 5% FDR are reported.
3.4 Details of how we account for known biases

Differential mutation rate across the genome
It is known that due to biological processes like replication time, the somatic mutation rate across the genome is not constant, and certain regions are likely to accumulate more somatic mutations. To address this bias we are calculating a background mutation rate for each CRD separately from its local spacers thus accounting for the difference in local mutation rates.

Different lengths of regulatory regions and spacers of a CRD
The total length of regulatory regions and the spacers of CRD may differ. This would pose a problem if the spacer regions were substantially shorter than the regulatory regions. In this case, our estimates of the background somatic mutation rate would be noisier than the mutation rate in the regulatory regions, since we are using the local spacers to estimate a background mutation rate for each CRD. To remedy this issue we progressively increase the number of spacer regions for a CRD by adding more flanking spacers, until the total length of the spacers is at least as long as the regulatory regions. This ensures that our estimates of background and observed mutation rates are comparable.

Different mutational signatures in cancers
Different types of cancer have different probabilities of a base being mutated given its local sequence context, referred to as mutational signatures. These are defined as the somatic mutation’s trinucleotide context, the base that is mutated and the immediate two flanking bases. This would create a bias if we do not consider the local context of a somatic mutation. If a stretch of DNA sequence contains a higher proportion of triplets that are more likely to be mutated in a given cancer, this region would seem to accumulate more mutations than a region, which is made up of triplets less likely to be mutated. Thus, we calculate a separate background mutation rate of each trinucleotide from a CRDs spacers and combine the trinucleotide mutation rates with its DNA sequence in the regulatory regions to calculate the total number of expected mutations.

Differential mutation rates between open and closed chromatin
The regulatory regions of a CRD are open chromatin regions whereas its spacers are closed chromatin, and we know that open and closed chromatin have different rates of somatic mutations. To account for this we devised a permutation scheme detailed in the previous section, which we run for each CRD separately. In short, this scheme creates a pseudo CRD at each iteration based on the sequence makeup of the CRD in question. We pick two random positions in the regulatory and spacer region of a random CRD and ask if these are mutated, and repeat this for all the bases of the CRD in question. This allows us to create a pseudo CRD, which preserves the differential mutation rate between open and closed chromatin, but breaks the clustering of mutations.

3.5 Selection of non-significant CRDs
We defined a stringent set of non-significant CRDs in both cancers by taking the CRDs that contained at least one somatic mutation in their regulatory regions, and had nominal and adjusted p-values greater than 0.4, resulting in 1271 CRDs in CLL and 10881 non-significant CRDs in CLL and skin cancer, respectively.

4 Significantly enriched pathways
ConsensusPathDB\(^3\) was used to find significantly enriched pathways using the Reactome\(^4\) and KEGG\(^5\) pathway databases. The pathways reported as significant have an FDR of 5% in CLL and due to the lower power in skin cancer the FDR is set at 20%.
5 Quantifying purifying selection action on somatic mutations.
The genome-wide estimates of purifying selection acting on each base of the human genome as calculated by LINSIGHT were downloaded as a BigWig file. For both CLL and skin cancer, two BED files were created, one with the somatic mutations observed in the regulatory regions of the significant CRDs (FDR = 5%) and one with the mutations in regulatory regions of non-significant CRDs (adjusted p-value > 0.4) with at least one somatic mutation. Finally bigWigAverageOverBed tool was used to calculate maximum probability of purifying selection acting on each of the mutations.

6 Assessment of mutational signatures
For each of the CRDs, we calculated the number of times a specific single nucleotide change has happened. In order to classify these into mutational signatures, we considered the six possible substitutions of the pyrimidine of the mutated Watson–Crick base pair (C to A, C to G, C to T, T to A, T to C, T to G) and the immediate 5’ and 3’ prime bases resulting in 96 mutational signatures. We then calculate the proportion of each of the mutational signatures in CRDs’ regulatory regions and their spacers. When we are assessing the enrichment of the mutational signatures found in the regulatory regions over the ones found in their spacers, we calculate the odds ratio of these two proportions, and test the significance of this odds ratio using the Fisher exact test (two-tailed test). The p-values obtained are subsequently corrected for multiple testing using Benjamini-Hochberg method.

7 Enrichment of transcription factor binding sites
We used the HOMER software package to identify transcription factor binding sites that are enriched around somatic mutations in CRDs with significantly excess of somatic mutations over the somatic mutations in non-significant CRDs. To this end, we created a set of sequences that were 31 bp in length that were centred around the somatic mutation position with 15 bp flanking it. We generated these sequences for the somatic mutations observed in significant CRDs and the non-significant CRDs. We then searched for transcription binding sites that are enriched in the sequences from the significant CRDs using the non-significant CRDs’ signatures as background. This was achieved using the known vertebrate transcription factor binding sites provided by the HOMER package.
Supplementary Figure 1 – a) The data types found in the SysGenetiX project. b) An example of the correlation structure among the chromatin marks in the population, where darker shades of blue represent higher correlation between pairs of chromatin marks and the cis regulatory domains (CRDs) identified depicted as black triangles.
Supplementary Figure 2 – The algorithm for calculating the nominal p-value for an excess of somatic mutations in a cis regulatory domain (CRD). The red boxes are the regulatory regions of the CRD and the blue lines represent the spacers from which the expected mutation rate for a given tri-nucleotide context is calculated. The yellow stars are somatic mutations.

1. Calculate the expected mutation rate from spaces with immediate 5’ and 3’ taken as context.

2. Given the base make up of the Regulatory regions calculate the expected number of mutations, using the null mutation rates of each tri-nucleotide context.

3. Calculate a p-value for excess somatic mutations observed in regulatory regions using a one-tailed Poisson distribution.
1. For each base in the regulatory region, select a random base in the regulatory region of a random module with the same context, and a base with the same context from the downstream spacer of the same random module.

2. Check if these pair of random bases are mutated, and iterate over all the bases in the regulatory region.
Supplementary Figure 4 – The cis-regulatory domain (CRD) in LCLs that is regulating the expression of the BIRC3 gene and has a significantly excess of somatic mutations in CLL. The red boxes represent the regulatory regions of the CRD, and the blue boxes are the spacers. The mutation rate calculated with a sliding window size of 101 bp is given as black bars. The gene structure in the region is represented by the dark blue boxes. This plot is generated using UCSC genome browser.
Supplementary Figure 5 – The cis regulatory domain (CRD) in fibroblasts that is regulating the expression of the LRRC37A3 gene and has a significantly excess of somatic mutations in skin cancer. The red boxes represent the regulatory regions of the CRD, and the blue boxes are the spacers. The mutation rate calculated with a sliding window size of 101 bp is given as black bars. The gene structure in the region is shown by the dark blue boxes. This plot is generated using UCSC genome browser.
Supplementary Figure 6 – The distributions of estimates of purifying selecting as calculated by LINSIGHT acting on somatic mutations in CRDs with significant excess of somatic mutations overlapping the IGHV locus vs. ones that are not-significant in the whole genome in (a) and only overlapping the IGHV locus in (b).
Supplementary Figure 7 – Number of somatic mutations observed per sample in CLL (a) and skin cancer (b). The two distributions are significantly different from each other (Mann-Whitney p < 2.2e-16).
**Supplementary Figure 8** – The mutational signatures observed in the regulatory regions vs. their spacer. Bars are coloured by the nucleotide change observed and stratified by the local context. In (a) and (c) the proportion of each mutational signature in the regulatory regions are presented as darker shades whereas the spacers are shown as lighter shades of CLL and skin cancer, respectively. In (b) and (d) odds ratio, proportion in regulatory over proportion in spaces, of a given mutational signature is plotted for CLL and skin cancer, respectively. The numbers above the bars are the Benjamini-Hochberg corrected $-\log_{10}(p\text{-value})$ of the observed enrichment or depletion where the significant classes are shown in red.
Supplementary Figure 9 – The distributions of number of bases overlapping the somatic mutation calls in the regulatory and spacer regions of the significant CRDs in CLL (a) and skin cancer (b). In both cancers, the coverage in regulatory regions is not significantly greater than the coverage in the spacer regions (one-tailed Mann-Whitney U test).
Supplementary Figure 10 – The correlation of p-values and odds ratios when including or excluding flanking spacers in the analysis for CLL in (a) and skin cancer in (b). The significant CRDs are shown and odds ratio is the mutation rate in regulatory regions over the mutation rate in corresponding spacers.
Supplementary Figure 11 – Statistics of significant vs. non-significant CRDs in CLL and skin cancer. (a) The distributions of the average correlation among the CHiP-seq peaks of the CRDs. (b) The distributions of genomic lengths of significant and non-significant CRDs including their internal spacers. (c) The distributions of the number of non-overlapping CHiP-seq peaks in significant CRDs vs. the non-significant ones. The boxplots are truncated. (d) Density plot of the genomic distance between the CRDs and the genes they regulate for significant and non-significant CRDs. (e) The distributions of the absolute value of the genomic distance between the CRDs and the genes they regulate. (d) The number of genes the significant and non-significant modules regulate.
Supplementary Figure 12 – The canonical AID mutation rate across the genome is given as a Manhattan plot, where alternating chromosomes are coloured as blue and grey. This was calculated using a sliding window size of 100 kb with a step size of 5 kb. In each window, to calculate the mutation rate, we counted the number of positions with the WRCy motif and how many of them overlapped with a canonical AID mutation. The genome was segmented based on this mutation rate using a HMM and the distinct regions that were labelled as having a similar mutation rate are presented as the coloured rectangles above the Manhattan plot. Each different colour corresponds to a distinct region with similar mutation rate.
Supplementary Figure 13 – The non-canonical AID mutation rate across the genome is given as a Manhattan plot, where alternating chromosomes are coloured as blue and grey. This was calculated using a sliding window size of 100 kb with a step size of 5 kb. In each window, to calculate the mutation rate, we counted the number of positions with the WA motif and how many of them overlapped with a non-canonical AID mutation. The genome was segmented based on this mutation rate using a HMM and the distinct regions that were labelled as having a similar mutation rate are presented as the coloured rectangles above the Manhattan plot. Each different colour corresponds to a distinct region with similar mutation rate.
Leveraging interindividual variability of regulatory activity refines genetic regulation of gene expression in schizophrenia

Maris Alver\textsuperscript{1,2,3,4}, Nikolaos Lykoskoufis\textsuperscript{1,2,3}, Anna Ramisch\textsuperscript{1,2,3}, Halit Ongen\textsuperscript{1,2,3}, Emmanouil T. Dermitzakis\textsuperscript{1,2,3,*}

1 Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland
2 Swiss Institute of Bioinformatics, University of Geneva, Geneva, Switzerland
3 Institute of Genetics and Genomics in Geneva, University of Geneva, Geneva, Switzerland
4 Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia

*Corresponding authors: M.A.: maris.alver@unige.ch and E.T.D.: emmanouil.dermitzakis@unige.ch
Department of Genetic Medicine and Development, University of Geneva
1 rue Michel-Servet, 1211 Geneva, Switzerland

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
ABSTRACT

Schizophrenia is a polygenic psychiatric disorder with limited understanding about the mechanistic changes in gene expression regulation. To elucidate on this, we integrate interindividual variability of regulatory activity with gene expression and genotype data captured from the prefrontal cortex of 272 cases and controls. We show that regulatory element activity is structured into 10,936 and 10,376 cis-regulatory domains in cases and controls, respectively, which display distinct regulatory element coordination structures in both states. By studying the interplay among genetic variants, gene expression and cis-regulatory domains, we ascertain that changes in coordinated regulatory activity tag alterations in gene expression levels ($p=8.62 \times 10^{-06}$, $OR=1.60$), unveil case-specific QTL effects, and identify regulatory machinery changes for genes affecting synaptic function and dendritic spine morphology in schizophrenia. Altogether, we show that accounting for coordinated regulatory activity provides a novel mechanistic approach to reduce the search space for unveiling genetically perturbed regulation of gene expression in schizophrenia.
INTRODUCTION

Schizophrenia (SCZ) is a severe mental illness that oftentimes leads to a lifetime of chronic disability. While several lines of evidence converge on the neurodevelopmental origin for the disorder and augment the impact of both genetic and environmental factors in disease aetiology, the pathophysiology of SCZ still remains incompletely understood with little progress in novel treatment development.\(^1,2\)

The last decades of extensive research has yielded valuable insights into the genomic and molecular underpinnings of SCZ. Large-scale genomic analyses triggered by prior studies of strong heritability estimates for SCZ (60-80%)\(^3,4\) have unveiled its highly polygenic architecture\(^5-7\). Complemented by gene expression and chromatin profiling analyses from dorsolateral prefrontal cortex (DLPFC), the SCZ risk variants have been shown to localize in functional regulatory genomic elements with approximately half of these displaying brain tissue-specific expression quantitative trait locus (eQTL) effects\(^5,6,8-10\), and to be enriched for open chromatin and evolutionary conserved regions\(^11,12\). Genes identified by differential expression and genome-wide association analyses (GWAS) are associated with brain developmental pathways, synaptic function, and immune response\(^13-15\). While immense efforts have resulted in comprehensive SCZ-specific resources for perturbed gene expression patterns and for fine-mapping and annotating discovered disease-associated genomic associations, little is known about the regulatory machinery changes and how genetic effects are propagated onto gene expression that drive molecular abnormalities in SCZ development.

Gene transcription profiles are defined by the activity of regulatory elements (REs) that overlap with open chromatin. Since the activity of REs are modulated by genetic variation, changes in chromatin accessibility result in gene expression variability and hence constitute as an intermediate phenotype for profiling eQTL effects on gene expression\(^16-20\). Systematic measurement of interindividual correlation between chromatin activity levels has revealed that the variability of nearby regulatory activity is structured into well-delimited sets of cis-regulatory domains (CRDs)\(^21\). The coordinated activity of REs within CRDs are under tight genetic control, mediate cis and trans effects of genetic variants onto gene expression and provide a higher-order structural resolution of functional regulatory associations\(^21\). Accounting for the three-dimensional (3D) genome organization in cis that captures concerted effect of regulatory activity could thereby facilitate a more robust signal detection for identifying disruption in regulatory function and for delineating deviations in gene expression cascades specific to disease. To build on this concept, we set out to analyse the interplay among genetic variants, coordinated regulatory activity and gene expression to characterize genetically perturbed regulatory machinery changes specific to SCZ. To this end, we integrated genome-wide genotyping data with RE activity levels (chromatin immunoprecipitation sequencing (ChIP-sequencing profiled for histone mark H3K27ac) and transcriptomic profiles (bulk RNA-sequencing) obtained from the DLPFC of SCZ cases and control subjects (dbGaP phs000979.v3.p2). At least two levels of
molecular data were available for 272 individuals: 98 SCZ cases and 174 controls (188 males and 84 females, 164 African Americans and 108 Europeans; Supplementary Fig. 1).

RESULTS

Distinct regulatory element coordination in SCZ

To study the coordination of REs, we systematically measured interindividual correlation between nearby chromatin peaks. We identified 10,938 CRDs in SCZ cases, 10,376 CRDs in controls and 11,374 CRDs in the combined set (i.e., across SCZ cases and controls), regrouping 28.9% (n=40,819), 31.4% (n=44,391) and 38.4% (n=54,278) of the peaks, respectively (Table 1), and capturing a higher-order structural resolution of regulatory activity. The majority of the CRDs contained two REs, while some captured correlated activity among >80 REs (mean number of REs per CRD 4.7; mean CRD length 138 kb in the combined set; Supplementary Fig. 6). As expected, we identified a high concordance of CRDs between SCZ cases and controls (42%; Supplementary Fig. 7), which represent a uniform coordination of regulatory activity in DLPFC in SCZ cases and controls. However, 58% of the CRDs showed distinctive RE coordination in SCZ cases and in controls, implying notable state-dependent differences in the regulatory machinery.

Given unique RE coordination in SCZ cases and controls, we sought to investigate the mechanism for SCZ-specific CRD formation. Specifically, we asked whether the chromatin peaks within SCZ-specific CRDs were differentially active or whether these had larger variance in activity in controls compared to SCZ cases. To this end, we considered CRDs in SCZ cases composed of peaks not part of any CRD in controls and compared single peak activities and mean correlation estimates among peaks per CRD between the two groups. We discovered that 53% of the peaks within SCZ-specific CRDs (3,540 peaks in 2,212 CRDs) were differentially active in SCZ cases at false discovery rate (FDR) 5% (Supplementary Fig. 8a). While the majority of the peaks (71%) showed lower activity in SCZ cases (Fig 1a), only a third of SCZ-specific CRDs had all underlying peaks differentially active between SCZ cases and controls (Supplementary Fig. 8b), implying that the regulatory activity originating from those genomic regions likely results in inhibition of downstream molecular cascades. The peaks of SCZ-specific CRDs displayed significantly higher mean correlation in SCZ cases compared to controls (Mann-Whitney U test p = 5.02e-47; Fig. 1bc), indicating that changes in the 3D structure of the genome, rather than differential activity, were responsible for SCZ-specific CRD formation. At FDR 5%, we identified eleven SCZ-specific CRDs to be associated with the expression of proximal genes, for example POU3F1 (also known as OCT6, transcriptional repressor for myelin-specific genes22), KIF5A (neuronal-specific vesicular transporter23), NECAB1 (Ca2+-binding in neurons24), and PDCD1LG2 (immune checkpoint receptor ligand25) (Supplementary Table 1). These associations exemplify coordinated regulatory changes specific to the disease state that affect or are affected by gene expression perturbations.
Changes in CRD activity track alternations in gene expression in SCZ

Having identified several distinct RE coordination structures in DLPDC in SCZ cases that were absent in controls, we focused next on CRDs that had the same structure of RE coordination across SCZ cases and controls (i.e., CRDs identified in the combined set). We set out to determine differences in regulatory activity between SCZ cases and controls and investigated their relation to genes that were differentially expressed (DEGs) between the two groups. At FDR 5%, we identified 1,141 CRDs (599 lower activity, 542 higher activity) and 1,363 genes (937 up-regulated, 426 down-regulated) to be differentially active and expressed in SCZ cases, respectively (Table 1, Supplementary Fig. 9, Supplementary Table 2, Supplementary Table 3). The differences in effect size for CRDs were subtle, reflecting a narrow variability range in regulatory activity (Supplementary Fig. 9a). The determined DEGs were in concordance with those previously reported in SCZ pathogenesis (Supplementary Fig. 10) and were significantly enriched for gene ontology (GO) terms related to sex-hormone and interferon-γ-mediated signalling, glucocorticoid receptor and glutamate receptor binding, axonogenesis and synapse assembly (Supplementary Table 4). The DEGs were significantly enriched for differentially active CRDs (Fisher’s exact test $p = 8.62e-06$, odds ratio 1.60) (Fig. 2a) with the majority of the genes (86%) showing the same direction of effect as the CRD in which the gene transcription start site (TSS) lied, indicating that deviations in gene expression track alterations in the regulatory machinery.

In addition to ascertaining that differentially active CRDs localized within the genomic proximity of DEGs, we additionally tested for direct association between gene expression and CRD activity (i.e., CRDs identified in the combined set). At FDR 5%, we identified in cis 95, 634 and 1,197 CRD-gene associations in SCZ cases, in controls, and in the combined set, respectively (Table 1, Supplementary Fig. 11, Supplementary Table 6). The majority of the genes were associated with a single CRD and the majority of the CRDs with a single gene with only a handful of CRDs being associated with up to ten different genes (Supplementary Fig. 12). Most gene TSSs clustered at CRD boundaries (Fig. 2bc), corroborating the proximal role of coordinated regulatory activity in gene transcription.

Genetic regulation of CRD activity and gene expression in SCZ

We next sought to study the genetic regulation of CRD activity and gene expression, search for SCZ-specific QTL effects and interrogate whether QTL effects colocalize with SCZ risk variants. At 5% FDR and in cis, we discovered 796 and 2,929 functionally independent CRD activity QTLs (aCRD-QTLs), and 867 and 6,166 functionally independent eQTLs in SCZ cases and controls, respectively (Table 1, Supplementary Table 6, Supplementary Table 7, Supplementary Fig. 13). The strength of the association was correlated with the genomic distance from the molecular phenotype (Supplementary Fig. 14). While almost all SCZ-identified QTL effects replicated in controls (Supplementary Fig. 15), 5% of aCRD-QTLs (n=40) and 15% of eQTLs (n=130) showed SCZ-specificity, i.e., these affected CRD activity or gene expression only in SCZ cases or displayed significant change in effect size compared to controls (Table 1, Fig. 3ab). The SCZ-specific
genotype-dependent variability in CRD activity and in gene expression imply context-dependent and pathway-activated gain in regulatory capacity. Results of gene enrichment analysis for genes associated with SCZ-specific eQTLs conform with posed hypotheses linking dysregulation of retinoid binding and adenosine deaminase activity with SCZ\textsuperscript{27,28} (Supplementary Table 8). Colocalization analyses for SCZ risk variants with aCRD-QTLs and with eQTLs showed modest yet proportionally similar enrichment for shared functional effects (1.6% for SCZ-identified aCRD-QTLs and 1.8% for SCZ-identified eQTLs). Interestingly, the aCRD-QTLs colocalized with different GWAS variants compared to eQTLs that shared a functional effect with SCZ risk variants (Supplementary Table 9).

To assess common genetic regulation of coordinated regulatory activity and gene expression, we investigated the association of aCRD-QTL and eQTL effects on the other molecular phenotype. Specifically, we correlated aCRD-QTLs with gene expression and eQTLs with CRD activity over gene-CRD associations detected across all samples at nominal significance. We identified that up to 46% of the eQTL and aCRD-QTL variants had an effect on CRD activity and gene expression, respectively (Supplementary Fig. 16). The considerable overlap between aCRD-QTL and eQTL effects for relevant CRD-gene pairs corroborates the functional interplay among genetic variants, CRDs and genes.

**Refining eQTL perturbations reveals regulatory machinery changes specific to SCZ**

Given the established interplay among genetic variants, genes and CRDs, we interrogated the functional directionality between them. We reasoned that the effect of a genetic variant on gene expression could either be mediated by or propagated to the changes in coordinated RE activity and that deviations in the regulatory machinery in SCZ cases compared to controls would imply molecular dysregulation specific to disease. To test this, we considered the previously discovered 1,197 CRD-gene pairs ascertained across SCZ cases and controls at FDR 5% and identified the same genetic variant (eCRD-QTL) that affected both molecular phenotypes and by that determined eCRDQTL-CRD-gene triplets for causal inference (Methods). Using Bayesian Networks (BN), we tested three relationship patterns: i) causal model in which the genetic variant affects first the CRD activity which then regulates the gene expression, ii) reactive model in which the genetic variant affects the gene expression which modulates the CRD activity, and iii) independent model in which the genetic variant affects the gene and the CRD independently; and studied these relationships separately in SCZ cases (n=59) and controls (n=105) (Supplementary Fig. 17). We discovered that at FDR 5%, 91.9% of the CRD-gene pairs had a cis-QTL effect (n=1,100; Table 1), indicating that the simultaneous change in CRD activity and gene expression was affected by the same nearby genetic driver. We observed more causal models in controls than in SCZ cases (Supplementary Fig. 18ab), which were likely driven by smaller SCZ sample size as reflected by the distribution of the probabilities for the most likely model for each triplet (Supplementary Fig. 18cd). The probability of the causal model increased the further the gene TSS was from the eCRD-QTL in both SCZ cases and controls (Supplementary Fig. 18ef), denoting the role of CRDs mediating the genetic effect onto distal genes.
To study the proportion of differential regulatory mechanisms between SCZ cases and controls and ascertain which molecular functions were affected by these changes, we first estimated the accuracy for BN results using bootstrapping to provide confidence for retrieved probabilities and next carried out gene enrichment analyses for genes associated with different regulatory mechanisms. The accuracy for inferring the most likely causal relationship for triplets was lower in SCZ cases (mean accuracy estimation 69.0%, sd = 16.6) than in controls (mean accuracy estimation 75.0%, sd = 18.0) (Supplementary Fig. 19ab). To exclude ambiguous signals, we proceeded with triplets that surpassed the accuracy estimation of 55% (748 triplets, 68% of studied triplets; Methods, Supplementary Fig. 19c, Supplementary Table 10). While two-thirds of the triplets displayed the same regulatory mechanism in SCZ cases and controls (Fig. 4a), one-third of studied triplets showed a change in directional effect from QTL variant onto molecular phenotype (Fig. 4b, Supplementary Fig. 19d). These deviations in regulatory mechanism in SCZ reflect gain or loss in the regulatory capacity that could either be driven by context-dependent or genetically predisposed developmental derailment of gene expression, or affected by external stimuli (e.g., treatment). The genes associated with change-associated triplets were significantly enriched for GO terms related to small GTPase binding, filopodium assembly and cellular lipid catabolic process (Supplementary Table 11), highlighting alterations in the regulatory machinery for gene expression affecting synaptic function and plasticity, and dendritic spine morphology in SCZ.

Perturbation in the regulatory mechanism of gene expression in SCZ is exemplified by a triplet consisting of an eCRD-QTL 7:5849993:C:G associated to gene FSCN1 and to a CRD composed of 5 REs (chr7:5623132-5705414) (Supplementary Fig. 20). FSCN1 is an actin-binding protein that is required for the formation of actin-based cellular protrusions and affects dendritic spine morphology29-32. Based on BN, the genetic variant affected first the activity of the CRD and then the expression of the gene in controls (causal model probability 0.99), whereas in SCZ cases the change in CRD activity was a reaction to gene expression (reactive model probability 0.77), indicating that the eCRD-QTL effect on FSCN1 expression was not mediated via the associated CRD activity in SCZ cases as seen for controls (Fig. 4d). This was further supported by significant downregulation of FSCN1 expression and CRD activity in SCZ cases compared to controls (FDR 5% p-value 0.02 and 0.005, respectively; Supplementary Fig. 20ab, Supplementary Table 2, Supplementary Table 3). Moreover, we identified an opposite direction of eCRD-QTL effect on FSCN1 expression in SCZ cases compared to controls (Fig. 4c), but not on CRD activity (Supplementary Fig. 20c).

Interestingly, this association was identified only in individuals with African American ancestry (Supplementary Fig. 20de) as the genetic variant was completely monomorphic in HBCC Europeans. The MAF spectrum of 12% in HBCC African Americans and 0% in HBCC Europeans is in concordance with population frequencies estimated in larger datasets (MAF 13% in Africans/African Americans and 0.4% in non-Finnish Europeans)33. These results indicate that the downregulation of FSCN1 expression in SCZ cases was driven by a different intermediary regulatory mechanism that deviated from one seen in controls and represents a dysregulated step within an abnormal molecular cascade affecting dendritic spine morphology in SCZ.
DISCUSSION

Deciphering regulatory mechanisms of gene expression that reflect molecular perturbation in SCZ are under extensive scrutiny yet are hindered by the complexity of the SCZ phenotype and scarcity of relevant molecular data. Studying regulatory activity that tracks changes in gene expression requires a higher order analysis approach for signal discernment due to narrow variability range in regulatory activity and extensive multiple testing burden\(^\text{12}\). Here we show that taking account of interindividual correlation between regulatory activity allows to refine changes in gene expression specific to disease, asserting that disease manifestation stems from dysregulated gene expression cascades that are steered by and propagated to the concerted action of REs. Interrogation of common genetic regulation of gene expression and CRD activity corroborated that correlated changes in gene expression and CRD activity are affected by the same genetic driver. Our results agree with findings showing considerable overlap between QTL effects on chromatin accessibility and gene expression\(^\text{12}\), that a single genetic variant drives the association between multiple chromatin peaks and a single gene\(^\text{20}\), and on the convergence of deviations detected in different molecular layers as seen for gene expression and methylation in SCZ\(^\text{34}\). Applying causal inference to study the causal relationships among genetic variants, genes and CRDs revealed regulatory machinery changes affecting synaptic function and dendritic spine morphology in SCZ which are in line with established molecular abnormalities identified for the disorder\(^\text{35-39}\). The deviations in regulatory mechanism reflect gain or loss in the regulatory capacity that could either stem from genetic predisposition, are acquired in disease progression or result from chronic pharmacology. Clear discernment of the proposed origins of effect was hampered due to small sample size and unavailability of relevant data yet allowed to draw the following conclusions.

First, while we found modest colocalization for detected QTLs with SCZ-predisposing genetic variants, the results reflect direct correlation between sample size and QTL signal detection\(^\text{40}\) and hence are in proportion to colocalization signals ascertained in previous findings\(^\text{9,12,26,41}\). Previous studies have highlighted the concordance of SCZ heritability enrichment for open chromatin regions in fetal and in SCZ DLPFC samples and the stability of methylation and expression feature deviations in fetal brain development that persist into adulthood for those affected by the disorder\(^\text{9,12,13,42}\), implying that a considerable proportion of signals detected in the current analysis do reflect brain development derailment due to genetic predisposition for SCZ. Second, while molecular data used in this study was extracted from bulk tissue, consistent comparison with signals identified in control samples provides confidence that the identified deviations in SCZ cases captured the most notable disease-specific molecular abnormalities in DLPFC. This is further supported by investigations revealing that DLPFC transcriptomic profiles are generally biased toward neuronal populations\(^\text{34}\) and that SCZ risk variants are overrepresented in neuronal vs non-neuronal open chromatin regions\(^\text{43}\). Third, studying QTL effects on gene expression and CRD activity separately in SCZ cases and controls allowed to discriminate context-dependent genetic effects on both molecular phenotypes, indicating gain in regulatory capacity that translated into gene expression and coordinated regulatory activity variability.
only in SCZ cases or showed significantly different or even discordant effect direction, as seen for FSCN1 gene, between the two groups. Fourth, a previous study that showed concordant DEG signals with those found in the current analysis, identified that differential gene expression in SCZ was not driven by antipsychotic intervention\cite{26}, ensuring that treatment effect was not the main trigger for DEG results between SCZ cases and controls in the current analysis. Lastly, inclusion of individuals of European and African American ancestry augmented signal identification and corroborates that the genetic basis of SCZ and its biology are broadly shared across populations\cite{7,44,45}.

Altogether, we have outlined that leveraging higher-order structural resolution of regulatory activity allows to reduce the search space for unveiling genetically perturbed regulation of gene expression specific to SCZ. We anticipate that cell-type specific gene expression and open chromatin exposure profiles in larger sample sets would allow better delimitation of CRD chromatin peak content, facilitate the identification of trans-regulatory hubs across different chromosomes as well as enhance more robust detection of origin effect for gene expression deviation, thereby increasing our understanding of perturbed functional pathways underlying schizophrenia and for prioritizing targets for experimental investigation and novel treatment development.
METHODS

Molecular and phenotype data

Molecular and phenotype data for the Human Brain Collection Core (HBCC) was accessed through dbGaP (study accession phs000979.v3.p2; request #88083 approved by NIH on January 31, 2020). All patients met DSM-IV criteria for a lifetime Axis I diagnosis of psychiatric disorders including schizophrenia or schizoaffective disorder. Controls had no history of psychiatric diagnoses or addictions. At least two levels of molecular data (i.e., whole-genome genotype, RNA-sequencing or ChIP-sequencing data) were available for 272 individuals. For genotype data, we determined the intersect of single nucleotide variant (SNV) content across three Illumina genotyping arrays (HumanHap650Y, Human1M-Duov3 and HumanOmni5M-Quad) after filtering the SNVs using standard procedure with PLINK v2.046 and imputed the derived genotype matrix using the Haplotype Reference Consortium reference panel47. Next, we applied post-imputation quality control filters by European and African American ancestral group separately and considered the union of filtered SNVs retrieved in both ancestry sets for the final SNV set. This yielded 8,245,179 biallelic SNVs in 272 individuals. Sequence data was mapped onto the human genome (hg19) with either BWA-MEM v0.7.1648 for ChIP-sequencing data or STAR49 for RNA-sequencing data. Gene expression was quantified using QTLtools50 and filtered for protein-coding and lincRNA genes and for the union quantifications detected in ≥50% in SCZ cases and in ≥50% in controls. This yielded 21,988 genes for 243 individuals. ChIP-sequencing peak calling and quantification was carried out with HOMER v14.11.151. We first determined ChIP-sequencing peak coordinates across SCZ cases and controls to get a population scale call set of ChIP-sequencing peaks and then quantified the peaks for each individual according to the identified peak coordinates. This yielded 141,219 ChIP-sequencing peaks for 193 individuals. To account for confounding factors in gene expression and ChIP-sequencing peak data, we regressed out ancestry (captured by principal component (PC) analysis on the genotype data), and technical variables (captured by PC analysis on the molecular phenotype data). For the latter, we used the number of PCs that maximized the number of QTLs discovery. Both gene expression and ChIP-sequencing data were normalized such that these matched a normal distribution with mean 0 and standard deviation 1.

Cis-regulatory domain calling and quantification

For CRD calling, we used the pipeline developed in Delaneau et al. 201921. First, we built a correlation map from chromatin data by systematically measuring interindividual correlation (i.e., Pearson correlation coefficient) between all possible pairs of ChIP-sequencing peak quantifications located on the same chromosome (within a 250-peak sliding window). Next, we applied hierarchical clustering on the data on a per chromosome basis to get a binary tree that regroups chromatin peaks for each chromosome depending on the correlation levels they exhibited. We relied on three empirical criteria for CRD calling: (i) overall correlation that requires the mean level of correlation within a CRD to be at least twice the background, (ii) edge correlation that requires the mean level of correlation of
the peaks at the CRD boundaries to be at least twice the background, and (iii) a requirement that the CRD covers at least two nonoverlapping regulatory elements. We quantified CRD activity on a per-individual basis by enumerating all ChIP-sequencing peaks per CRD and taking the mean of all single peak quantifications per individual to retrieve a single quantification value for each individual. CRDs called in SCZ cases were used for characterizing SCZ-specific CRDs. CRDs identified in the combined set were used for the rest of the downstream analyses.

**CRD structure analysis**

For determining CRD sharing between SCZ cases and controls, we compared ChIP-sequencing peak correlation maps between SCZ cases and controls and called a CRD shared if ≥50% of the peaks overlapped. To assess the features of SCZ-specific CRDs, we considered only CRDs in SCZ cases composed of peaks not regrouping into any CRD in controls. For peak activity estimation, we used ChIP-sequencing peak quantifications uncorrected for covariates and applied a Mann-Whitney U test per peak activity between SCZ cases and controls. Significant differences between SCZ cases and controls were determined at FDR 5% using R/qvalue package. For estimating peak activity correlation per SCZ-specific CRD between SCZ cases and controls, we used ChIP-sequencing peak quantifications corrected for biological and technical covariates, calculated the mean Pearson correlation estimate between peak activities per CRD and used Mann-Whitney U test for comparing these correlation estimates between SCZ cases and controls.

**Differential CRD activity and differential gene expression analysis**

Differential CRD activity analysis and differential gene expression analysis were carried out using DESeq2. Significant associations were determined at FDR 5%. For differential CRD activity analysis, we used unnormalized ChIP-sequencing peak read counts obtained with HOMER and summed these up per CRD using the correlation map identified in the combined set. For differential gene expression analysis, we used RNA-sequencing read counts. To identify covariates for correction, we carried out an association testing i) between all available biological and technical covariates and disease status (SCZ/CTL) (Mann-Whitney U test), and ii) between all available biological and technical covariates and individual ChIP-sequencing peak activity or gene expression quantifications (linear regression) and calculated r² estimate to identify the proportion of true associations. R/clusterProfiler package was used for gene set enrichment analysis.

**Mapping molecular quantitative trait loci (QTLs)**

QTL mapping was carried out using the standard procedure implemented in the QTLtools software package. Specifically, we performed 1,000 permutations to correct for the number of genetic variants being tested in cis per molecular phenotype (+/- 1 Mb window) and corrected for the number of molecular phenotypes being tested genome-wide using false discovery rate (FDR). To identify multiple QTLs with independent effects on a molecular phenotype, we used the conditional analysis approach based on a forward-backward scan implemented in QTLtools. For SCZ-specific
QTL discovery, we considered QTL effects identified in SCZ cases and for each variant-phenotype pair ran a linear regression including genotype, disease status (SCZ/CTL), and covariates, and tested for significance of a genotype * disease status interaction on molecular phenotype (gene expression or CRD activity). This was followed with FDR 5% correction for the number of QTLs tested. We assessed the likelihood of a shared functional effect between SCZ risk variants from four GWAS studies\(^5\)\(^-\)\(^7\),\(^45\) and SCZ-identified QTLs using regulatory trait concordance (RTC)\(^53\),\(^54\).

**CRD and gene association**

We used QTLtools cis permutation pass\(^50\) to identify CRDs associated with a gene in a +/-1 Mb window from a transcription start site of a gene. We performed these analyses i) to identify genes associated with SCZ-specific CRDs using CRDs identified in SCZ cases, and ii) to capture comparable associations for SCZ cases and controls using CRDs detected in the combined set. We tested QTL effects for association with the other molecular phenotype (i.e., aCRD-QTLs with gene expression eQTLs with CRD activity) via CRD-gene associations detected across all samples at nominal significance.

**Causal inference**

To quantify gene-CRD pairs identified as significant at FDR 5% across SCZ cases and controls, we used PC analysis-based dimensionality reduction. For each gene-CRD pair, we aggregated gene expression with CRD activity and used the coordinates on PC1 as new pseudo-phenotypes for QTL mapping in a cis window. This effectively gave us eCRDQTL-CRD-gene triplets consisting of a genetic variant, a CRD and a gene, all associated with each other.

We applied a Bayesian Network approach to infer the most likely causal relationship for eCRDQTL-CRD-gene triplets common to SCZ cases and controls and conducted the analyses separately in SCZ cases and in controls. This approach allowed to estimate the most likely network from which the observed data originates by calculating the posterior probabilities for the three possible causal models\(^57\): i) causal model in which the genetic variant affects first the CRD and then the gene, ii) reactive model in which the genetic variant affects the gene and then the CRD, iii) independent model in which the genetic variant affects the gene and the CRD independently.

To provide confidence for retrieved probabilities, we carried out 100 bootstrapping runs for each tested triplet separately for SCZ cases and controls using sampling with replacement. We estimated how many times the most probable model across bootstrapping runs for each triplet was the same as in the original Bayesian Network results and filtered out all triplets that fell below a confidence threshold of 55%: this corresponds to the lower quartile value in SCZ cases.\(^56\)

\(R\)/clusterProfiler package\(^56\) was used for gene set enrichment analysis for genes that belonged to triplets showing directional change from eCRD-QTL onto gene expression/CRD activity between SCZ cases and controls. We considered two scenarios: i) causal model in controls, but reactive/independent in SCZ cases; ii) reactive/independent in controls, but causal in SCZ cases.
REFERENCES


ACKNOWLEDGEMENTS

This research was supported by the Intramural Research Program of the NIMH (NCT00001260, 900142), the Estonian Research Council ETAg grant no PUTJD901 (to M.A.) and the National Centre of Competence in Research (NCCR) grant “Synapsy – The Synaptic Bases of Mental Diseases” (no 51NF40-158776) from the Swiss National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Table 1. Molecular phenotype associations at FDR 5% in SCZ cases and controls. The columns indicate the molecular phenotype and type of association, and the numbers of identified associations and sample set used. *Gene-CRD associations found using CRDs identified in the combined set or CRDs identified only in SCZ cases.

<table>
<thead>
<tr>
<th>Molecular phenotype and type of association</th>
<th>Tested sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENE EXPRESSION</td>
<td></td>
</tr>
<tr>
<td>Differentially expressed genes</td>
<td>1,363 (937 ↑, 426 ↓)</td>
</tr>
<tr>
<td>eQTL</td>
<td>867</td>
</tr>
<tr>
<td>SCZ-specific eQTL</td>
<td>130</td>
</tr>
<tr>
<td>CRDs</td>
<td></td>
</tr>
<tr>
<td>cQTL</td>
<td>3,239</td>
</tr>
<tr>
<td>CRDs</td>
<td>10,938</td>
</tr>
<tr>
<td>Differentially active CRDs</td>
<td>1,141 (542 ↑, 599 ↓)</td>
</tr>
<tr>
<td>aCRD-QTL</td>
<td>796</td>
</tr>
<tr>
<td>SCZ-specific aCRD-QTL</td>
<td>40</td>
</tr>
<tr>
<td>QTL-CRD-gene</td>
<td></td>
</tr>
<tr>
<td>Gene-CRD pairs</td>
<td>95/119*</td>
</tr>
<tr>
<td>eCRDQTL-CRD-gene triplets</td>
<td>38</td>
</tr>
</tbody>
</table>

CC-BY 4.0 International license

It is made available under a CC-BY 4.0 International license.
Fig. 1. Features of SCZ-specific CRDs. (a) Difference in medians per peak activity between SCZ cases and controls as a function of the strength of association given in \(-\log_{10} p\)-values; purple dots denote peaks that are differentially active between SCZ cases and controls at FDR 5% (3,450 peaks) with light purple and dark purple indicating lower and higher median activity, respectively, in SCZ cases compared to controls. (b) Comparison of differences in per CRD peak activity correlation estimates between SCZ cases and controls for SCZ-specific CRDs (3,078 CRDs). (c) Example region of a correlation structure between 250 peaks on chromosome 5 in SCZ cases and controls, revealing a well-delimited SCZ-specific CRD that is composed of 5 regulatory elements.
Fig. 2. Association between CRDs and genes. (a) TSSs of differentially expressed genes (DEG) are localized within differentially active CRDs significantly more often than expected by chance (Fischer’s exact test $p = 8.72 \times 10^{-6}$, odds ratio $1.60$); coloured dots denote DEGs identified at FDR 5%; purple dots mark DEGs, turquoise dots mark DEGs with TSS within differentially active CRD. (b) Distribution of gene-to-CRD distances for genes localizing outside the associated CRD boundary (545 gene-CRD associations). (c) Distribution of the relative position of gene TSS to the boundary of an associated CRD (652 gene-CRD associations).
Fig. 3. Example of a SCZ-specific (a) aCRD-QTL and (b) eQTL effect. Genotype-dependent effect on CRD activity and gene expression identified only in SCZ cases.
Fig. 4. Regulatory mechanisms for eCRDQTL-CRD-gene triplets in SCZ cases and controls.: Comparison of the direction of effect from eCRD-QTL onto gene expression and CRD activity for tested triplets between SCZ cases and controls: (a) triplet count for models showing the same regulatory mechanism in SCZ cases and controls, and (b) triplet count for models showing a change in the regulatory mechanism between SCZ cases and controls (term "different" indicates either reactive or independent model; light blue bar indicates triplets, for which the causal model (i.e., mediation via CRD for QTL effect) was not identified in SCZ cases nor in controls). (c) Genotype-dependent effect for eCRD-QTL 7:5849993:C:G on FSCN1 expression. (d) Distinct regulatory mechanism of genetic regulation on gene expression for SCZ cases and controls for a triplet consisting of an eCRD-QTL 7:5849993:C:G, gene FSCN1 and a CRD composed of 5 REs on chr7:5623132-5705414; the probabilities based on Bayesian Networks for each tested model is given above schematics; shading of the colour for the gene and for the CRD indicates strength in expression and activity, respectively.
Supplementary Information

Leveraging interindividual variability of regulatory activity refines genetic regulation of gene expression in schizophrenia

Maris Alver\textsuperscript{1,2,3,4}, Nikolaos Lykoskoufis\textsuperscript{1,2,3}, Anna Ramisch\textsuperscript{1,2,3}, Halit Ongen\textsuperscript{1,2,3}, Emmanouil T. Dermitzakis\textsuperscript{1,2,3}

\textsuperscript{1} Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland
\textsuperscript{2} Swiss Institute of Bioinformatics, University of Geneva, Geneva, Switzerland
\textsuperscript{3} Institute of Genetics and Genomics in Geneva, University of Geneva, Geneva, Switzerland
\textsuperscript{4} Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia

This PDF file includes:
Materials and Methods
Figures S1 to S20
References
DATA PREPARATION

General overview
Molecular data for the Human Brain Collection Core (HBCC) within the Division of Intramural Research Programs (DIRP) at the National Institute of Mental Health (NIMH) was accessed through dbGaP (study accession phs000979.v3.p2; request #88083-1 approved by NIH on January 31st, 2020; https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000979.v3.p2).

Briefly, post-mortem dorsolateral prefrontal cortex (DLPFC) brain tissues were obtained under the protocols approved by the CNS IRB with the permission of the next-of-kin through the Offices of the Chief Medical Examiners in the District of Columbia, Northern Virginia, and Central Virginia, and from the University of Maryland Brain and Tissue Bank and the Stanley Medical Research Institute. Clinical characterization, neuropathological screening, toxicological analyses, and the dissections of the brain region were performed as described previously1. All patients met DSM-IV criteria for a lifetime Axis I diagnosis of psychiatric disorders including schizophrenia or schizoaffective disorder. Controls had no history of psychiatric diagnoses or addictions. While the HBCC cohort consists of individuals with different ancestral background, we opted for using samples from the two largest self-reported ancestral groups: African American and European.

Phenotype data
At least two levels of molecular data (i.e., whole-genome genotype, RNA-sequencing or ChIP-sequencing data) were available for 272 individuals: 98 SCZ cases and 174 controls, 188 males and 84 females, 164 African Americans and 108 Europeans; controls <16 years of age were excluded (Supplementary Fig. 1).

Genotype data
Subjects were genotyped with three Illumina whole-genome genotyping arrays: HumanHap650Y, Human1M-Duov3 and HumanOmniSM-Quad. Raw .idat files were downloaded from dbGaP and converted to .gtc and .vcf files using Illumina Array Analysis Platform Genotyping Command Line Interface (iaap-cli and gtc2vcf (webpage: https://github.com/freeseek/gtc2vcf), respectively. Array-based quality control was carried out with PLINK v2.02 (webpage: https://www.cog-genomics.org/plink/2.0/) using the following criteria: i) exclusion of individuals with genotype call rate <95%; and exclusion of single nucleotide variants (SNVs) with call rate <95%, Hardy-Weinberg equation (HWE) <10e-5, minor allele frequency (MAF) <0.01, and with ambiguous genotypes (AT and GC SNVs); ii) confirmation of the match for genotype and phenotype sex, removal of outliers who deviated +/- 3 standard deviations from the samples’ heterozygosity rate mean as well as verification that the data did not contain closely related individuals (PI_HAT >0.2).

Prior to imputation, we determined the intersect of SNVs across three genotyping arrays, filtered each array for the uniform content (506,548 SNVs) and merged the data into one matrix. We used the Haplotype Reference Consortium reference panel3 (webpage: http://www.haplotype-consortium.org/) for array imputation with the following parameters: build hg19, reference panel apps@hrc-r1.1, population mixed, phasing eagle. After imputation, we filtered out SNVs with low imputation quality score R2 <0.3 and applied HWE <1e-6 and MAF <0.05 filters by European and African American ancestral group separately. For the final SNV set, we considered the union of filtered SNVs retrieved in both ancestry sets (8,245,179 biallelic SNVs). We tested for an association between imputed SNVs and genotyping array to confirm that SNV data were not biased towards any genotyping array. No bias towards any genotyping array was detected (Supplementary Fig. 2).

We used the 1000 Genome Project data4 as reference to exclude samples that showed differential ancestry background than European or African/Admixed American based on the principal component analysis (PCA) (Supplementary Fig. 3).
RNA-sequencing data

Raw .fastq files (2 x 125bp pair-end reads) were downloaded from dbGaP. RNA-seq mapping was done with STAR 2-pass mapping using GENCODE hg19 reference genome annotations. RNA-seq data quality was assessed using QTLtools bamstat mode (webpage: https://qtltools.github.io/qtltools/) for i) the number of mapped sequencing reads passing mapping quality filter, ii) number of mapped sequencing reads falling within the GENCODE hg19 annotations, and iii) the number of GENCODE hg19 annotations covered by at least one sequencing read.

Gene expression was quantified by counting the number of RNA-seq reads that were in correct orientation, that had a mapping quality score above 225 and that did not contain more than 16 mismatches in both ends of the fragment with the reference genome. Each gene in each sample was normalized to get an RPKM value using the formula RPKM = (read_count x 1e9) / (total_mapped_reads x gene_length). We filtered for protein-coding and lincRNA genes and considered the union quantifications detected in >=50% in SCZ cases and in >=50% in controls. This yielded 18,258 protein-coding genes and 3,730 lincRNAs (total of 21,988 genes) for 243 individuals.

ChIP-sequencing data

Raw .fastq files (2 x 75bp pair-end reads; profiled for histone mark H3K27ac) were downloaded from dbGaP. Reads were aligned to reference genome hg19 with BWA-MEM v0.7.16 using an alignment score threshold of 10. For quality control, all ChIP-seq experiments were processed through Phantompeakqualtools v1.16 to generate two quality metrics: normalized strand cross-correlation (NSC) and relative strand cross-correlation (RSC) (Supplementary Fig. 4). These metrics use the cross-correlation of stranded read density profiles to measure enrichment independently of peak calling. RSC <0.8 and NSC <1.05 indicate low signal to noise ratio. The quality tag (based on thresholded RSC) was >2 (very high) for all experiments.

Peak calling and quantification were carried out with HOMER v14.11.1 (webpage: http://homer.ucsd.edu/homer/). Briefly, we first determined ChIP-seq peak coordinates across SCZ cases and controls to get a population scale call set of ChIP-seq peaks and then quantified the peaks for each individual according to the identified peak coordinates. Specifically, we first built a population call set of ChIP-seq peaks by aggregating 1e6 ChIP-seq reads from 74 SCZ cases and 74 controls together in a unique BAM file. Next, we carried out the actual peak calling onto the derived consensus peak set (derived unique BAM file) for all individuals (n=193: 74 SCZ cases and 119 controls) using HOMER findPeaks mode and parameters -style histone -o auto. This yielded 141,219 ChIP-seq peaks for each individual (mean peak length 2137 base pairs (bp)). Next, we quantified the peaks by obtaining per-peak read counts per sample using the peak coordinates from the consensus peak set and using the HOMER script annotatePeaks.pl with the following options: -noann -nogene -size given. This script counted the number of ChIP-seq reads falling within the peak coordinates. Read counts were subsequently normalized for a total of 10 million mapped reads per sample.

Mislabelling detection

Given that many analyses relied on testing genotype versus sequence data, we looked at the concordance between both to ensure there was no sample mislabeling and all individuals had both sequence and genotype data available. We used QTLtools mbv mode and .vcf and .bam file for each individual and assessed the concordance at heterozygous and homozygous genotypes between genotype and RNA-sequencing and between genotype and ChIP-seq sequencing data. No such errors were detected.
COMPUTATIONAL AND STATISTICAL ANALYSES

Cis-regulatory domain (CRD) calling

For CRD calling, we used the pipeline developed in Delaneau et al. 2019 \(^1\) (webpage: https://github.com/odelaneau/clomics). We started with building a correlation map by systematically measuring interindividual correlation between all possible pairs of ChIP-seq peak quantifications (i.e., retrieved Pearson correlation coefficients using corrected and rank-normal transformed data matrix (see final paragraph of this section for specifics). Next, we applied an agglomerative hierarchical clustering of the data on a per chromosome basis. Specifically, we started with each ChIP-seq peak being assigned to its own cluster and iteratively merged clusters as we moved up in the hierarchy. To determine the pair of clusters to be merged, we maintained throughout the procedure a correlation matrix that corresponded to the matrix of squared correlations between ChIP-seq peaks, searched the pair of clusters that exhibited the highest value as well as constantly updated the correlation matrix when merging clusters together. For minimizing computational cost, we searched for the maximal correlation between clusters that were not separated by more than 250 rows or columns (i.e., peaks or groups of peaks in the correlation matrix), allowing to store and update the diagonal parts of the correlation matrix and merging together proximal features of the genome. This strategy resulted in a binary tree that regrouped all ChIP-seq peaks from the same chromosome in which each node delimited a set of highly correlated ChIP-seq peaks.

CRDs were called by identifying the minimal set of internal nodes that captured most of the overall correlation mass (i.e., cumulative sum of squared correlation). To retain an internal node as a CRD, three criteria needed to be fulfilled: i) CRDs regrouped only highly correlated ChIP-seq peaks: the mean absolute correlation between all possible pairs of ChIP-seq peaks within a CRD had to be at least twice as high as the mean correlation between all ChIP-seq peaks in the chromosome; ii) CRDs had well-defined boundaries: the mean absolute correlation between all pairs of ChIP-seq peaks involving either the first or the last ChIP-seq peak (on the basis of their genomic location) had to be at least twice as high as the same value derived for the first and last peaks on the chromosome; iii) CRDs captured distal coordination between at least two overlapping regulatory regions: ChIP-seq peaks had to cover at least two non-overlapping regulatory regions. These three criteria were implemented into an algorithm that processed each binary tree starting from the root node (node regrouping all peaks of a chromosome) and recursively traversed the internal nodes of the tree until an internal node fulfilled all three criteria. Then, declared the internal node and all the peaks downstream as a CRD, stopped to go deeper by ignoring the children of this node and carried on with other internal nodes in the tree.

This pipeline was applied using only SCZ ChIP-seq quantifications (n=74), only control ChIP-seq quantifications (n=119) and using ChIP-seq quantifications in the combined set (i.e., SCZ cases and controls together; n=193), such that we started with three separate correlation maps and in the end had called SCZ-specific CRDs, controls-specific CRDs as well as uniform CRDs across all samples. For correcting ChIP-seq quantifications, we identified the optimal number of PCs that captured variability in ChIP-seq data in SCZ cases, in controls and across samples via QTL mapping (see section QTL mapping for molecular phenotypes) by considering variable number of PCs as covariates. After finding the optimal configuration (giving the best QTL discovery power; Supplementary Fig. 5a-c), we corrected ChIP-seq peak quantifications for 3 genotype PCs and 10 ChIP-seq PCs in SCZ cases (n=74), 3 genotype PCs and 20 ChIP-seq PCs in controls (n=119), and 3 genotype PCs and 30 ChIP-seq PCs across samples (n=193). All three resulting data matrices were rank-normal transformed separately.

CRD activity quantification

For CRD activity quantification, we applied a dimensionality reduction approach, i.e., we enumerated all ChIP-seq peaks per CRD, and took the mean of all single peak quantifications per individual to retrieve a single quantification value for each individual. We used the ChIP-seq peak correlation map retrieved in the combined set (i.e., across SCZ cases and controls; n=11,374 CRDs). For SCZ-specific CRD structure analyses, we retrieved CRD activity quantifications for SCZ cases only using the CRDs identified in SCZ cases (n=10,938 CRDs). The resulting vectors were again rank-normalized such that these matched a
normal distribution with mean 0 and standard deviation 1 and consisted of one row per CRD and one column per individual.

Quantitative Trait Loci (QTL) mapping for molecular phenotypes (ChIP-seq peak activity, gene expression, CRD activity)

For each molecular phenotype, we first enumerated all genetic variants within +/- 1 Mb and then tested each one of these variants for association with the phenotype and only retained the best hit (i.e., with the smallest nominal p-value). Secondly, we adjusted the best nominal p-value for the number of variants being tested by permutations. Specifically, we randomly shuffled the phenotype quantifications 1,000 times and retained the best association p-values for each permuted data set, which effectively gave 1,000 null p-values of associations. Third, to correct for the number of molecular phenotypes being tested whole genome (e.g., number of genes, peaks, CRDs), we used a false discovery rate (FDR) correction approach and declared phenotype-variant pairs at FDR 5% threshold as significant. These steps were carried out with QTLtools cis mode.

To discover multiple QTLs with independent effects on a given molecular phenotype, we used the conditional analysis approach implemented in QTLtools. Briefly, this approach is based on a forward-backward scan of the cis-window around the phenotypes to automatically learn the number of independent QTLs and to identify the most likely candidate variants, while controlling for a given FDR.

For SCZ-specific QTL discovery, we considered QTL effects identified in SCZ cases (796 aCRD-QTLs and 867 eQTLs) and for each variant-phenotype pair ran a linear regression including genotype, disease status (SCZ/CTL), and covariates, and tested for significance of a genotype * disease status interaction on molecular phenotype (gene expression or CRD activity). This was followed with FDR 5% correction for the number of QTLs tested.

CRD structure analysis

For determining CRD sharing between SCZ cases and controls, we compared ChIP-seq peak correlation maps between SCZ cases and controls and called a CRD shared if ≥50% of the peaks overlapped between the reference and the query correlation map.

To assess the features of SCZ-specific CRDs, we considered only CRDs in SCZ cases composed of peaks not regrouping into any CRD in controls. These formed 28% of the CRDs identified in SCZ cases (3,078 CRDs composed of 6,650 peaks). For underlying peak activity estimation, we used ChIP-seq peak quantifications normalized for 10 million reads per sample, uncorrected for any covariates and applied a Mann-Whitney U test per peak activity between SCZ cases and controls. Significant differences between SCZ cases and controls were determined at FDR 5% using R/qvalue package.

For confirming whether the peaks within SCZ-specific CRDs showed different correlation structures in SCZ cases vs controls and were not driven by the mean background correlation estimate ascertained separately in SCZ cases and controls in CRD calling, we used ChIP-seq peak quantifications corrected for biological and technical covariates (3 genotype PCs and 10 ChIP-seq PCs in SCZ cases, and 3 genotype PCs and 20 ChIP-seq PCs in controls as outlined in CRD calling section). The corrected data matrices were rank-normal transformed separately. We calculated the mean Pearson correlation estimate between peak activities per CRD separately in SCZ cases and controls (i.e., in controls measured the correlation estimate between peaks per SCZ-specific CRDs) and used Mann-Whitney U test for comparing identified mean correlation estimates between SCZ cases and controls.

CRD and gene association

Briefly, we considered normalized CRD activity quantifications (final step in CRD activity quantification section) and corrected and normalized gene expression quantifications and used QTLtools cis permutation pass to identify CRDs associated with a gene in a +/-1 MB window from a gene’s transcription start site. We performed these analyses to i) identify genes associated with SCZ-specific CRDs, and ii) capture comparable associations for SCZ cases and controls using the same CRD annotations, i.e., CRDs identified in the combined set. The first approach was performed in SCZ cases only (n=59) using CRD activity quantifications retrieved based on the CRDs identified in SCZ cases (10,938 CRDs; Supplementary
Table 1). The second approach was done for SCZ cases (n=59), for controls (n=105) and across samples (n=164; disease status (SCZ/CTL) considered as a covariate) using CRD activity quantifications retrieved based on the CRDs identified in the combined set (Supplementary Table 5).

Specifically, to capture technical and biological variability in gene expression data, we residualized for ancestry, using 3 genotype PCs, and for the number of optimal RNA-seq PCs that allowed to discover the maximum number of eQTLs. This was done similarly as for the ChiP-seq data by doing association testing at variable number of PCs. Gene expression quantifications were corrected for 3 genotype PCs and 10 RNA-seq PCs in SCZ cases, 3 genotype PCs and 30 RNA-seq PCs in controls, and 3 genotype PCs and 40 RNA-seq PCs in the combined set (Supplementary Fig. 5d-f). The resulting matrices were rank-normal transformed. Next, we enumerated all CRDs within +/-1 Mb of gene’s transcription start site, tested their activity for association with gene expression and stored the best hit together with the nominal p-value. We adjusted the nominal p-value for the number of CRDs being tested in cis using permutation and corrected for the number of genes being tested using the \textit{R/qvalue} package\textsuperscript{12}. We determined gene-CRD associations at FDR 5% as significant.

Differential CRD activity and differential gene expression analysis

Both differential CRD activity and differential gene expression analyses were carried out using DESeq2\textsuperscript{13}. Significant associations were determined at FDR 5% (Supplementary Table 2, Supplementary Table 3). For differential CRD activity analysis, we used unnormalized ChiP-seq peak read counts obtained with HOMER (annotatePeaks.pl with options -noann -nogene -size given -raw)\textsuperscript{9} and summed these up per CRD using the ChiP-seq peak correlation map identified in the combined set (11,374 CRDs). For differential gene expression analysis, we used RNA-seq read counts.

To identify covariates for correction, we carried out association testing i) between all available biological and technical covariates and diagnosis status (Mann-Whitney U test), and ii) between all available biological and technical covariates and individual ChiP-seq peak activity and gene expression quantifications (linear regression) and calculated \(\pi_1\) estimate\textsuperscript{12} to identify the proportion of true associations. We identified the following covariates for differential CRD activity analysis: sex, age at death, genotype PC1, library batch, GC content in sequencing data, empirical insert size, 15bp repeat in sequencing data. We identified the following covariates for differential gene expression analysis: sex, age at death, genotype PC1, genotype PC2, post-mortem interval, brain pH, brain weight, RNA integrity number, total RNA yield, A260/A280 ratio, GC content in sequencing data, transcript integrity number, empirical insert size, 15bp repeat in sequencing data, date of sequencing.

\textit{R/clusterProfiler} package\textsuperscript{14} was used for gene set enrichment analysis. We considered genes that were either i) significantly differentially expressed (regardless of direction of effect), ii) significantly down-regulated or iii) significantly up-regulated. Significant associations were determined at FDR 5% (Supplementary Table 4).

Association of aCRD-QTLs and eQTLs with the other molecular phenotype

We tested QTL effects for association with the other molecular phenotype (i.e., gene expression with aCRD-QTLs and CRD activity with eQTLs) via CRD-gene nominal associations using CRD identified in the combined set). Specifically, for each gene we identified in cis window all associated CRDs at nominal pass, and for each CRD determined in cis window all associated genes at nominal pass. Using these intermediate associations, we could look whether e.g., an eQTL also affects the CRD that the targeted eGene is associated with, and vice versa, whether the aCRD-QTL affects the gene that the impacted aCRD is associated with. Proportion of sharing was estimated using \(\pi_1\) estimate\textsuperscript{12}.

Colocalization with SCZ GWAS variants

Briefly, we assessed the likelihood of a shared functional effect between SCZ risk variants from four GWAS studies\textsuperscript{15-18} and SCZ-identified QTLs (796 aCRD-QTLs and 876 eQTLs) using regulatory trait concordance (RTC). Specifically, we considered independent hits from four GWAS studies and applied the RTC algorithm. This algorithm assesses the likelihood of a shared functional effect between a GWAS variant and a QTL variant by quantifying the change in the statistical significance of the QTL after correcting the
QTL phenotype (gene expression or CRD activity) for the genetic effect of the GWAS variant and comparing its correction impact to that of all other SNPs in the interval\(^{19,20}\). We applied a cut-off of RTC ≥ 0.9 for determining a shared functional effect. The output files indicate the union results across four GWAS studies. When accounting for LD, we see 16 independent shared effects between GWAS and eQTL variants and 12 independent shared effects between GWAS and aCRD-QTL variants (Supplementary Table 9).

**CRD-gene pair quantification**

To quantify the 1,197 gene-CRD pairs we identified as significant at FDR 5% across SCZ cases and controls (n=164), we used PCA-based dimensionality reduction. For each gene-CRD pair, we aggregated gene expression with CRD activity and used the coordinates on PC1 as new pseudo-phenotypes. For identifying genetic variants that affect both the CRD activity and the gene expression (per gene-CRD pair), we used the new derived pseudo-phenotypes and carried out an eCRD-QTL (genetic variant that affects both the CRD and gene) discovery analysis in cis across all samples (n=164) using permutation.

**Causal inference for determining causal relationships for eQTL-CRD-CRD-gene triplets**

We applied a Bayesian Network approach to infer the most likely causal relationship for eCRDQTL-CRD-gene triplets common to SCZ cases and controls (1,100 triplets) and conducted the analyses separately in SCZ cases (n=59) and in controls (n=105). This approach allowed to estimate the most likely network from which the observed data originates. The starting point is always the genetic variant as this does not change (genome is fixed). We explored three distinct models (topologies): i) causal model in which the genetic variant affects first the CRD and then the gene, ii) reactive model in which the genetic variant affects the gene and then the CRD, iii) independent model in which the genetic variant affects the gene and the CRD independently (Supplementary Fig. 17). For each triplet we built a 59 x 3 and 105 x 3 data matrix for SCZ cases and controls, respectively, that contained normalized quantifications, and calculated the likelihood of three possible Bayesian Network topologies using R/bnlearn package\(^{21}\). We converted the likelihoods to posterior probabilities, assuming a uniform prior probability for three possible models.

**Bootstrapping**

To estimate the accuracy for the Bayesian Network results and provide confidence for retrieved probabilities, we used bootstrapping. We carried out 100 bootstrapping runs for each tested triplet separately for SCZ cases (n=59) and controls (n=105) using sampling with replacement. For accuracy estimation, we calculated how many times the most probable model across bootstrapping runs for each triplet was the same as in the original Bayesian Network results. We filtered out all triplets that fell below a confidence threshold of 55%: this corresponds to the lower quartile value in SCZ cases (Supplementary Fig. 19ab).

**Gene set enrichment for model-change associated triplets**

Used R/clusterProfiler package\(^{14}\) for gene set enrichment analysis for genes that belonged to triplets showing directional change from eCRD-QTL onto gene expression/CRD activity between SCZ cases and controls. We considered two scenarios: i) causal model in controls, but reactive/independent in SCZ cases (n=147); ii) reactive/independent in controls, but causal in SCZ cases (n=87).
Supplementary Fig. 1. Overview of the phenotypic characteristics of the HBCC cohort. Mean age 51 years (sd = 14.4) for schizophrenia (SCZ) cases and 42.2 years (sd = 16.4) for controls.
Supplementary Fig. 2. Association testing between imputed SNVs and genotyping array. For each genotyping array, an association test was carried out using logistic regression for array of interest (either (a) Illumina Human1M-Duov3, (b) Illumina HumanOmni5M-Quad or (c) Illumina HumanHap650Y) vs other two arrays. Manhattan and QQ-plots depict the distribution of imputed SNVs per chromosome as a function of -log10 p-values and expected vs observed p-values, respectively. Post-imputation quality control (MAF <0.05 and HWE <1e-6) was applied by ancestry and the union SNV content across ancestry sets was considered for this and all downstream analyses. No bias towards any genotyping arrays was detected.
Supplementary Fig. 3. Principal component (PC) analysis of genotype data. Coordinates of (a) PC1 vs PC2, (b) PC1 vs PC3 and (c) PC2 vs PC3 in reference to the 1000 Genome Project samples. HBCC cohort samples are coloured in pink and cluster at PC coordinates represented by the African and European 1000 Genome Project super populations.
Supplementary Fig. 4. PhantomPeakQualTools quality metrics. (a) Normalized Strand Cross-Correlation coefficient (NSC) distribution and (b) Relative Strand Cross-correlation coefficient (RSC) distribution for H3K27ac peaks in HBCC samples. NSC <1.05 and RSC <0.8 indicate low signal to noise.
Supplementary Fig. 5. Optimization of QTL discovery. Number of chromatin-QTLs (cQTLs) and expression-QTLs (eQTLs) discovered (a,d) in SCZ cases, (b,e) in controls (CTL) and (c,f) in the combined set as a function of the number of genotype (GT) principal components (PCs) and ChIP-seq and RNA-seq PCs used to residualize ChIP-seq peak quantification and gene expression quantification data, respectively. The PCs that allowed the discovery of the maximum number of QTLs were retrained for downstream analyses.
Supplementary Fig. 6. Regulatory element (RE) content of CRDs. Number of regulatory elements per CRD (a) in SCZ cases (n=74), (b) in controls (CTL, n=119), and (c) in the combined set (n=193). Mean number of REs per CRD was 3.7, 4.3 and 4.7 in SCZ cases, in controls and in the combined set, respectively. Mean CRD length was 137,017 bp, 135,734 bp and 138,144 base pairs in SCZ cases, in controls and in the combined set, respectively.
Supplementary Fig. 7. Fraction of CRD peak content sharing between SCZ cases and controls (CTL). A CRD was deemed shared between SCZ cases and controls in case $\geq 50\%$ of ChIP-seq peaks of the reference CRD were present in a CRD from the query state. Forty-two percent of CRDs detected in controls (n=119) were also detected in SCZ cases (n=74) and vice versa, 34% of CRDs detected in SCZ cases were also detected in controls.
Supplementary Fig. 8. Differential peak activity between SCZ cases and controls (CTL). Only peaks clustering into SCZ-specific CRDs (i.e., CRDs composed of peaks not part of any CRD in controls) were considered. (a) P-value distribution of peaks differentially active between SCZ cases and controls. (b) Proportion of differentially active peaks (3,540 peaks) between SCZ cases and controls per SCZ-specific CRDs at FDR 5%. One-third of SCZ-specific CRDs (1,056 CRDs and 2,242 peaks) were forming due to all underlying peaks showing significantly different peak activity in SCZ cases compared to controls.
Supplementary Fig. 9. Volcano plots outlining (a) differential CRD activity and (b) differential gene expression between SCZ cases and controls. Differentially active CRDs (n=1,141; 599 with lower activity and 542 with higher activity) and differentially expressed genes (n=1,363; 937 up-regulated and 426 down-regulated), respectively, at FDR 5% are highlighted in purple.
Supplementary Fig. 10. Replication of differential gene expression analysis results for SCZ in previously published findings. Differentially expressed genes identified at FDR 5% were in concordance with findings published by (a) Jaffe et al., 2018 and by (b) Fromer et al., 2016 based on \( \pi_1 \) estimate.
Supplementary Fig. 11. Gene-CRD associations. P-value distribution for gene-CRD associations (a) in SCZ cases (n=59), (b) in controls (n=105) and (c) in the combined set (n=164). At FDR 5%, 95, 634 and 1,197 CRD-gene associations were identified in SCZ cases, in controls and in the combined set, respectively. Disease status (SCZ/CTL) was considered as a covariate for CRD-gene pair detection across in the combined set.
Supplementary Fig. 12. Gene-CRD associations. Number of genes and CRDs as a function of the number of CRDs and genes they were associated with, respectively, identified in the combined set (n=1,197 gene-CRD associations).
Supplementary Fig. 13. QTL discovery. P-value distribution for QTL and CRD activity/gene expression associations (a,c) in SCZ cases and (b,d) in controls. At 5% FDR and in cis, 796 and 2,929 functionally independent aCRD-QTLs, and 867 and 6,166 functionally independent eQTLs in SCZ cases and controls, respectively, were discovered.
**Supplementary Fig. 14. CRD/gene distance in base pairs from associated QTL.** Genomic distance between genetic variant and CRD/gene as a function of the strength of association given in -log10 p-values (a,c) for SCZ cases and (b,d) for controls.
Supplementary Fig. 15. Proportion of sharing aCRD-QTL and eQTL effects between SCZ cases and controls based on π1 estimate. P-value distributions of SCZ-identified (a) aCRD-QTL and (c) eQTL effects tested in controls, and control-identified (b) aCRD-QTL and (d) eQTL effects tested in SCZ cases.
Supplementary Fig. 16. Proportion of sharing aCRD-QTL and eQTL effects for gene expression and for CRD activity, respectively, based on π1 estimate. P-value distribution of (a) SCZ-identified and (b) controls-identified aCRD-QTLs tested for gene expression over CRD-gene associations identified at nominal significance level. P-value distribution of (c) SCZ-identified and (d) controls-identified eQTLs tested for CRD activity over gene-CRD associations identified at nominal significance level.
Supplementary Fig. 17. Bayesian Network approach. Schematic of models considered in Bayesian Networks to infer the most likely causal relationship for eCRDQTL-CRD-gene triplets in SCZ cases and controls. eCRD-QTL denotes a genetic variant that affects the activity of a CRD and the expression of a gene that show a significant association with each other.
Supplementary Fig. 18. Causal inference estimation for eCRDQTL-CRD-gene triplets. Counts of the most probable model for each triplet (n=1,100) for (a) SCZ cases and (b) controls. Distribution of the probabilities for the most probable model for each triplet for (c) SCZ cases and (d) controls. Mean probabilities for each model as a function of the distance in base pairs between eCRD-QTL and the target gene for (e) SCZ cases and (f) controls.
Supplementary Fig. 19. Regulatory mechanism for eCRDQTL-CRD-gene triplets in SCZ cases and controls. Distribution of accuracy estimation (%) denoting how often the most probable model across bootstrapping runs for each triplet was the same as in the original Bayesian Network analysis for SCZ cases and controls (a) across models and (b) by model. (c) Triplet counts as a function of accuracy estimation; purple colour denotes triplet counts for SCZ cases, grey for controls and red indicates triplet counts at the intersect of accuracy estimation for SCZ cases and controls. (d) Comparison of the direction of effect from eQTL-CRD onto molecular phenotypes between SCZ cases and controls for 748 triplets that surpassed the accuracy estimation of 55%.
Supplementary Fig. 20. Example of mechanistic change in the regulation of gene expression between SCZ cases and controls for a triplet consisting of an eCRD-QTL 7:584993:C:G, gene FSCN1 and a CRD composed of 5 REs on chr7:5623132-5705414. Distribution of (a) FSCN1 gene expression and (b) CRD activity for SCZ cases and controls. (c) Genotype-dependent effect for eCRD-QTL 7:584993:C:G on CRD activity. Genotype-dependent effect for eCRD-QTL 7:584993:C:G on FSCN1 expression by ancestry group (d) for SCZ cases and (e) for controls.
REFERENCES