Title
Deep Characterization of the Contribution of Short Tandem Repeats Across Tissues

Author
Fotsing, Stephanie Feupe

Publication Date
2018

UNIVERSITY OF CALIFORNIA SAN DIEGO

Deep Characterization of the Contribution of Short Tandem Repeats Across Tissues

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

in

Bioinformatics and System Biology with a Specialization in Biomedical Informatics

by

Stephanie Feupe Fotsing

Committee in charge:

Professor Melissa Gymrek, Chair
Professor Vineet Bafna, Co-Chair
Professor Vikas Bansal
Professor Hannah Carter
Professor Lucila Ohno-Machado

2018

This Dissertation of Stephanie Feupe Fotsing is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____
Co-chair

____
Chair

University of California San Diego
2018

DEDICATION

I dedicate this thesis to my family. My success is mostly yours.

To my husband Joseph Fotsing for your love, patience and unwavering support throughout this journey. You've been my rock throughout and held my hand when I couldn't stand. I remain forever grateful and couldn't have asked for a better partner 13 years ago.

To my children Leandra and Niels-Gabriel for lightening my days. We've shared this entire journey and throughout; your smiles and complete dependence have made this entire experience most memorable. You are my first pride. Love always.

A special dedication in memory of my beautiful mother Julienne Feupe to whom I owe every bit of my existence and the person I have become. You were my cheerleader, confidant, best friend and muse, but did not live long enough to see this work to its end. Your caring push for tenacity still rings in my ears. I can hear you sing the songs of joy at this moment.

To my father Martin Feupe for your unconditional love. Your words of encouragements and constant reminder of your pride still fill my ears. With your always happy personality, you've shown me over and over what courage and perseverance look like.

To my brothers, sisters and friends for your love and encouragements.

With all things, God blessed me abundantly!

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ACKNOWLEDGEMENTS

My life's circumstances and values have dictated which one of my priorities including this dissertation became number one at a given moment. In any case, my completion of this dissertation is a result of me juggling those priorities and by God's grace, not alone. Throughout this journey, I have encountered many mentors who made this journey a fruitful life teaching experience, where my PhD comes as a single unit in a big package. This diverse set of experienced researchers, mentors and friends have offered me a large landscape of skills sets and different styles of teaching, mentoring and research that I can leverage at different times in my future.

I would like to first thank Dr. Melissa Gymrek for all the support, the resources and guidance she provided to the completion of this work, from theme generation to paper writing, revision and submission. She has the teaching version of Short tandem repeats and human genetics made easy. Her mentoring, at time hands-on or hands-off depending on the circumstances, helped ease my way into Short tandem repeats and ultimately to the material here presented.

I would like to thank Dr. Vineet Bafna for the support, guidance and mentoring at all times, even when none of my goals coincided with his line of research. He helped me ease my way into human genetics with various exploratory projects and collaborations. As advisor and mentor, I couldn't ask for a better one. My admiration grew in folds as I worked with him.

I am forever grateful for Dr. Lucila Ohno-Machado for the encouragements, unwavering support throughout my doctorate program. Thank you for helping navigate this section of my life with confidence. Her mentorship and push have been most valuable and always timely. She believed in me when I needed it the most.

I thank my thesis committee Dr. Melissa Gymrek, Dr. Vineet Bafna, Dr. Hannah Carter, Dr. Lucila Ohno-Machado and Dr. Vikas Bansal for your advice, recommendation and reviews.

I want to extend my thanks to the whole Department of Biomedical Informatics (DBMI) group, the Bafna lab members, and the Gymrek lab members for the meaningful input, discussion and insights and most of all just being awesome people. Sorry for the many missed social events. Thank you for being amazing colleagues. Heartfelt gratitude to Dr. Hyeon-Eui Kim, Dr. Jiang Xiaoqian for their insights, guidance and great discussion. Thank you to Dr. Roy Ronen for introducing me to genetic analysis, Dr. Jing Zhang, for being there and forming the mom club of 2 in the program. I can't forget the playdates with the kids while we compared notes on how much this is hard, we would always end with "it is going to be fine."

Finally, I extend a heartfelt gratitude to my friends and family here in the US and back home in Cameroon for providing me of a great foundation for success and for being my staunch of supporters and fans.

Chapter 2 contains parts of material from Shubham Saini, Ileena Mitra, Nima Mousavi, Stephanie Feupe Fotsing, Melissa Gymrek - A Reference Haplotype Panel for Genome-wide Imputation of Short-Tandem Repeats. The dissertation author was responsible for this part of the analysis, but not the primary author of the paper.

Chapter 3 is in full adapted from material in submission for publication as it may appear. Stephanie Feupe Fotsing, Alon Goren, Melissa Gymrek "Multi-tissue analysis reveals short tandem repeats as ubiquitous regulators of gene expression and complex traits". The dissertation author was the primary author of this paper.

VITAE

2004 Professional B.Sc Computer Networking, African Institute of Computer Science

2011 MSc. in Bioinformatics and Medical informatics, San Diego State University

2018 Ph.D. Bioinformatics and Systems Biology with Specialization in Biomedical Informatics, University of California San Diego

PUBLICATIONS

Fotsing, Feupe Stephanie, Alon Goren, Gymrek Melissa "Multi-tissue analysis reveals short tandem repeats as ubiquitous regulators of gene expression and complex traits" Manuscript in review

Saini Shubham, Mitra Ileena, Mousavi Nima, **Fotsing Feupe Stephanie**, Gymrek Melissa "A Reference Haplotype Panel for Genome-wide Imputation of Short-Tandem Repeats" Nature Cammunications, Sept 2018

Shenvi Edna C., **Feudjio Feupe Stephanie**, Yang Hai, El-Kareh Robert "Closing the loop": a mixed-methods study about resident learning from outcome feedback after patient handoffs, Diagnosis (Berl). 2018 Sep 21; PMID:30240357; doi: 10.1515/dx-2018-0013.

Kuo Tsung-Ting, Huh Jina, Kim Jihoon, El-Kareh Robert, Singh Siddharth, **Feudjio Feupe Stephanie**, Kuri Vincent, Lin Gordon, Day Michele E., Ohno-Machado Lucila, Hsu Chun-Nan "The Impact of Automatic Pre-annotation in Clinical Note Data Element Extraction - the CLEAN Tool" eprint arXiv:1808.03806, August 2018

Jiang Xiaoqian, Wang Shuang, **Stephanie F. Feupe**, Ohno-Machado Lucila, Privacy-protecting analysis of distributed big data: A practical solution for sharing patient data while maintaining privacy protections. Biomedical Computation review, Spring 2016 Volume 12, Issue1 ISSN 1557-3192

Kim H, Bell E, Kim J, Sitapati A, Ramsdell J, Farcas C, Friedman D, **F. Feupe S**, Ohno-Machado L. "iCONCUR: informed consent for clinical data and bio-sample use for research" J. Am Med Inform Assoc. 2017 Mar; 24(2): 380–387 - doi: 10.1093/jamia/ocw115

Doan S, Lin KW, Conway M, Ohno-Machado L, Hsieh A, **F. Feupe S**, Garland A, Ross MK, Jiang X, Farzaneh S, Walker R, Alipanah N, Zhang J, Xu H, Kim HE. "PhenDisco: phenotype discovery system for the database of genotypes and phenotypes. J Am Med Inform Assoc. 2014 Jan- Feb; 21 (1): 31-6.

F. Feupe, Stephanie, Frias P. F., Mednick SC, McDevitt EA, Heitzman ND. Nocturnal continuous glucose and sleep stage data in adults with type 1 diabetes in real-world conditions. J Diabetes Sci Technol. 2013 September 1; 7(5): 1337-45.

FIELDS OF STUDY

Major fields: Biomedical informatics

Studies in Human Genomics
Professors Vineet Bafna and Melissa Gymrek

Studies in Biomedical informatics
Professor Lucila Ohno-Machado

ABSTRACT OF THE DISSERTATION

Deep Characterization of the Contribution of Short Tandem Repeats Across Tissues

By

Stephanie Feupe Fotsing

Doctor of Philosophy in Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics

University of California San Diego, 2018

Professor Melissa Gymrek, Chair
Professor Vineet Bafna, Co-Chair

High-Throughput Sequencing (HTS) and Genome-Wide Association Studies (GWAS) studies have given us unprecedented insights into the influence of Single Nucleotide Variants (SNV) and Copy Number Variants (CNV) on different phenotypes including gene expression, diseases, and complex traits. However, how other complex genetic variations such as Short Tandem Repeats (STRs) in the genome may affect gene expression remains largely unknown. Identifying and genotyping these types of variants from short DNA sequencing reads or low coverage data present difficult bioinformatics challenges. Additionally, traditional association tests must be modified to handle highly multi-allelic loci such as STRs. Several studies have examined the effect of STRs on gene expression genome-wide. However, these studies were restricted to a single cell type such as whole blood or lymphoblastoid cell lines (LCLs) and had limited power to detect associations due to low quality genotypes. Thus, the results of these studies have had limited biological insights and interpretation in different contexts.

In this dissertation, we address the importance of incorporating STRs in causal screening and large-scale medical genetics studies. We perform the first and largest yet characterization of STRs that contribute to gene expression variation across multiple tissues. To assure robust and reliable outcomes and insights, we leverage data from the GTEx project, which has collected high coverage whole genome sequencing data and RNA-sequencing across dozens of tissues, for more than 600 individuals. Our work confirms a clear contribution of STRs to gene expression regulation, with 25,554 eSTRs identified across 17 tissues. Of these, 14% are identified as high confidence causal variants after fine-mapping against nearby SNPs. eSTRs are highly enriched at predicted promoter and enhancer regions and for motifs with high GC-content. We identified a subset of eSTRs capable of forming G-quadruplexes (G4), a highly stable DNA secondary structure known to be involved in gene regulation. We show that long G4-forming STRs tend to increase expression of nearby genes, potentially by lowering the free energy of promoter regions and promoting RNA polymerase II stalling. Finally, we identify high-confidence eSTRs that likely underlie previously identified genetic associations with complex phenotypes including schizophrenia and blood-related traits.

CHAPTER 1: INTRODUCTION

1.1 Problem statement

In human genetics, there is a great interest in understanding how genetic variation affects phenotypes and traits. Advances in research and technology have improved tremendously in the past decades and thus improved the quality and types of research questions asked. This has helped uncover causal variant(s) to disease phenotypes and understand genome function, i.e. the molecular implication of given types of variants. Many large-scale studies have helped answer some of these questions, including Genome Wide Association Studies (GWAS).

The Human genome project [1,2] started in 1990 and, as the first large-scale sequencing project, was published a decade later, in the dawn of the years 2000. Its publication generated new waves of biological questions for understanding how our DNA works and how genes and their functions are regulated. In 2005, the first GWAS study screening for SNPs associated with age-related macular degeneration [3] was published. Since then, hundreds of GWAS studies have been conducted at larger scales, continually adding to the catalog of association of genetic variants and genes to specific traits. Interestingly, most GWAS hits are located in non-coding regions [4,5], which leads to the hypothesis that they play a role in gene regulation [5] by either transcriptional regulation, noncoding RNA function, and/or epigenetic regulation. Many bioinformatics methods have helped elucidate the likely mechanisms affected by these GWAS hits in noncoding regions. These in vitro or in vivo experiments targeting the suspected mechanism include mice testing in T2D [6], fine mapping methods like eQTL analysis [7], prioritization methods and more [8].

expression Quantitative Traits Loci (eQTLs) analysis is one such method used to assess the role of non-coding variants in gene expression. eQTL analysis consists of the identification variants that explain a fraction of a given gene expression, using various methods including association tests and regression analysis. Briefly, variability in gene expression can explain differences in population

[9–11], difference in phenotypes like diseases: (eg. Celiac disease, Crohn's disease, asthma,etc.) [12–14]; complex traits like skin color [15] and others. While GWAS studies are useful for identifying variants or regions associated with traits. eQTL analysis helps identify the underlying gene(s) affected by the said variants or regions, hence explaining the underlying relation between the variants (or regions) and the trait or susceptibility to the trait. These analyses have led to better interpretation of GWAS studies and the identification of new functional loci. However, there are many diseases for which the underlying causes are still to be elucidated, and traits to be explained.

One critical limitation of these large-scale studies has been their exclusive focus on single nucleotide variations. The human genome harbors multiple types of variants. Short Tandem Repeats (STRs) for example, are repetitive short segments of DNA with unit length of 1 up to 10 bps depending on the study that defines it [16–18]. They have been linked to multiple phenotypes and diseases. Large scale studies on these variants would be extremely informative.

In addition, eQTLs studies particularly have interrogated in large gene expression from one single tissue or single cell; mostly blood related. Consequently, they have missed the contribution of other types of genetic variation on one hand and the overlooked the specificity of many eQTLs as some genes are expressed in some tissue types and not in others.

In this dissertation I highlight the importance of including Short Tandem Repeats (STRs) in causal variant screening and large-scale studies. As a solution, I provide a deep characterization of STRs and expression STRs (eSTRs) from their distribution and location in the genome and motif enrichment analysis. Diverse regulatory roles of eSTRs are highlighted and hypothesized.

1.2 An overview of the different types of variants in research

Back to basics, the common pipeline for any type of functional analysis involving genetic variants can be summarized as follow: Variant identification and profiling – variant filtering – functional analysis – candidates' variant selection (see Figure 1.1). It is well known that there are different major types of genetic variation in the genome: (1) Single nucleotide variations or SNVs which often include small insertion and deletion (indels) – (2) structural variation or SVs which include a wide range of large sequence variation – and (3) short tandem repeats (STRs) and variable number tandem repeats (VNTRs). These variants are markers that can be surveyed in order to: (1)- find the causal variants to diseases or traits, (2)- understand their roles in the genome and (3)- characterize or identify a whole population or an individual.

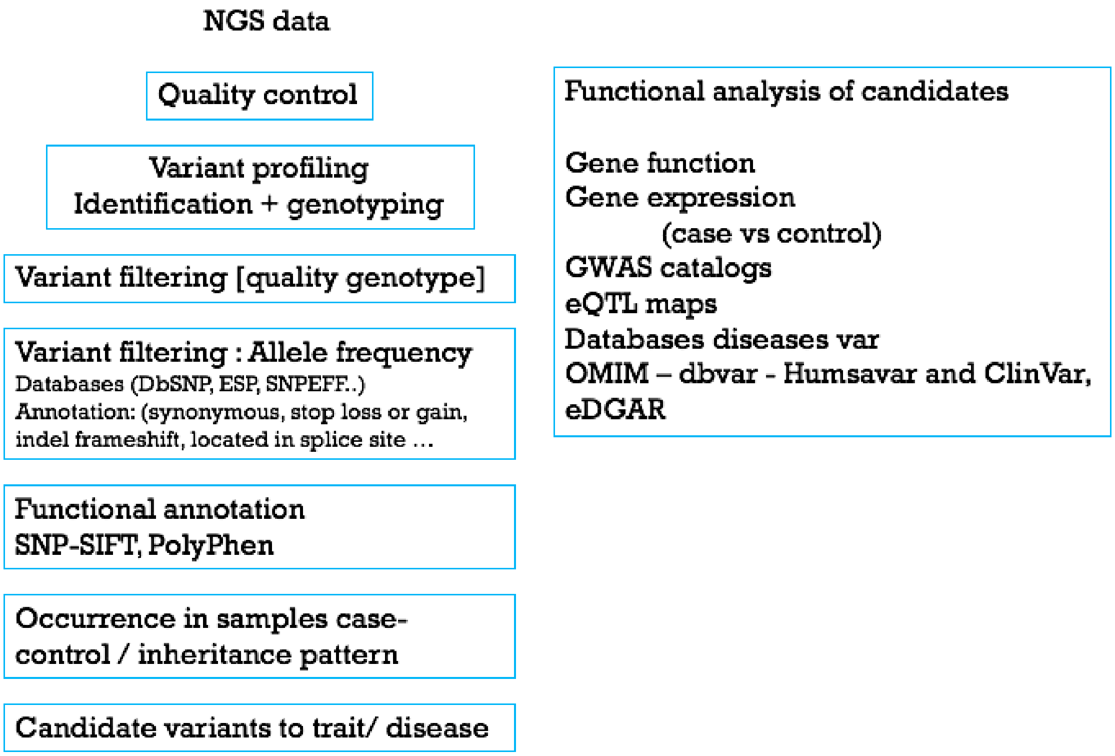

NGS data
Quality control
Variant profiling
Identification + genotyping
Variant filtering [quality genotype]
Variant filtering : Allele frequency
Databases (DbSNP, ESP, SNPEFF..)
Annotation: (synonymous, stop loss or gain,
indel frameshift, located in splice site …
Functional annotation
SNP-SIFT, PolyPhen
Occurrence in samples case-
control / inheritance pattern
Candidate variants to trait/ disease
Functional analysis of candidates
Gene function
Gene expression
(case vs control)
GWAS catalogs
eQTL maps
Databases diseases var
OMIM – dbvar - Humsavar and ClinVar,
eDGAR

Figure 1.1: Standard pipeline for candidates' variant calling

1.2.1 Single Nucleotide variations

Single nucleotide variations or SNVs are DNA changes in which a single base — A, T, C or G — differs from the reference [19]. Often studied with small insertion and small deletion (indels), SNVs have been extensively profiled from next generation sequencing, characterized, studied and their influence on phenotypes is now easily interpreted at the molecular level, both for single and cohort samples. Based on GWAS, there are tens of thousands of SNVs implicated in various diseases and traits, for example Psoriasis [20], Parkinson disease [21,22], Hodgkins_lymphoma and susceptibility to Hodgkins_lymphoma [22]. They have also been the most studied and the most convenient studied polymorphic markers to map disease-causing and disease susceptibility genes as well as those related to drug responsiveness. eQTLs studies have elucidated the influence of SNVs on gene expression dosage. In combination, GWAS and eQTL analysis have helped characterize single nucleotide variants to the extent of creating a big catalog of SNVs-traits association [21,22] used today in various research and clinical settings for functional analysis research, diagnostic purposes and treatment design. A somewhat standard panel for diseases testing has been created for many diseases and used by multiple companies and clinics for diseases diagnosis and disease susceptibility screening. Examples would include the most used panel in cancer research, the breast cancer panel that includes to this day almost 20 SNVs for BRCA1/2 [23]; or, the panels of SNVs used in cardiology for arrhythmia (SNVs in up to 36 genes like ANK6, CACNA1C, KCNE2, etc.), Cardiomyopathy (SNVs in up to 55 genes including MYPN, MYOZ2, MYH7, ACTC1, etc.) and many more. While this has been going on, studies on genetic variants of larger sizes such as structural variations (SV), short tandem repeats (STRs) and variable number tandem repeats (VNTRs), have been lagging.

1.2.2 Structural variations

Structural variations are variation that involve changes of a larger segment of DNA (>50bps). SVs have been proven to cause diseases [24,25], hence the relevance of their inclusion in human genomic and causal variant screening. They have more complex structures and include large insertions, inversions, translocations and copy number variants (CNV), which in turn include large deletions, duplications and multi-allelic copy number variation (mCNVs). Each type of variation has distinct genomic features at detailed sequence level and thus requires different algorithm and sequencing consideration to identify and genotype accurately for downstream analysis.

CNVs are the most common type of SVs surveyed in studies, leaving the other SV types unexplored. CNVs have been well studied, their influence on phenotypes fully characterized globally [26,27] and on many disease phenotypes like malaria, epilepsy, type 2 diabetes, etc. [25,26,28–30] through GWAS; thus they are easily included in genetic studies for clinical phenotypes. SVs have also been linked to diverse phenotypes, but not all types are usually integrated into reference panels of genetic variation screening or large studies.

However, since 2010, there have been efforts from different bioinformatics labs and companies to integrate these variants in genetic studies and research, hence the development of integrated maps for structural variants. These panels include one for large multiallelic copy number variations in humans [31] by the Broad Institute, and an integrated SV map from 26 human populations [32]. In 2016, the second phase of the Genome of the Netherlands (GoNL) project [33] produced a well characterized, haplotype-resolved, SV-integrated reference panel[34], yielded from high coverage sequencing data from hundreds of unrelated family trios and twins [33,35]. From the latter study, almost 200 SNPs previously associated with diseases and other phenotypes were found to be in strong linkage disequilibrium with SVs. These efforts have made more appealing the integration of structural variants in large studies, functional impact studies, such as the most recent survey of the influence of structural variation on gene expression [36].

1.2.3 Short tandem repeats

STRs are distributed throughout the genome and highly variable from one individual to the other. Throughout this dissertation, they will be defined as repeating sequences of unit length up to 6 bps. The mutational characteristics of STRs are the reasons why they have been used for DNA fingerprinting and forensic analysis for decades. STRs have been associated with quite a lot of complex phenotypes such as the Huntington disease [37], oculopharyngeal muscular dystrophy [38], multiple ataxia, Fragile X [39] and many others. While large scale studies have been useful in uncovering the role of multiple genetic variation, (mostly SNV and less often structural variation), they have certainly missed the contribution of STRs. This exclusion has been the clear result of the complexity of these variants and the difficulties associated with their discovery, genotyping or profiling. The past few years have been marked by multiple efforts to profile STRs both on the technological side and also on the research side. In 2014, the first most comprehensive reference panel for STRs was published [40], giving way soon after to the first glimpse at expression STRs [16] and their suggested roles in other phenotypes. The latter study, although very informative and groundbreaking, had some limitation. The use of low coverage data from the 1000G project reduced the power of STR profiling and thus of the analysis itself. Gene expression data from a lymphoblastoid cell line (LCL) only was used, hence the necessity for more analysis to confirm or generalize the results.

1.3 Solution layout

In this dissertation I acknowledge that a good comprehensive catalog of STR variation that is implicated in gene expression fluctuation is fundamental for population studies and adds to the knowledge of demographic history and genotype-phenotype association. We highlight the importance of including STRs in causal variant screening and provide a solution by providing a deep characterization of these variants and their role in regulating gene expression in different contexts. We leverage on the rich dataset from the GTEx project which includes high coverage whole genome sequence data from 650 samples and gene expression (RNA-Seq) from dozens of tissues. We interrogate gene expression from seventeen (17) tissues with big enough samples for statistical power.

In chapter 2, we highlight what is known about STRs and the caveats of common practice of expression quantitative traits analysis and causal variant identification that are limited to the most common and easy to characterize variants like SNVs and CNVs.

In chapter 3, we identify expression STRs (eSTRs) and perform a comprehensive analysis of their distribution in the genome, and their functional role. We then begin to elucidate potential mechanisms by which eSTRs affect gene regulation through eSTRs motifs enrichment for guanine-rich motifs with the potential of forming G-quadruplexes (G4) complexes. These are secondary structures formed by sequences containing consecutive runs of guanines nucleotides, with functional roles in DNA replication and gene transcription [41,42].

Finally, in Chapter 4, we summarize the dissertation, propose future directions from this research and conclude

References

1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C,

Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. Science. 2001 Feb 16;291(5507):1304–51.

2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowki J, International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.

3. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. Complement factor H polymorphism in age-related macular degeneration. Science. 2005 Apr 15;308(5720):385–9.

4. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015 Oct 15;24(R1):R102–10.

5. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol. 2012 Nov;30(11):1095–106.

6. Dajani R, Li J, Wei Z, March ME, Xia Q, Khader Y, Hakooz N, Fatahallah R, El-Khateeb M, Arafat A, Saleh T, Dajani AR, Al-Abbadi Z, Abdul Qader M, Shiyab AH, Bateiha A, Ajlouni K, Hakonarson H. Genome-wide association study identifies novel type II diabetes risk loci in Jordan subpopulations. PeerJ. 2017 Aug 17;5:e3618.

7. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. Philos Trans R Soc Lond B Biol Sci. 2013 May 6;368(1620):20120362.

8. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet. 2013 Nov 7;93(5):779–97.

9. Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. Gene-expression variation within and among human populations. Am J Hum Genet. 2007 Mar;80(3):502–9.

10. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG. Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet. 2007 Feb;39(2):226–31.

11. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavaré S, Deloukas P, Dermitzakis ET. Population genomics of human gene expression. Nat Genet. 2007 Oct;39(10):1217–24.

12. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, Vermeire S, Dewit O, de Vos M, Dixon A, Demarche B, Gut I, Heath S, Foglio M, Liang L, Laukens D, Mni M, Zelenika D, Van Gossum A, Rutgeerts P, Belaiche J, Lathrop M, Georges M. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet. 2007 Apr 20;3(4):e58.

13. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E, Heinzmann A, Simma B, Frischer T, Willis-Owen SAG, Wong KCC, Illig T, Vogelberg C, Weiland SK, von Mutius E, Abecasis GR, Farrall M, Gut IG, Lathrop GM, Cookson WOC. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 2007 Jul 26;448(7152):470–3.

14. Hunt KA, Zhernakova A, Turner G, Heap GAR, Franke L, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, Panesar D, Gwilliam R, Takeuchi F, McLaren WM, Holmes GKT, Howdle PD, Walters JRF, Sanders DS, Playford RJ, Trynka G, Mulder CJJ, Mearin ML, Verbeek WHM, Trimble V, Stevens FM, O'Morain C, Kennedy NP, Kelleher D, Pennington DJ, Strachan DP, McArdle WL, Mein CA, Wapenaar MC, Deloukas P, McGinnis R, McManus R, Wijmenga C,

van Heel DA. Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet. 2008 Apr;40(4):395–402.

15. Liu F, Visser M, Duffy DL, Hysi PG, Jacobs LC, Lao O, Zhong K, Walsh S, Chaitanya L, Wollstein A, Zhu G, Montgomery GW, Henders AK, Mangino M, Glass D, Bataille V, Sturm RA, Rivadeneira F, Hofman A, van IJcken WFJ, Uitterlinden AG, Palstra R-JTS, Spector TD, Martin NG, Nijsten TEC, Kayser M. Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. Hum Genet. 2015 Aug;134(8):823–35.

16. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, Erlich Y. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016 Jan;48(1):22–9.

17. Press MO, Carlson KD, Queitsch C. The overdue promise of short tandem repeat variation for heritability. Trends Genet. 2014 Nov;30(11):504–12.

18. Fan H, Chu J-Y. A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics. 2007 Feb;5(1):7–14.

19. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011 Aug 18;12(9):628–40.

20. de Cid R, Riveira-Munoz E, Zeeuwen PLJM, Robarge J, Liao W, Dannhauser EN, Giardina E, Stuart PE, Nair R, Helms C, Escaramís G, Ballana E, Martín-Ezquerra G, den Heijer M, Kamsteeg M, Joosten I, Eichler EE, Lázaro C, Pujol RM, Armengol L, Abecasis G, Elder JT, Novelli G, Armour JAL, Kwok P-Y, Bowcock A, Schalkwijk J, Estivill X. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat Genet. 2009 Feb;41(2):211–5.

21. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014 Jan;42(Database issue):D1001–6.

22. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017 Jan 4;45(D1):D896–901.

23. Evans DG, Brentnall A, Byers H, Harkness E, Stavrinos P, Howell A, FH-risk study Group, Newman WG, Cuzick J. The impact of a panel of 18 SNPs on breast cancer risk in women attending a UK familial screening clinic: a case-control study. J Med Genet. 2017 Feb;54(2):111–3.

24. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437–55.

25. Sener EF. Association of Copy Number Variations in Autism Spectrum Disorders: A Systematic Review. Chinese Journal of Biology [Internet]. 2014 Nov 16 [cited 2018 Aug 17];2014. Available from: https://www.hindawi.com/journals/cjb/2014/713109/

26. Fu R, Mokhtar SS, Phipps ME, Hoh B-P, Xu S. A genome-wide characterization of copy number variations in native populations of Peninsular Malaysia. Eur J Hum Genet. 2018 Jun 1;26(6):886–97.

27. Repnikova EA, Rosenfeld JA, Bailes A, Weber C, Erdman L, McKinney A, Ramsey S, Hashimoto S, Lamb Thrush D, Astbury C, Reshmi SC, Shaffer LG, Gastier-Foster JM, Pyatt RE. Characterization of copy number variation in genomic regions containing STR loci using array comparative genomic hybridization. Forensic Sci Int Genet. 2013 Sep;7(5):475–81.

28. Monlong J, Girard SL, Meloche C, Cadieux-Dion M, Andrade DM, Lafreniere RG, Gravel M, Spiegelman D, Dionne-Laporte A, Boelman C, Hamdan FF, Michaud JL, Rouleau G, Minassian BA, Bourque G, Cossette P. Global characterization of copy number variants in epilepsy patients from whole genome sequencing. PLoS Genet. 2018 Apr 12;14(4):e1007285.

29. de Jesús Ascencio-Montiel I, Pinto D, Parra EJ, Valladares-Salgado A, Cruz M, Scherer SW. Characterization of Large Copy Number Variation in Mexican Type 2 Diabetes subjects. Sci Rep. 2017 Dec 6;7(1):17105.

30. Cheeseman IH, Miller B, Tan JC, Tan A, Nair S, Nkhoma SC, De Donato M, Rodulfo H, Dondorp A, Branch OH, Mesia LR, Newton P, Mayxay M, Amambua-Ngwa A, Conway DJ, Nosten F, Ferdig MT, Anderson TJC. Population Structure Shapes Copy Number Variation in Malaria Parasites. Mol Biol Evol. 2016 Mar 1;33(3):603–20.

31. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA. Large multiallelic copy number variations in humans. Nat Genet. 2015 Mar;47(3):296–303.

32. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, Konkel MK, Malhotra A, Stütz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Mu XJ, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer E-W, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, 1000 Genomes Project Consortium, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO. An integrated map of structural variation in 2,504 human genomes. Nature. 2015 Oct 1;526(7571):75–81.

33. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet. 2014 Aug;46(8):818–25.

34. Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R, Renkens I, Coe BP, Deelen P, de Ligt J, Lameijer E-W, van Dijk F, Hormozdiari F, Genome of the Netherlands Consortium, Uitterlinden AG, van Duijn CM, Eichler EE, de Bakker PIW, Swertz MA, Wijmenga C, van Ommen G-JB, Slagboom PE, Boomsma DI, Schönhuth A, Ye K, Guryev V. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. Nat Commun. 2016 Oct 6;7:12989.

35. Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, Ye K, Guryev V, Vermaat M, van Dijk F, Francioli LC, Hottenga JJ, Laros JFJ, Li Q, Li Y, Cao H, Chen R, Du Y, Li N, Cao S, van Setten J, Menelaou A, Pulit SL, Hehir-Kwa JY, Beekman M, Elbers CC, Byelas H, de Craen AJM, Deelen P, Dijkstra M, den Dunnen JT, de Knijff P, Houwing-Duistermaat J, Koval V, Estrada K, Hofman A, Kanterakis A, Enckevort D van, Mai H, Kattenberg M, van Leeuwen EM, Neerincx PBT, Oostra B, Rivadeneira F, Suchiman EHD, Uitterlinden AG, Willemsen G, Wolffenbuttel BH, Wang J, de Bakker PIW, van Ommen G-J, van Duijn CM. The Genome of the Netherlands: design, and project goals. Eur J Hum Genet. 2014 Feb;22(2):221–7.

36. Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y. The impact of structural variation on human gene expression. bioRxiv [Internet]. 2016; Available from: http://www.biorxiv.org/content/early/2016/06/09/055962.abstract

37. Sutherland GR, Richards RI. Simple tandem DNA repeats and human genetic disease. Proc Natl Acad Sci U S A. 1995 Apr 25;92(9):3636–41.

38. Piccolo G, Cortese A, Tavazzi E, Piccolo L, Sassone J, Ciammola A, Alfonsi E, Colombo I, Moggio M. Late onset oculopharyngeal muscular dystrophy with prominent neurogenic features and short GCG trinucleotide expansion. Muscle Nerve. 2011 Jan;43(1):141–2.

39. Richards RI, Holman K, Yu S, Sutherland GR. Fragile X syndrome unstable element, p (CCG) n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. Hum Mol Genet. 1993;2(9):1429–35.

40. Willems T, Gymrek M, Highnam G, 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014 Nov;24(11):1894–904.

41. Lin C, Yang D. Human Telomeric G-Quadruplex Structures and G-Quadruplex-Interactive Compounds. Methods Mol Biol. 2017;1587:171–96.

42. Hänsel-Hertsch R, Beraldi D, Lensing SV, Marsico G, Zyner K, Parry A, Di Antonio M, Pike J, Kimura H, Narita M, Tannahill D, Balasubramanian S. G-quadruplex structures mark human regulatory chromatin. Nat Genet. 2016 Oct;48(10):1267–72.

CHAPTER 2: REVIEW AND IMPLICATION OF THE INCLUSION STR IN STUDIES

2.1 Barriers to the inclusion of STRs in studies

Here we define STRs as sequences of up to 6 bps units that repeat themselves multiple times. Their total lengths may vary from 10 bps (for shorter mononucleotides or dinucleotide repeats) to a few hundred. They are difficult to genotype in general, but even more difficult are mononucleotide repeats. Much recently, most algorithms [1,2] required reads spanning the STRs in order to confirm the actual length of STRs. They are distributed throughout the genome and highly polymorphic, which confers them the advantage of being used for DNA fingerprinting and forensic analysis. The exceptions to this rule are mononucleotides, which are less polymorphic, as variation in these sequences can lead to nonsynonymous and frameshift mutations, especially in coding regions, which have the reputation of being deleterious. Mononucleotides are selected against longer sizes [3] and areas around these repeats can be highly polymorphic [4]. Variations in STR repeat numbers are likely to be identified as indels hence the term "false indels." In addition, unless using amplification-free (PCR-free) sequencing method, their polymorphic nature will more likely introduce stutter noise, [1,2] which are errors in STR length introduced by DNA slippage. However, PCR-free is not always possible for single cell analysis and targeted-enrichment protocols.

According to the public library of medicine (PUBMED), in the past five years, there have been more than five hundred publications that address short tandem repeats in all aspects of research, including algorithm development for identification or profiling, methods for analyses, implication in diseases and cancers, and method and algorithm improvement and reviews. This is an improvement compared to the previous years, and we reiterate their importance. In this Chapter, we present a case for an updated focus specifically on short tandem repeats and an imperative investigation of their role in gene regulation. We highlight that they carry information that cannot be overlooked, and that the list of complex phenotypes driven by STRs continues to grow.

2.2 Overview on homopolymers: a complex class of STRs

Mononucleotide repeats or homopolymers are sequences made of repeating unique nucleotides noted poly(dA).poly(dT) and poly(dG).poly(dC). In general, poly(dA) and poly(dT) are used interchangeably and so are poly(dG) and poly(dC); herein named poly-A and poly-G. Homopolymers constitute the largest subset of microsatellites in the human genome followed by dinucleotides. Just like STRs of larger unit size, homopolymers are difficult to genotype and interpret, but their boundaries are easier to define. Mononucleotide repeats existence across species have been well documented, and well characterized in prokaryotes [5] and various Eukaryotes [3]. Their role in the human genome, however, is not very well understood, but several implications in different functional mechanisms in cells have been documented and hypothesized. We sought to assess what is currently known so far of this type of repeats in human. In addition, with the ever-increasing availability of high coverage sequencing data and better methods for genotyping these repeats, we conducted a brief survey and characterization of mononucleotide repeats in the human genome using the extensive resource, the GTEx dataset. We evaluated the distribution of the homopolymers, and we investigated their overall influence on gene expression in different tissues, which led to infer their potential role in gene regulation.

2.2.1 Characteristics of mononucleotides in literature

The accurate identification of homopolymer length can be difficult. However, thanks to advances in sequencing technology and gold standards, significant progress has been made in this field. They are the most abundant type of repeats [6].

From the biological standpoint, mononucleotide repeats are less polymorphic than other STRs of larger unit sizes. They are under natural selection against longer sizes, especially in coding regions[3]. They are difficult for cells to replicate accurately and more prone to insertion and deletion variations [4]. Different hypotheses have been enunciated as the cause for their abundance in the human genome, including slipped-strand mispairing and indel slippage. In slipped-strand mispairing, a misalignment of the two DNA strands by repeat step-size (here 1) multiple is more likely to occur. And once it does, a DNA replicating enzyme can correct it by either adding or removing the repeats (mismatch repair). On the other hand, during DNA polymerase-mediated DNA duplication, the repetitive sequence with mismatches is inserted or deleted: this is called indel slippage.

On the functional front, mononucleotide repeat variations are largely implicated in cancer and tumor genesis [7–10]. Mononucleotide repeats have been linked to multiple cancers [8–10]. They are of high occurrence in, and characteristic of, cancerous tumors [11,12]. Scientists have leveraged this characteristic to evaluate microsatellite instability in tumors and thus detect DNA mismatch repair-deficiency [9,10]. In fact, a brief search of the terms "mononucleotide repeats" or "homopolymers" on PUBMED quickly yields hundreds of cancer-related results. Broken down by cancer types, colorectal cancer is by far the leader. One hypothesizes that mononucleotide repeats play a big role in this specific cancer type. It turns out in fact that different panels of mononucleotide repeat markers have been created for the identification of colorectal tumors and to predict the outcome of such cancers [7–10,13]. Additionally, research showed that mononucleotide repeats can disrupt gene expression in many ways [3,14], including forming binding sites upstream (close to transcription Start Sites TSS) of many housekeeping genes [14]. Another speculated mechanism is by acting as promoters, especially the poly-A tracts [15]. They also play a role in disease occurrence and susceptibility [7,13,16]. Variations in these repeats are likely to create frameshifts and or non-synonymous variants in coding regions [12,17], thus altering the function of the gene itself. Additionally, specific human repetitive sequences, such as Alu elements, contain long poly-A tracts [18] called A-tails for which the length determines whether the Alu element is active or not (i.e. its retro-positional capability [19]). Poly-G tracks not only contribute to the overall GC richness, but may also form strong hairpins and other complex secondary structures [20,21].

2.2.2 Distribution of homopolymers and their influence on gene expression

Up until now, no study has systematically reviewed homopolymers distribution and their role in the human genome. Here, we used the dataset from the Genotype-Tissue Expression (GTEx) project [22] to perform a genomic review of the distribution of homopolymers the human genome and later, in Chapter 3, they are included as part of all STRs in the evaluation of the role of STRs in gene regulation. A succinct summary of their potential influence on gene expression variability is presented here.

(a) Distribution of homopolymers in the genome

We used the HipSTR tool [1] to genotype mononucleotide STRs in 650 high coverage (30x) whole genome sequencing from GTEx. After quality control of the calls, HipSTR identified and genotyped 828,971 homopolymers runs across samples, which represented about 52.25% of all genotyped STRs. Their total lengths varied between 10 and 71 bps. In terms of their distribution throughout the genome, the most abundant homopolymer type, as expected, were the poly-A tracts accounting for 99.5% of all homopolymers and 0.5% (7,437) for poly-C tracks. By location, 50.74% of all homopolymers were found in introns of genes, less than 0.1% in coding regions and only 1.5% in UTR regions. Up to 75% of UTR mononucleotides repeats were located in 3-prime UTRs, which may be where they contribute to gene regulation. We confirmed that less than 2% of homopolymers are found within the gene transcript, which makes sense for protein conservation and selection against polymorphisms.

(b) Summarizing e-homopolymers in the genome

Almost 73% were removed for downstream analysis for quality control reasons, such as call rate (threshold of 80%), failing Hardy Weinberg Equilibrium (HWE), for being next to (in continuation of) another STR, or falling in segmental duplication regions (this is a rare occurrence). This filtering proved to be important for the identification of high-quality eSTRs mononucleotide repeats. After this careful filtering, a small fraction 12,822 (<1.6%), were screened for potential eQTLs (e-homopolymers) across 17 tissues. Using a linear regression model relating the average homopolymer length, to nearby (cis) gene expression, and accounting for covariates (race, confounding factor in expression). We identified 10,250 mononucleotide repeats as e-homopolymers for 7070 genes at 10% FDR. Almost 31.6% (3237) of them were shared by two or more tissues.

To delineate between tagging effect and possibly causal e-homopolymers, we used a two-layer verification (ANOVA and CAVIAR) as fully described in Chapter 3. Briefly, with ANOVA, we compared two linear models: the SNP-only model vs. SNP+STR model, using the anova_lm function in the python statsmodels.api.stats module. This told us whether adding eSTRs to the model, improve the model. With CAVIAR, we fine-mapped eSTR signals against the top 100 eSNPs and other STRs within 100 kb upstream and downstream of each gene. We set criteria for causal e-homopolymer as follow: if the e-homopolymer adds more information to explain gene expression variability by ANOVA ($P<0.05$) after FDR correction, and the posterior probability for being the causal variant >10% by CAVIAR. Following these criteria, we found that 11.6% of e-homopolymers (1189) were identified as causal, with 57 at > 90% causality score per CAVIAR. See summary in Table 2.1).

Table 2.1: Table summary of e-homopolymer identification

Tissue	**Sample size**	**Number of e-homopolymers**	**Causal e-homopolymers**	**Causal with score>50%**
Nerve Tibial	265	1690	175	129
Thyroid	262	1620	164	140
Skin Sun-Exposed	297	1494	146	111
Artery Tibial	276	1375	136	109
Esophagus Mucosa	255	1363	140	114
Adipose Subcutaneous	270	1358	113	104
Muscle Skeletal	343	1319	121	89
Transformed Fibroblasts	225	1246	117	87
Lung	259	1199	116	115
Esophagus Muscularis	214	1149	123	111
Whole Blood	336	1035	76	75
Artery Aorta	191	972	112	86
Skin Not-Sun-Exposed	209	956	98	80
Adipose Visceral	193	714	77	72
Heart Ventricle	199	686	102	85
Brain Cerebellum	107	588	67	72
Brain Caudate	108	272	30	45

e-homopolymers were enriched for being within 1.5 Kb of TSS regions, more than 2 folds for causal homopolymers eSTRs. Poly-C were rare but more likely to be eSTRs, contributing to GC-rich enrichment. These results highlight the non-negligible potential for homopolymers in gene regulation role and thus the importance of further screening and their inclusion in downstream analysis like causal screening and large-scale studies. Examples of top causal e-homopolymers included a Poly-A in the 3' UTR region of SLC15A4 gene on chromosome 12 and a Poly-C in the promoter region of the CRMP1 gene on chromosome4. (Figure 2.1).

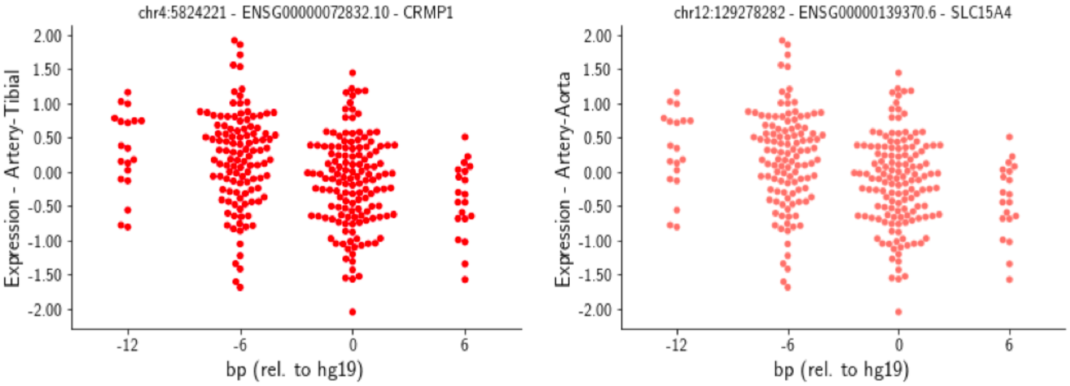


Figure 2.1: Example putative causal e-homopolymer. Left and right plots give HipSTR STR dosage (red and salmon) vs. normalized gene expression (y-axis). STR dosage is defined as the average length difference from hg19. One dot represents one sample. P-values are obtained using linear regression of genotype vs. gene expression.

2.3 Causal variant screening: limitation on variant types surveyed

Findings from large-scale studies have been used and applied in the development of new clinical practices. It is in these clinical setting that patients with rare diseases, un-named diseases, or unexplained diseases come with the hope of finding, if not a cure, at least an explanation for their conditions. For cases where hypotheses have already been generated from the symptoms, reported cases and previous associations in the literature like genome wide association studies are available; the recourse to targeted sequencing or exome sequencing may be used to verify the suspected culprits. However, in many cases, whole genome sequencing is the way to start in order to generate these hypotheses.

The common practice then is to use genome sequencing of the patient and family members, preferably parents and or sibling(s) if a pattern of inheritance or family history is suspected. In these scenarios, next-generation sequencing has been the main method for acquiring genetic data. Briefly, the common practice of causal or candidate variants screening, no matter whether SNVs, SVs or STRs, involves the following steps with quality control and sanity checks at each step:

- Variant identification and profiling or genotyping
- Variant filtering
- Candidate variants selection
- Functional analysis of candidates usually supported by functional annotation with databases like the Online Mendelian Inheritance in Man (OMIM) database, the Human Gene Mutation Database (HGMD), the archive of relationships between sequence variation and human phenotype (CLINVAR). Other software and tools for annotation include the program for predicting the impact of variants by Sorting Intolerant From Tolerant (SIFT), the Genomic Evolutionary Rate Profiling (GERP), the tool for predicting the possible impact of amino acid substitution on the sequence structure and protein function (Polyphen), etc.
- Validation using model organism, in vitro or iPSC

Functional analysis is a critical part of the whole process. Overall, the diversity in variant types and structure requires a separate pipeline for each type of variation to be surveyed. As mentioned in the beginning, the regular practice only focuses on SNPs markers and ignores complex variants and variants of larger size. But such practice limitations.

As a use case example of such failure, consider a patient presenting symptoms of an uncategorized or ambiguous “Spino-cerebellar ataxia” (SCA) symptoms. SCA is a well-known neurological disease among neurologists and clinicians, a very heterogeneous condition. While patients all share a common trait: (problems with coordination and balance - ataxia), their clinical diagnosis is a difficult task for clinical neurologists because of their high heterogeneity in genotypes and phenotypes as well as a stacking overlap between phenotypes (here symptoms) in different types and subtypes [23]. Today, there are over 35 types of SCA and even more, if counting the subtypes, identified with different causes. But the majority have been linked to multiple STRs, VNTRs in different genes and chromosomes [23]. Only a few would be identified with a SNV panel (like the rare SCA11 caused by small intragenic indels) [24] or a structural variant panel (like SCA15 caused by large deletions)[25,26]. In our uncategorized ataxia case, a full screen that includes all variant types would be necessary. This is only known because of the different studies conducted in this well-documented condition [23,25–29]. The same cannot be said for other diseases. Because the genome harbors different types of variants, which disrupt processes and molecular mechanisms differently, this makes the case for not limiting causal variants screening to SNVs and or SVs which may often lead to dead ends.

2.4 The importance of tissue and cell type in eQTLs

Gene regulation is a complex machinery influenced by feedback loops and many other factors. In fact, transcription regulation is context specific, as a result of different factors coming together such as the DNA sequence itself, the nucleosome positioning, DNA binding proteins like transcription factors, histone modification, and other non-coding RNAs [30]. These elements together modulate gene transcription differently depending on the cell type, hence the term “Cell specific” transcription [31]. For these reasons, the regulatory mechanisms that lead to a disease can be tissue- or cell-specific as opposed to being observed across different tissues. In the latter case, the relevant regulatory activity may not always be detected in the tissue most relevant to the pathobiology of a causal gene. An example is the work that was recently done by Hao Mei et al.[32]. They looked at T2D associated gene expression across 44 tissues and found that expression of genes can be tissue specific. While association signals of some genes can be strong in some tissues, other tissues may just negate the existence of any association to the trait altogether, hence the importance of multiple tissue analysis. The hypothesis is that it is not always possible to identify tissues that are most relevant for a given disease or phenotype, but we can identify the variant(s) or loci associated to the diseases-causing gene(s) in different tissue types, hence identifying the relevant regulatory mechanism that leads to the disease/phenotype. It then becomes possible to investigate new relevant regulatory activity. In the next chapter, we explore patterns of gene expression across different tissues and characterize expression STRs.

2.5 Inclusion in large scale studies

The purpose of large scale studies like eQTL and GWAS is to leverage data from a population to identify association and best-case scenario causal relations between genetic markers and a given trait that, in the eQTL case, would be gene expression. Large scale studies have led to the creation of reference panels [33–35], a key to linkage disequilibrium-based imputation. Imputation is a statistical method for inferring unknown genotypes from known genotypes, which in turn improves the power of combining test across studies and the power of other GWAS studies [36]. From GWAS studies, catalogs for variants linked to multiple traits are also created [37,38] for use by clinical and translational bioinformaticians for the identification of causal variants to diseases, to understand various biological mechanisms that lead to diseases, and thus infer targeted therapies and treatments. eQTLs analysis, on the other hand, has led to an even better understanding of these association between genes and traits or variants and traits [39–41]. Succinctly, many variants associated with traits or diseases fall within non-coding and regulatory regions and will influence gene expression that way. The role of eQTL analysis is to shed the light on those regulatory mechanisms. It promises characterization of functional sequence variation and/or the understanding of basic processes of gene regulation.

Today, we have a large amount of literature on eQTLs SNVs, but very few on larger size variants like STRs and SVs. In addition, most eQTLs are conducted in the context of a particular condition only survey gene expression from the tissue or cell type directly impacted or suspected to be relevant to the trait or condition. We pointed earlier in section 2.4 the importance of surveying gene expression in different tissues, as available for a clearer mapping.

To even the plain field of causal variant screening and functional research, it is imperative to integrate all types of variants. Such integration is only possible if proper characteristics of the variants are available for reliable interpretation. Tools used for variants characterization include: databases that contain the reference human genome sequence, a map of variants and a set of algorithms and technologies that can quickly and accurately analyze, in context (a quantifiable phenotype), whole-genome samples. The result is a set of genetic variations that contribute to the phenotype variation, like the onset of a disease.

2.6 Improving large-scale study through imputation

Generally, GWAS studies survey hundreds and sometimes thousands of samples. To limit the cost of WGS for all individuals in the study, as would be ideal, a large panel of SNVs (which are highly polymorphic single nucleotide variations) are genotyped across all individuals. However, GWAS power is a function of both the number of samples in the study and the number of loci or variants genotyped in the study, i.e., the larger the better. Imputation is a statistical method that uses information from a reference panel which either (1) contains a larger number of genotyped loci or (2) covers a larger portion of the genome sequence - to infer unknown genotypes of skipped loci. This operation adds to the number of loci in the study and thus improves the power of GWAS studies [36]. It also improves the power of combining tests across studies. With a reference panel for STRs, association studies can now be conducted and improved with imputation. Following the recent wave of interest in these variants, we recently conducted a preliminary and successful eQTL analysis on simulated phenotypes and found that STR imputation improved the power to the detect association in association studies [42]. We conducted a short eQTLs analysis using real data that appears in Saini et al.,2018 [42] and found that STR imputation could identify STR associations using real phenotypes, as described in [42] and below. This study also showed that STRs could be imputed from SNPs and vice-versa.

We conducted a short eQTLs analysis using real data from the GTEx [22] project and imputed STRs genotypes for chromosome 21. STR genotypes for real data were obtained using the HipSTR tool. Imputed STRs genotyped were obtained using SNP data from the same samples. We focused on a linear additive model relating STR dosage, defined as the average allele length, since the majority of known functional STRs follow similar models (e.g.,[43–46]) and nearby gene expression, given the large number of reported associations between STR length and expression of nearby genes in cis[44,47] (termed eSTRs). To this end, we analyzed eSTRs from samples in the GTEx dataset for which RNA-sequencing, WGS, and SNP array data were available. As a test case, we imputed STR genotypes using SNP data for chromosome 21 and tested for association with genes expressed in whole blood. For comparison, we additionally performed each association using genotypes obtained from WGS using the HipSTR tool[1]. A total of 2,452 (STR x gene) tests were performed in each case. Association p-values were similarly distributed across both analyses and showed a strong departure from the uniform distribution expected under a null hypothesis of no STR association (Figure 2.2A). For all nominally significant associations ($P<0.05$), effect sizes were strongly correlated when using imputed vs. HipSTR genotypes ($r=0.99$; $p=1.01x10^{-79}$, $n=97$). Furthermore, effect sizes obtained from imputed data were concordant with previously reported effect sizes in a separate cohort using a different cell type (lymphoblastoid cell lines) ($r=0.79$; $p=0.0042$, $n=11$) (Figure 2.2B). We identified genes for which the STR is most likely the causal variant and tested whether STR imputation had greater power to identify causal eSTRs compared to SNP-based analyses. We used ANOVA model comparison to determine genes for which the STR explained additional variation over the top SNP. We compared two linear models: Y~eSNP (SNP-only model) vs. Y~eSNP+eSTR (SNP+STR model) using the anova_lm function in the python statsmodels.api.stats module. Additionally, we used CAVIAR v1.0 [48] to further fine-map eSTR signals against the top 100 eSNPs within 100kb of the TSS and TES of each gene. Pairwise-LD between the eSTR and eSNPs was estimated using Pearson correlation between SNP dosages (0, 1, or 2) and STR dosages (sum of the two repeat allele lengths). We identified 3 genes with ANOVA p-value $P<0.05$ for which the STR was the top variant returned by CAVIAR. One example, a CG-rich STR in the promoter of CSTB, was previously demonstrated to act as an eSTR [49] and expansions of this repeat are implicated in myoclonus epilepsy [50]. In each case, imputed STR genotypes were more strongly associated with gene expression compared to the best tag SNP (Figure 2.2C-D).

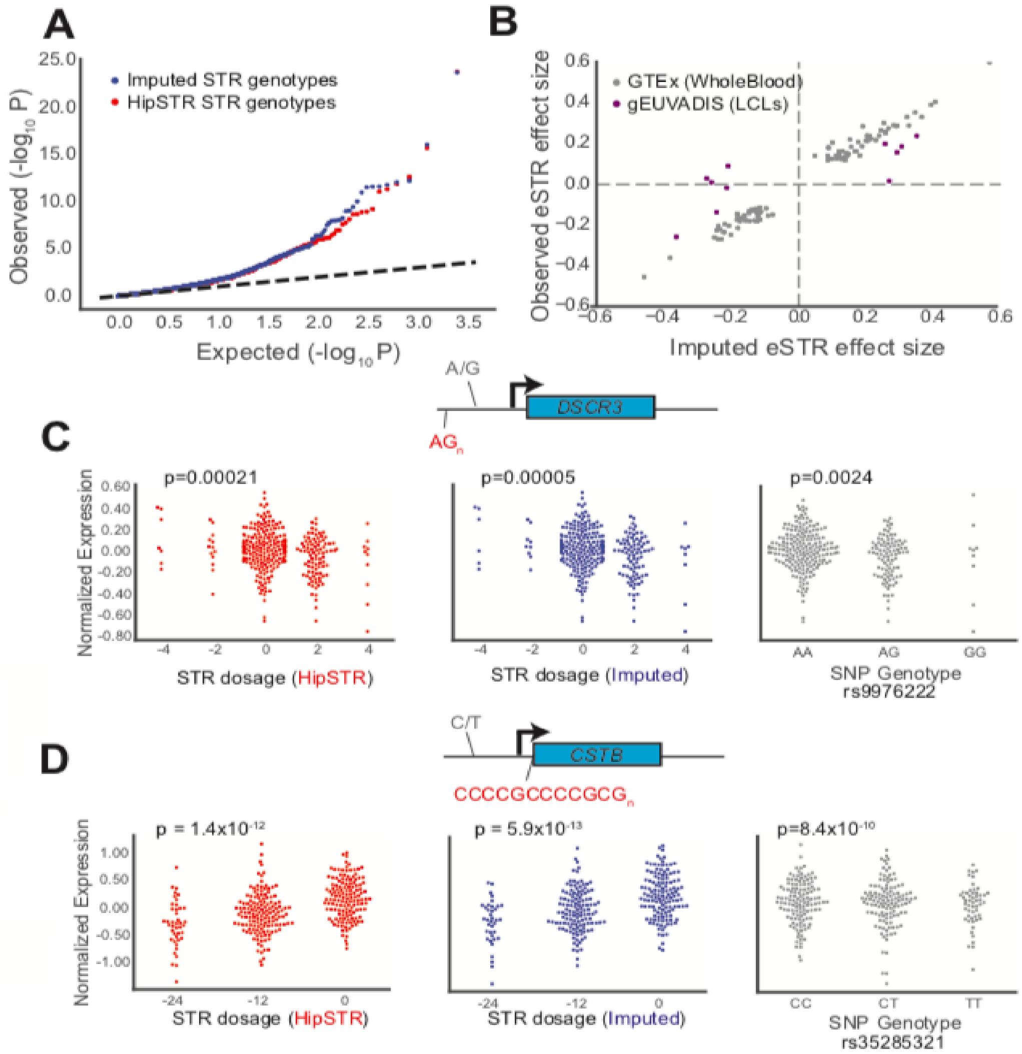


Figure 2.2: STR imputation improves power to detect STR associations. A. Quantile-quantile plot for eSTR association tests. Each dot represents a single STR × gene test. The x-axis gives the expected log10 p-value distribution under a null model of no eSTR associations. Red and blue dots give log10 p-value for association tests using HipSTR genotypes and imputed STR genotypes, respectively. The black dashed line gives the diagonal. B. Comparison of eSTR effect sizes using observed vs. imputed genotypes. Each dot represents a single STR × gene test. The x-axis gives effect sizes obtained using imputed genotypes. Gray dots give the effect size in GTEx whole blood using HipSTR genotypes. Purple dots give effect sizes reported previously in lymphoblastoid cell lines. C. and D. Example putative causal eSTRs identified using imputed STR genotypes. Left, middle, and right plots give HipSTR STR dosage (red), imputed STR dosage (blue), and the best tag SNP genotype (gray) vs. normalized gene expression, respectively. STR dosage is defined as the average length difference from hg19. One dot represents one sample. P-values are obtained using linear regression of genotype vs. gene expression. STR and SNP sequence information is shown for the coding strand. Gene diagrams are not drawn to scale.

Conclusion

STRs have been linked to multiple disease phenotypes and gene expression variability. However, it is crucial to properly characterize and understand their implication and gene regulation in different contexts in humans. Herein, we presented the case for an imperative full view on STRs and the implication of such characterization in research and the clinical setting. We also show the importance of including mononucleotides STRs in such analysis.

Chapter 2 section 2.6 contains parts of material from Shubham Saini, Ileena Mitra, Nima Mousavi, Stephanie Feupe Fotsing, Melissa Gymrek - A Reference Haplotype Panel for Genome-wide Imputation of Short-Tandem Repeats. The dissertation author was responsible for this part of the analysis, but not the primary author of the paper.

References

1. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Methods. 2017 Jun;14(6):590–2.

2. Bahlo M, Bennett MF, Degorski P, Tankard RM, Delatycki MB, Lockhart PJ. Recent advances in the detection of repeat expansions with short-read next-generation sequencing. F1000Res [Internet]. 2018 Jun 13;7. Available from: http://dx.doi.org/10.12688/f1000research.13980.1

3. Gu T, Tan S, Gou X, Araki H, Tian D. Avoidance of long mononucleotide repeats in codon pair usage. Genetics. 2010 Nov;186(3):1077–84.

4. Siddle KJ, Goodship JA, Keavney B, Santibanez-Koref MF. Bases adjacent to mononucleotide repeats show an increased single nucleotide polymorphism frequency in the human genome. Bioinformatics. 2011 Apr 1;27(7):895–8.

5. Coenye T, Vandamme P. Characterization of mononucleotide repeats in sequenced prokaryotic genomes. DNA Res. 2005;12(4):221–33.

6. Feng W, Zhao S, Xue D, Song F, Li Z, Chen D, He B, Hao Y, Wang Y, Liu Y. Improving alignment accuracy on homopolymer regions for semiconductor-based sequencing technologies. BMC Genomics. 2016 Aug 22;17 Suppl 7:521.

7. Takehara Y, Nagasaka T, Nyuya A, Haruma T, Haraga J, Mori Y, Nakamura K, Fujiwara T, Boland CR, Goel A. Accuracy of four mononucleotide-repeat markers for the identification of DNA mismatch-repair deficiency in solid tumors. J Transl Med. 2018 Jan 12;16(1):5.

8. Hienonen T, Sammalkorpi H, Enholm S, Alhopuro P, Barber TD, Lehtonen R, Nupponen NN, Lehtonen H, Salovaara R, Mecklin J-P, Järvinen H, Koistinen R, Arango D, Launonen V, Vogelstein B, Karhu A, Aaltonen LA. Mutations in Two Short Noncoding Mononucleotide Repeats in Most Microsatellite-Unstable Colorectal Cancers. Cancer Res. 2005 Jun 1;65(11):4607–13.

9. Pagin A, Zerimech F, Leclerc J, Wacrenier A, Lejeune S, Descarpentries C, Escande F, Porchet N, Buisine M-P. Evaluation of a new panel of six mononucleotide repeat markers for the detection of DNA mismatch repair-deficient tumours. Br J Cancer. 2013 May 7;108:2079.

10. Alves IT, Cano D, Böttcher R, van der Korput H, Dinjens W, Jenster G, Trapman J. A mononucleotide repeat in PRRT2 is an important, frequent target of mismatch repair deficiency in cancer. Oncotarget. 2017 Jan 24;8(4):6043–56.

11. Yamamoto H, Perez-Piteira J, Yoshida T, Terada M, Itoh F, Imai K, Perucho M. Gastric cancers of the microsatellite mutator phenotype display characteristic genetic and clinical features. Gastroenterology. 1999 Jun;116(6):1348–57.

12. Schwartz S Jr, Yamamoto H, Navarro M, Maestro M, Reventós J, Perucho M. Frameshift mutations at mononucleotide repeats in caspase-5 and other target genes in endometrial and gastrointestinal cancer of the microsatellite mutator phenotype. Cancer Res. 1999 Jun 15;59(12):2995–3002.

13. Brennetot C, Buhard O, Jourdan F, Flejou J-F, Duval A, Hamelin R. Mononucleotide repeats BAT-26 and BAT-25 accurately detect MSI-H tumors and predict tumor content: implications for population screening. Int J Cancer. 2005 Jan 20;113(3):446–50.

14. Aporntewan C, Pin-on P, Chaiyaratana N, Pongpanich M, Boonyaratanakornkit V, Mutirangura A. Upstream mononucleotide A-repeats play a cis-regulatory role in mammals through the DICER1 and Ago proteins. Nucleic Acids Res. 2013 Oct;41(19):8872–85.

15. Zhou Y, Bizzaro JW, Marx KA. Homopolymer tract length dependent enrichments in functional regions of 27 eukaryotes and their novel dependence on the organism DNA (G+C)% composition. BMC Genomics. 2004 Dec 14;5:95.

16. Kim MS, Yoo NJ, Lee SH. Frameshift mutation at mononucleotide repeat in ERCC5 in gastric carcinomas with microsatellite instability. Pathology. 2009;41(4):394–5.

17. Ikenoue T, Togo G, Nagai K, Ijichi H, Kato J, Yamaji Y, Okamoto M, Kato N, Kawabe T, Tanaka A, Matsumura M, Shiratori Y, Omata M. Frameshift mutations at mononucleotide repeats in RAD50 recombinational DNA repair gene in colorectal cancers with microsatellite instability. Jpn J Cancer Res. 2001 Jun;92(6):587–91.

18. Roy AM, Carroll ML, Nguyen SV, Salem AH, Oldridge M, Wilkie AO, Batzer MA, Deininger PL. Potential gene conversion and source genes for recently integrated Alu elements. Genome Res. 2000 Oct;10(10):1485–95.

19. Roy-Engel AM, Salem A-H, Oyeniran OO, Deininger L, Hedges DJ, Kilroy GE, Batzer MA, Deininger PL. Active Alu element “A-tails”: size does matter. Genome Res. 2002 Sep;12(9):1333–44.

20. Sengar A, Heddi B, Phan AT. Formation of G-Quadruplexes in Poly-G Sequences: Structure of a Propeller-Type Parallel-Stranded G-Quadruplex Formed by a G15 Stretch. Biochemistry. 2014 Dec 16;53(49):7718–23.

21. Lewis FD, Wu Y, Zhang L. Reversible formation of DNA G-quadruplex hairpin dimers from stilbenediether conjugates. Chem Commun . 2004 Mar 5;0(6):636–7.

22. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013 Jun;45(6):580–5.

23. Tan EK, Ashizawa T. Genetic testing in spinocerebellar ataxias: defining a clinical role. Arch Neurol. 2001 Feb;58(2):191–5.

24. Bauer P, Stevanin G, Beetz C, Synofzik M, Schmitz-Hübsch T, Wüllner U, Berthier E, Ollagnon-Roman E, Riess O, Forlani S, Mundwiller E, Durr A, Schöls L, Brice A. Spinocerebellar ataxia type 11 (SCA11) is an uncommon cause of dominant ataxia among French and German kindreds. J Neurol Neurosurg Psychiatry. 2010 Nov;81(11):1229–32.

25. van de Leemput J, Chandran J, Knight MA, Holtzclaw LA, Scholz S, Cookson MR, Houlden H, Gwinn-Hardy K, Fung H-C, Lin X, Hernandez D, Simon-Sanchez J, Wood NW, Giunti P, Rafferty I, Hardy J, Storey E, Gardner RJM, Forrest SM, Fisher EMC, Russell JT, Cai H, Singleton AB. Deletion at ITPR1 underlies ataxia in mice and spinocerebellar ataxia 15 in humans. PLoS Genet. 2007 Jun;3(6):e108.

26. Marelli C, van de Leemput J, Johnson JO, Tison F, Thauvin-Robinet C, Picard F, Tranchant C, Hernandez DG, Huttin B, Boulliat J, Sangla I, Marescaux C, Brique S, Dollfus H, Arepalli S, Benatru I, Ollagnon E, Forlani S, Hardy J, Stevanin G, Dürr A, Singleton A, Brice A. SCA15 due to large ITPR1 deletions in a cohort of 333 white families with dominant ataxia. Arch Neurol. 2011 May;68(5):637–43.

27. Lee Y-C, Tsai P-C, Guo Y-C, Hsiao C-T, Liu G-T, Liao Y-C, Soong B-W. Spinocerebellar ataxia type 36 in the Han Chinese. Neurol Genet. 2016 Jun;2(3):e68.

28. Edener U, Bernard V, Hellenbroich Y, Gillessen-Kaesbach G, Zühlke C. Two dominantly inherited ataxias linked to chromosome 16q22.1: SCA4 and SCA31 are not allelic. J Neurol. 2011 Jul;258(7):1223–7.

29. Holmes SE, O'Hearn E, Margolis RL. Why is SCA12 different from other SCAs? Cytogenet Genome Res. 2003;100(1-4):189–97.

30.Venters BJ, Pugh BF. How eukaryotic genes are transcribed. Crit Rev Biochem Mol Biol. 2009 Jun;44(2-3):117–41.

31. Gibcus JH, Dekker J. The context of gene expression regulation. F1000 Biol Rep. 2012 Apr 2;4:8.

32. Mei H, Li L, Liu S, Jiang F, Griswold M, Mosley T. Tissue Non-Specific Genes and Pathways Associated with Diabetes: An Expression Meta-Analysis. Genes [Internet]. 2017 Jan 21;8(1). Available from: http://dx.doi.org/10.3390/genes8010044

33. Chou W-C, Zheng H-F, Cheng C-H, Yan H, Wang L, Han F, Richards JB, Karasik D, Kiel DP, Hsu Y-H. A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples. Sci Rep. 2016 Dec 22;6:39313.

34. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061–73.

35. International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec 18;426(6968):789–96.

36. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387–406.

37. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014 Jan;42(Database issue):D1001–6.

38. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017 Jan 4;45(D1):D896–901.

39. Liang L, Morar N, Dixon AL, Lathrop GM, Abecasis GR, Moffatt MF, Cookson WOC. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. Genome Res. 2013 Apr;23(4):716–26.

40. Koopmann TT, Adriaens ME, Moerland PD, Marsman RF, Westerveld ML, Lal S, Zhang T, Simmons CQ, Baczko I, dos Remedios C, Bishopric NH, Varro A, George AL Jr, Lodder EM, Bezzina CR. Genome-wide identification of expression quantitative trait loci (eQTLs) in human heart. PLoS One. 2014 May 20;9(5):e97380.

41. Gillies CE, Putler R, Menon R, Otto E, Yasutake K, Nair V, Hoover P, Lieb D, Li S, Eddy S, Fermin D, McNulty MT, Nephrotic Syndrome Study Network (NEPTUNE), Hacohen N, Kiryluk K, Kretzler M, Wen X, Sampson MG. An eQTL Landscape of Kidney Tissue in Human Nephrotic Syndrome. Am J Hum Genet. 2018 Aug 2;103(2):232–44.

42. Saini S, Mitra I, Mousavi N, Fotsing SF, Gymrek M. A reference haplotype panel for genome-wide imputation of short tandem repeats [Internet]. bioRxiv. 2018 [cited 2018 Aug 30]. p. 277673. Available from: https://www.biorxiv.org/content/early/2018/07/24/277673

43. Hefferon TW, Groman JD, Yurk CE, Cutting GR. A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. Proc Natl Acad Sci U S A. 2004 Mar 9;101(10):3504–9.

44. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, Erlich Y. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016 Jan;48(1):22–9.

45. Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M. A polymorphic microsatellite that mediates induction of PIG3 by p53. Nat Genet. 2002 Mar;30(3):315–20.

46. Shimajiri S, Arima N, Tanimoto A, Murata Y, Hamada T, Wang K-Y, Sasaguri Y. Shortened microsatellite d (CA) 21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. FEBS Lett. 1999;455(1-2):70–4.

47. Quilez J, Guilmatre A, Garg P, Highnam G, Gymrek M, Erlich Y, Joshi RS, Mittelman D, Sharp AJ. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. Nucleic Acids Res. 2016 May 5;44(8):3750–62.

48. CAVIAR -- CAusal Variants Identication in Associated Regions [Internet]. [cited 2018 Aug 20]. Available from: http://genetics.cs.ucla.edu/caviar/

49. Borel C, Migliavacca E, Letourneau A, Gagnebin M, Béna F, Sailani MR, Dermitzakis ET, Sharp AJ, Antonarakis SE. Tandem repeat sequence variation as causative Cis-eQTLs for protein-coding gene expression variation: The case of CSTB. Hum Mutat. 2012;33(8):1302–9.

50. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015 May 8;348(6235):648–60.

CHAPTER 3: DEEP CHARACTERIZATION OF THE CONTRIBUTION OF SHORT TANDEN REPEATS ACROSS

3.1 Abstract

The human genome harbors multiple types of variants. Large-scale studies and traits focused studies have identified thousands of genetic variants, mostly Single Nucleotide Polymorphisms (SNPs) that contribute to gene expression variation in dozens of tissues and cell types. This has given ways to great interpretations of genetic variation and their role in number of diseases and traits. However, more complex variants such as repetitive sequences and longer size variants have been ignored due to bioinformatics limitations; keeping a hole in the understanding of the role of these other variants in many complex traits. Short Tandem Repeats (STRs) in particular, represent one of the most complex genetic variant classes. They are difficult to genotype from short reads and/or low coverage sequencing data. They are highly multi-allelic and often cast aside in past studies. STRs have been linked to dozens of diseases and shown to contribute gene expression variation. Amid multiple hypotheses on the role of STRs in regulating gene expression, characterizing these variants is becoming imperative.

The interpretation of the role of STRs in gene expression has so far faced several limitations including short read lengths and low coverage sequencing, which significantly decrease the quality of genotypes and thus the power to detect true associations. Moreover, regulatory elements have been proven to be tissue-specific or cell-specific at times; as a result, conclusions from such studies are limited and less likely to be used in different contexts. Understanding of STRs effects on gene regulation thus requires analysis in biologically relevant contexts.

Here, we looked at the impact of STRs on gene expression across 17 tissue types using high coverage, whole genome sequencing, and gene expression data for 650 samples from the Genotype-Tissue Expression (GTEx) project. We identified more than 25K expression STRs (eSTRs) that

affect 11,810 genes across 17 tissues. More than 32.5% of these eSTRs are shared by two or more tissues. We used statistical fine-mapping techniques to identify more than 3 thousand high-confidence fine-mapped expression STRs and showed that these fine-mapped eSTRs are highly enriched for features characteristic of regulatory regions. Characterization of fine-mapped signals reveals potential regulatory mechanism for eSTRs, namely G-quadruplex secondary structures formation. Finally, we identified almost a dozen GWAS signals for Schizophrenia, Autism Spectrum Disorder, and Height for which eSTRs were the most likely driving force of the association signal.

3.2 Introduction

Genome-Wide Association Studies (GWAS) and expression Quantitative Traits Loci (eQTL) studies have largely focused on bi-allelic variants, such as Single Nucleotide Polymorphisms (SNPs) or Structural Variants (SVs). However, these variants fail to explain the majority of heritability for most complex traits. STRs consist of short repeated motifs of 1-6bp in tandem and comprise more than 1.4 million loci genome-wide. They have been implicated in dozens of Mendelian disorders, including Huntington's Disease and Fragile X Syndrome [1–3]. Additionally, dozens of single-gene studies have shown that STRs may regulate gene expression through a variety of mechanisms [4]. However, due to their rapid mutation rates, STRs are often only in weak linkage disequilibrium with SNPs and thus are not captured by array-based studies. Additionally, due to the difficulty in accurately genotyping STRs from short reads, they have largely been filtered from studies based on next-generation sequencing.

We and others recently demonstrated that STRs contribute to a significant fraction of the heritability of gene expression [5]. However, these studies faced important limitations. Short read lengths and low coverage sequencing resulted in low-quality genotypes, reducing power to perform association studies. Additionally, the analysis was restricted to a single cell type, limiting the ability to derive meaningful biological insights and relevance to most complex traits.

The mechanism by which STRs control gene expression in human is not well understood. There is an abundance of STRs in promoter regions of gene in the human genome [6] and it has been hypothesized that STRs may influence gene expression through different mechanism [4] including (a) by forming transcription binding sites, as is the case with Gilbert's syndrome [7,8]; (b) by inducing unusual DNA secondary structures such as Z-DNA, a nucleosome formation blocker[9,10] or H-DNA triplex structure, a transcriptional activation element [10]; (c) by affecting spacing between regulatory elements [11,12]; (d) by modulating epigenetic properties such as DNA methylation or Heterochromatin-zation [4,13]; (e) by forming toxic RNA and protein aggregates .

A clear understanding of the role of STRs in gene regulation finds application in causal variant identification studies and the interpretation of (GWAS) results. In this study, we aimed to thoroughly characterize STRs with a significant contribution to gene expression variation (a.k.a. eSTR) in the human genome. We first identify eSTRs across 17 tissues using high coverage whole genome sequencing from 650 human samples and RNA-Seq from the Genotype-Tissue Expression (GTEx) project. We used fine mapping techniques to delineate between association signal as a result of tagging effect and eSTRs likely to be causal to gene expression variation. To understand the mechanism of transcriptional regulation, we performed enrichment analysis and interrogated eSTRs based on their localization and sequence composition. We reproduced examples in the literature and identified eSTRs likely to be drivers of association to traits from previous GWAS studies.

3.3 Results

3.3.1 Profiling expression STRs across 17 human tissues

We performed a genome-wide analysis to identify associations between STR repeat length and expression of nearby genes (expression STRs, or "eSTRs", which we use to refer to a unique STR by gene association). We focused on 652 samples included in the Genotype-Tissue Expression (GTEx) [14] dataset for which both high coverage Whole Genome Sequencing (WGS) and RNA-sequencing of multiple tissues were available. The WGS cohort consisted of 561 individuals with reported European ancestry, 75 of African ancestry, and 8, 3, and 5 of Asian, Amerindian, and Unknown ancestry, respectively. HipSTR[15] was used to obtain STR genotypes for each sample. Resulting genotype calls were subjected to stringent filtering (Methods). After filtering, 175,226 STRs remained for downstream analysis. To identify eSTRs, we performed a linear regression between average STR length and normalized gene expression for each STR within 100kb of a gene, controlling for sex, population structure, and technical covariates (Methods, Supplementary Figures S3.1.1, S3.1.2). The analysis was restricted to 17 tissues where we had data for at least 100 samples (Figure 3.1A, Table 3.1) and to genes with median RPKM greater than 0. As a control, for each STR-gene pair, we performed a permutation analysis in which sample IDs were shuffled. On average we tested 16,065 genes per tissue and altogether, an average of 278,521 STR by gene tests were performed for each tissue.

eSTR analysis identified 25,561 unique eSTRs across 11,810 genes in at least one tissue at a gene-level FDR of 10%. Of these, 32.5% (8,417) were shared by two or more tissues and 469 were shared by 10 or more tissues (Supplementary Figure S3.1.3), consistent with previous findings for SNP eQTLs [16]. P-values from the permuted controls followed a uniform distribution as expected under the null hypothesis of no association (Figure 3.1B). Tibial nerve had the most identified eSTRs (4,352) compared to the least in the two brain tissues (1,551 and 675 for Cerebellum and Caudate, respectively), as expected due to differences in sample sizes across tissues. Effect sizes were strongly correlated across tissues (Figure 3.1C), with related tissues, such as subcutaneous adipose vs. visceral adipose showing the strongest concordance.

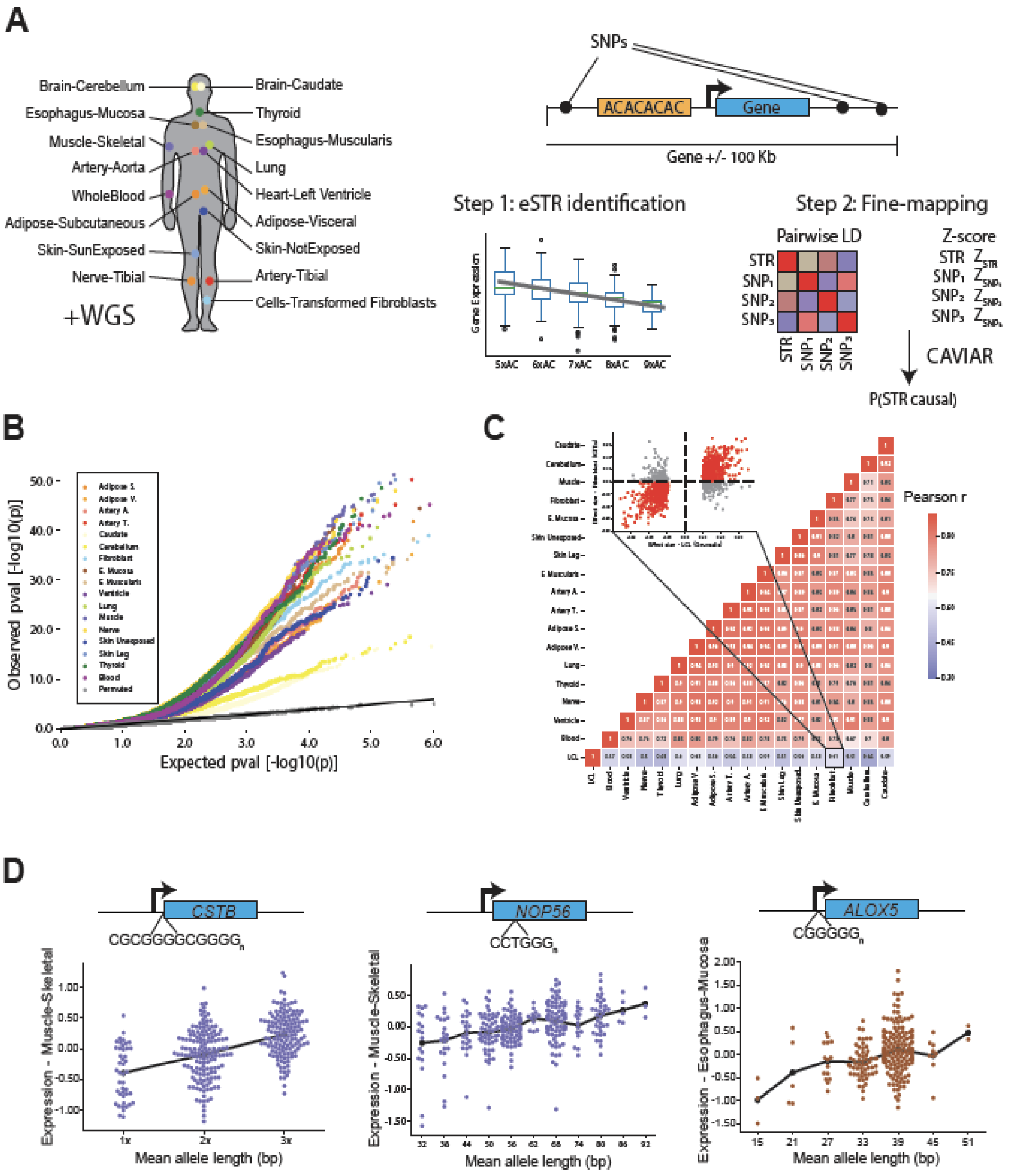
We evaluated our results by comparing to eSTRs we previously reported using an orthogonal cohort (gEUVADIS[17]) with lymphoblastoid cell line (LCL) expression data[18]. Effect sizes from LCLs were significantly correlated with effect sizes in all tissues ($p<0.01$ for all tissues, mean

Pearson r=0.55) (Figure 3.1C, inset), albeit to a lesser extent than for pairwise tissue comparisons within the GTEx cohort, with 74.4% of eSTRs showing the same direction of effect on average. We additionally tested previously reported eSTRs from single-gene studies in the GTEx cohort. Most of these were originally tested using in vitro constructs in cell lines rather than primary tissues, and thus may not recapitulate in vivo conditions. Still, six of eight examples had nominally significant eSTRs ($p<0.01$) in at least one tissue analyzed (Supplementary Table S3.1.1).

Table 3.1: Summary of cross-tissue eSTR identification.

Tissue	Sample size	# eSTRs (FDR>10%)	# Candidate causal eSTRs
Tibial Nerve	265	4,352	312
Thyroid	262	4,105	330
Sun-exposed Skin	297	3,827	298
Subcutaneous Adipose	270	3,587	292
Tibial Artery	276	3,454	269
Esophagus Mucosa	255	3,461	286
Skeletal Muscle	343	3,370	272
Transformed Fibroblasts	225	3,088	248
Lung	259	2,989	277
Esophagus Muscularis	214	2,824	256
Whole Blood	336	2,585	206
Aorta Artery	191	2,396	234
Non-exposed Skin	209	2,386	200
Visceral Adipose	193	1,840	175
Heart-Left Ventricle	199	1,731	212
Brain-Cerebellum	107	1,551	191
Brain-Caudate	108	675	103

Figure 3.1: Cross-tissue identification of eSTRs. A. eSTR discovery in 17 tissues. For each STR within 100kb of a gene, we tested for association between length of the STR and expression of the gene in 17 tissues profiled by GTEx. For each gene, CAVIAR was used to fine-map the effects of eSTRs vs. the top 100 *cis* SNPs on gene expression. CAVIAR takes as input pairwise variant LD and effect sizes (Z-scores) and outputs a posterior probability of causality for each variant. B. eSTR association results. The quantile-quantile plot compares observed p-values for each STR-gene test vs. the expected uniform distribution. Gray dots denote permutation controls. The black line gives the diagonal. C. eSTR effect sizes are correlated across tissues and studies. Each cell gives the Pearson correlation between each pair of tissues. The bottom row represents effect sizes obtained previously using the gEUVADIS LCLs in Gymrek *et al.*[18]. The inset compares effect sizes in LCLs to effect sizes for transformed fibroblasts obtained in that study. Red dots denote eSTRs with the same direction of effect in each case. D. Example putative causal eSTRs. For each plot, x-axis gives STR dosage as the average allele length relative to hg19 and the y-axis gives normalized expression. Each point represents a single individual. Black lines give the mean expression for each mean allele length. Gene diagrams are not drawn to scale.


A
Brain-Cerebellum
Esophagus-Mucosa
Muscle-Skeletal
Artery-Aorta
WholeBlood
Adipose-Subcutaneous
Skin-SunExposed
Nerve-Tibial
+WGS
Brain-Caudate
Thyroid
Esophagus-Muscularis
Lung
Heart-Left Ventricle
Adipose-Visceral
Skin-NotExposed
Artery-Tibial
Cells-Transformed Fibroblasts
SNPs
ACACACAC
Gene
Gene +/- 100 Kb
Step 1: eSTR identification
Gene Expression
5xAC
6xAC
7xAC
8xAC
9xAC
Step 2: Fine-mapping
Pairwise LD
STR
SNP1
SNP2
SNP3
Z-score
STR Z_{STR}
SNP1 Z_{SNP_1}
SNP2 Z_{SNP_2}
SNP3 Z_{SNP_3}
CAVIAR
P(STR causal)
B
Observed pval [-log10(p)]
Expected pval [-log10(p)]
0.0
10.0
20.0
30.0
40.0
50.0
1.0
2.0
3.0
4.0
5.0
6.0
Adipose S.
Adipose V.
Artery A.
Artery T.
Caudate
Cerebellum
Fibroblast
E. Mucosa
E. Muscularis
Ventricle
Lung
Muscle
Nerve
Skin Unexposed
Skin Leg
Thyroid
Blood
Permuted
C
Caudate
Cerebellum
Muscle
Fibroblast
E. Mucosa
Skin Unexposed
Skin Leg
E Muscularis
Artery A.
Artery T.
Adipose S.
Adipose V.
Lung
Thyroid
Nerve
Ventricle
Blood
LCL
Pearson r
0.90
0.75
0.60
0.45
0.30
D
CSTB
CGCGGGGCGGGG$_n$
Expression - Muscle-Skeletal
1.00
0.50
0.00
-0.50
-1.00
1x
2x
3x
Mean allele length (bp)
NOP56
CCTGGG$_n$
Expression - Muscle-Skeletal
0.50
0.00
-0.50
-1.00
-1.50
32
38
44
50
56
62
68
74
80
86
92
Mean allele length (bp)
ALOX5
CGGGGG$_n$
Expression - Esophagus-Mucosa
2.00
1.50
1.00
0.50
0.00
-0.50
-1.00
-1.50
15
21
27
33
39
45
51
Mean allele length (bp)

eSTRs identified above could potentially be explained by tagging nearby causal variants such as Single Nucleotide Polymorphisms (SNPs). We employed two fine-mapping approaches to identify candidate causal eSTRs unlikely to be simply tagging additional variants. First, we used the ANOVA model comparison to determine whether eSTRs explained additional variability in expression of the target gene beyond the best eSNP (Methods). On average across tissues, 27.9% of eSTRs improved the model (10% FDR), consistent with previous results obtained in LCLs (Gymrek et al. 2016) (23%). Next, to determine whether the eSTR could be simply tagging alternative SNPs besides the best eSNP, we employed CAVIAR [19], a statistical fine-mapping framework for identifying causal variants. CAVIAR models the relationship between local LD-structure and association scores to quantify the posterior probability of causality for each variant. We used CAVIAR to fine-map eSTRs against the top 100 associated SNPs and all STRs tested for each gene (Methods, Figure 3.1A Step2). On average across tissues, 14.3% of eSTRs had the highest causality scores of all variants tested.

We identified a group of 3,474 unique high-confidence eSTRs across 3,046 genes (Table 3.1, Supplementary Table S3.1.2) corresponding to the top 14% of eSTRs (Supplementary Figure S3.1.4) with CAVIAR posterior scores of at least 10%. We hereby refer to these STRs as “causal eSTRs”, with the caveat that further validation would need to be performed to definitely determine these to be causally related to gene expression. Multiple STRs with known disease implications were captured by this list. For example, a CG-rich repeat upstream of Cystatin B gene (CSTB) which has been previously implicated in myoclonus epilepsy[12] and as a causal eQTL for CSTB gene [20] was identified as an eSTR in 13 tissues (Figure 3.1D left, CAVIAR score >0.99 in seven tissues). In many cases, top causal eSTRs were highly multi-allelic loci showing clear linear trends between repeat length and expression. For example, an intronic GGCCTG repeat in the Nucleolar Protein 56 gene (NOP56) implicated in spinocerebellar ataxia 36[21,22] was identified as a causal eSTR in skeletal muscle and 5 other tissues and harbored 11 common alleles (Figure 3.1D middle). In another example, a CGGGGG repeat in the promoter of ALOX5 gene was identified as a causal eSTR (CAVIAR p=0.38) (Figure 3.1D right). This repeat was previously shown to regulate ALOX5 expression in leukocytes [23] and is associated with reduced lung function [23,24].

3.3.2 eSTRs are strongly enriched in predicted promoters and enhancers

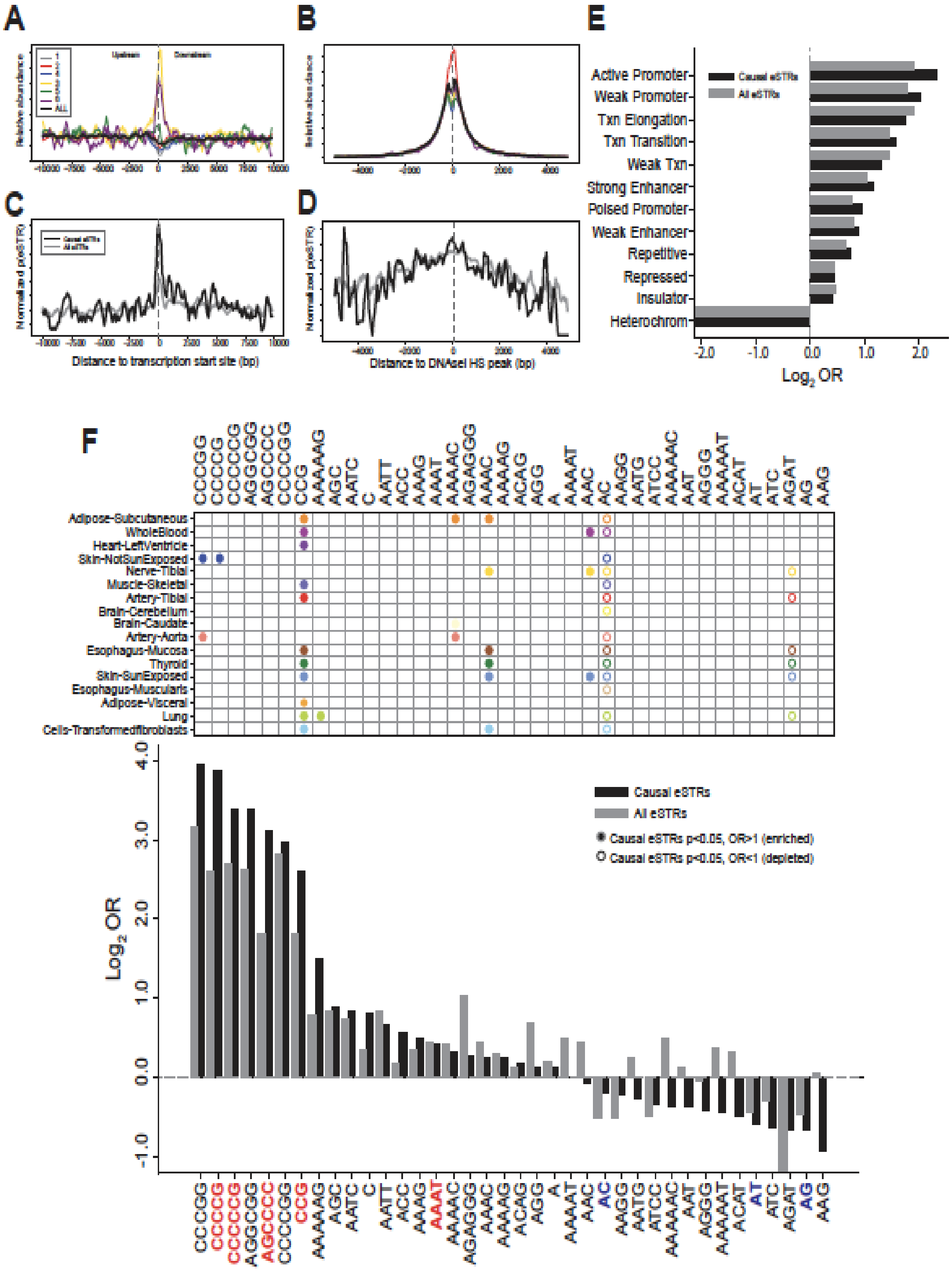
We next sought to characterize properties of eSTRs that might give insights into their biological function. As expected, the majority of eSTRs are located in non-coding regions (Supplementary Table S3.2.1) (54.6% and 32% in introns and intergenic regions, respectively), with only 75 eSTRs (16 causal) falling in coding regions (Supplementary Table S3.2.2). eSTRs were strongly enriched at 5' UTRs (OR=3.3; Fisher's two-sided p=1.9e-37), 3' UTRs (OR=3.1; Fisher's two-sided p=3.1e-77), and within 3kb upstream of transcription start sites (OR=3.2; Fisher's two-sided p<1e-300). These enrichments increased as a function of CAVIAR causality score (Supplementary Figure S3.2.2), suggesting that eSTRs act primarily through regulatory mechanisms and that CAVIAR posterior probabilities accurately identify the eSTRs most likely to be causal.

We examined the distribution of STRs in regulatory regions in more details by visualizing the localization of STRs around Transcription Start Sites (TSSs) and regulatory regions predicted based on DNAseI hypersensitive (DNAseI HS) sites identified by ENCODE[25]. Overall, STRs are depleted directly at TSSs (Figure 3.2A), although certain classes of motifs are highly enriched, including tri-, penta-, and hexa-nucleotide repeats as well as STRs with GC-rich motifs (Supplementary Figure S3.2.1). Strand-specific localization patterns in promoters, but not enhancer, regions showed striking differences for different STR motifs, with G-rich motifs enriched on the coding strand and A and T rich most prevalent on the coding strand upstream and downstream of the TSS, respectively (Supplementary Figure S3.2.1) consistent with previous observations [26]. STRs, particularly dinucleotides, are prevalent in enhancer regions, consistent with a similar finding in

Drosophila melanogaster [27] (Figure 3.2B). STRs closest to TSSs, and to a lesser extent near DNAseI HS peaks, were most likely to act as eSTRs (Figure 2C-D). Furthermore, eSTRs were strongly enriched in regions predicted by ChromHMM[28] to be regulatory or transcribed across multiple cell types (Figure 2E, Supplemental Figure S3.2.2, Supplementary Table 3.2.3) and in transcription factor binding sites profiled by ENCODE (Supplementary Table 3.2.4) (Methods). For example, eSTRs were enriched in active promoters (Fisher exact test odds ratio=3.75, two-sided p=4.4e-106), weak promoters (Fisher exact test odds ratio=3.48, two-sided p=4.7e-139), and strong enhancers (Fisher exact test odds ratio=2.07, two-sided p=4.7e-84). Enrichments in promoter, and to a lesser extent enhancer, regions strengthened when considering only causal eSTRs (Figure 3.2C-E, Annex Figure S3.2.2). For example, causal eSTRs were 5-fold enriched in active promoters (Fisher exact test two-sided p=4.7e-46) compared to 3.75-fold for all eSTRs (Supplementary Table 3.2.3). Taken together, these results suggest that CAVIAR posterior probabilities accurately identify the eSTRs most likely to be causal variants.

eSTRs were strongly enriched for repeats with CG-rich motifs across all tissues (Figure 3.2F, Supplementary Table 3.2.5, Methods), and these enrichments were again strengthened when restricting to causal eSTRs. For example, the motifs CCCGG, CCCCG, and CCCCCG were 15.5, 14.6, and 10.3-fold enriched, respectively, in causal eSTRs (Fisher's exact test two-sided p<10-3). Notably, while other motifs, including AAAT (Fisher exact test odds ratio=1.33, two-sided p=8.5e-5) and AAAAAG (Fisher exact test odds ratio=2.8, two-sided p=2.75e-3), few individual motifs showed tissue-specific enrichment in causal eSTRs (Figure 3.2F). Overall, eSTRs were equally likely to show increasing vs. decreasing trends between repeat copy number and expression (binomial p=0.27, 50.3% positive and p=0.31, 50.9% positive for all eSTRs and causal eSTRs, respectively). When restricting to motifs enriched in causal eSTRs (Bonferonni corrected p<0.05), eSTRs were enriched for positive (increasing) effect sizes (p=0.010, 61% positive).

Figure 3.2: Characterization of eSTRs in predicted regulatory regionsA. Density of STRs around transcription start sites. The y-axis give the relative number of STRs in each 100bp bin around the TSS. Negative numbers denote upstream regions and positive numbers denote downstream regions. B. Density of STRs around DNAseI HS sites. Plots are centered at ENCODE DNAseI HS clusters and give the relative number of STRs in each 50bp bin. For A. and B. black line denotes all STRs and colored lines denote repeats with different motif lengths (gray=homopolymers, red=dinucleotides, gold=trinucleotides, blue=tetranucleotides, green=pentanucleotides, purple=hexanucleotides). C. Relative probability to be an eSTR around TSSs. The y-axis gives the probability of an STR in each bin to be an eSTR, normalized by the genome-wide average probability to be an eSTR. The gray line gives the probability of an STR in each bin to be a causal eSTR. D. Relative probability to be an eSTR around DNAseI HS sites. For A-D, values were smoothed by taking a sliding average of each four consecutive bins. E. Enrichment of eSTRs in ChromHMM states. Bars give the $\log_2$ odds ratio from performing Fisher's exact test comparing eSTRs to all STRs (gray bars) or all causal eSTRs to all STRs (black bars). Enrichment p-values are given in Supplementary Table CHROMHMMENRICH. F. Motif enrichment at eSTRs. The x-axis gives all motifs for which there were at least 3 causal eSTRs across all tissues. The y-axis of the bottom plot gives $\log_2$ odds ratios from performing Fisher's exact test comparing eSTRs to all STRs (gray bars) or all causal eSTRs to all STRs (black bars). Bolded motifs indicate motifs that were enriched (red) or depleted (blue) across all causal eSTRs (Bonferonni corrected $p<0.05$). The top panel denotes motifs that were enriched (closed circles) or depleted (open circles) in causal eSTRs in each tissue (Bonferonni corrected $p<0.05$).


A
Relative abundance
Upstream
Downstream
ALL
-10000
-7500
-5000
-2500
0
2500
5000
7500
10000
B
Relative abundance
-4000
-2000
0
2000
4000
C
Normalized p(eSTR)
Causal eSTRs
All eSTRs
Distance to transcription start site (bp)
D
Normalized p(eSTR)
Distance to DNAseI HS peak (bp)
E
Causal eSTRs
All eSTRs
Active Promoter
Weak Promoter
Txn Elongation
Txn Transition
Weak Txn
Strong Enhancer
Poised Promoter
Weak Enhancer
Repetitive
Repressed
Insulator
Heterochrom
-2.0
-1.0
0.0
1.0
2.0
Log2 OR
F
CCCGG
CCCCG
CCCCCG
AGGCGG
AGCCCC
CCCCGG
CCG
AAAAAG
AGC
AATC
C
AATT
ACC
AAAG
AAAT
AAAAC
AGAGGG
AAAC
AAAAG
ACAG
AGG
A
AAAAT
AAC
AC
AAGG
AATG
ATCC
AAAAAC
AAT
AGGG
AAAAAT
ACAT
AT
ATC
AGAT
AG
AAG
Adipose-Subcutaneous
WholeBlood
Heart-LeftVentricle
Skin-NotSunExposed
Nerve-Tibial
Muscle-Skeletal
Artery-Tibial
Brain-Cerebellum
Brain-Caudate
Artery-Aorta
Esophagus-Mucosa
Thyroid
Skin-SunExposed
Esophagus-Muscularis
Adipose-Visceral
Lung
Cells-Transformedfibroblasts
Log2 OR
4.0
3.0
2.0
1.0
0.0
-1.0
Causal eSTRs
All eSTRs
● Causal eSTRs p<0.05, OR>1 (enriched)
○ Causal eSTRs p<0.05, OR<1 (depleted)

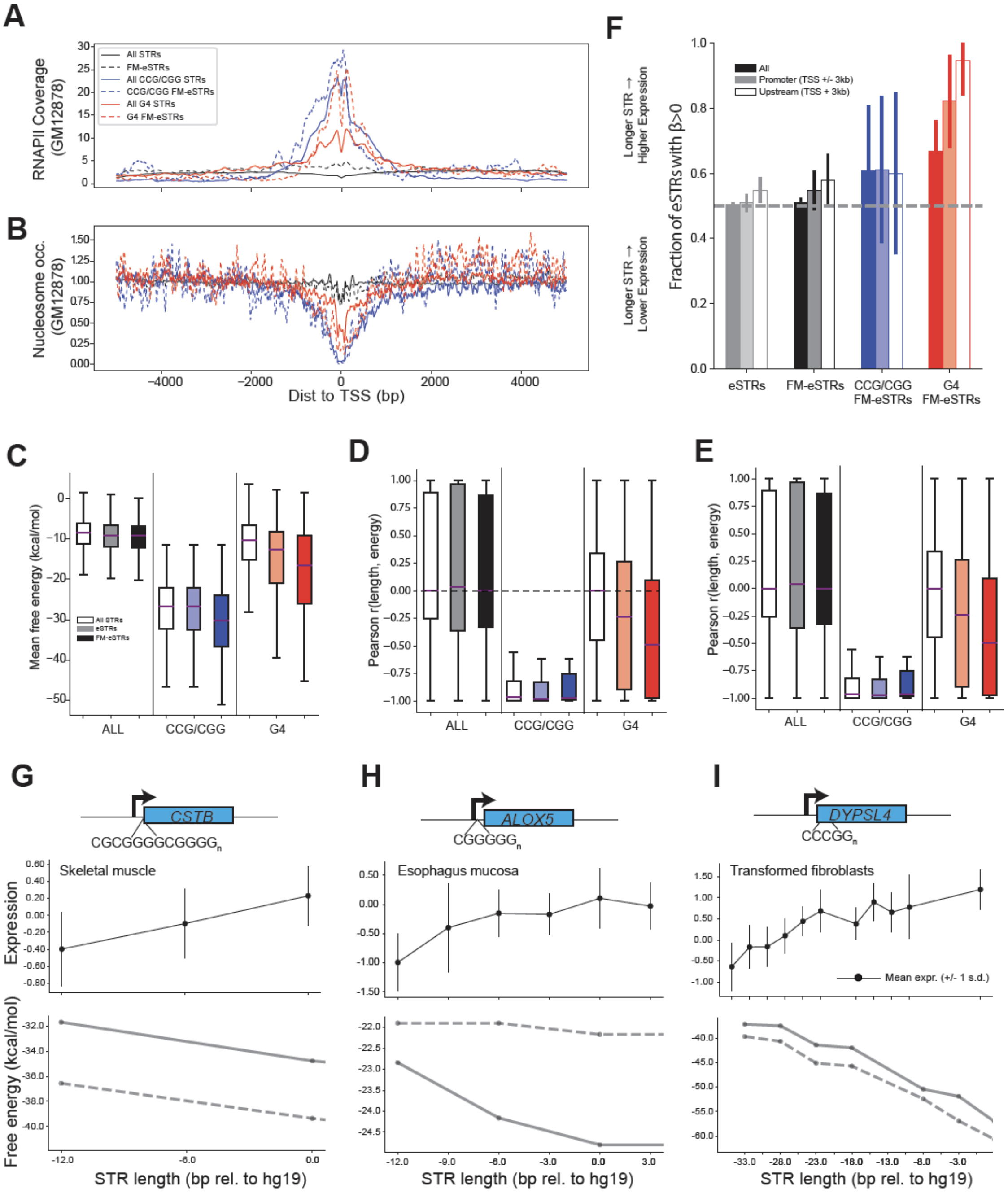
3.3.3 CG-rich eSTRs in promoters enhance transcription through secondary structures formation

The most robust eSTR signals identified above consisted of GC-rich repeats located in promoter regions. Nearly all of the motifs (e.g. CCCCGn, CCCCCGn, AGCCCCn) most strongly enriched in causal or well fine-mapped and herein (FM) eSTRs form sequences with the ability to form G4 quadruplexes[29]. These are DNA secondary structural features characterized by G-rich sequences may regulate gene expression through a variety of mechanisms. G4 motifs may act as nucleosome exclusion signals [29–31]. Enhanced G4 formation in promoter and 5'UTR regions was recently shown to increase transcriptional activity of nearby genes [30]. Furthermore, G4 motifs have been shown to correlate with promoter-proximal transcriptional pausing downstream of TSSs [32] and form highly stable structures with low free energy[33] to promote gene expression. We thus hypothesized that the effects of CG-rich eSTRs may be in part due to formation of non-canonical nucleic acid secondary structures that modulate DNA or RNA stability as a function of repeat number. Based on previous findings, we expected that longer STR alleles at these repeats would increase DNA stability and in turn result in lower nucleosome occupancy and higher transcription. We considered two classes of eSTRs: those following the standard G4 motif ($G_3N_{1-7}G_3N_{1-7}G_3N_{1-7}G_3$) (Supplementary Table S3.3.1) and CCG/CGG repeats, which are abundant in 5'UTR regions and may form hairpins or other secondary structures in single-stranded RNAor DNA, but do not meet the standard G4 definition. Both classes of GC-rich repeats were associated with higher RNAPII (Figure 3.3A) and lower nucleosome occupancy (Figure 3.3B) compared to all STRs, with more extreme differences when considering only causal eSTRs.

Next, we used Mfold [35] to calculate the free energy of each STR and 50bp of its surrounding context on either the template or non-template strand of DNA or RNA (Methods). When considering template-strand DNA sequences, both G4 STRs, and to a greater extent CCG repeats, were associated with overall lower free energy (greater stability) compared to all STRs in promoter regions (Kolomogorov-Smirnov [KS] two-sided p=8.3e-140 and p=3.5e-54 for CCG and G4 repeats, respectively; Figure 3.3C). Overall, FM eSTRs were fell in regions with significantly lower free energy compared to all STRs (KS two-sided p=4.6e-17, 1.1 fold decrease compared to all STRs), but this effect was strongest for G4 STRs (KS two-sided p=6.3e-8, 1.6 fold decrease compared to all G4). We next tested whether modulating the number of repeats at each STR affected the predicted DNA stability. Overall, FM eSTRs were about equally likely to increase or decrease free energy (47% with negative effects, defined as Pearson r<0), whereas 100% of CCG and 70% of G4 FM eSTRs showed negative correlations between free energy and repeat length (Figure 3.3D), significantly more than the 50% expected by chance (binomial two-sided p=4.8e-7 and p=5.3e-4 for CCG and G4, respectively). The magnitude of the change in energy across common allele lengths was strongest for GC-rich FM eSTRs (Figure 3.3E). Similar trends were observed when considering structures formed on the non-template DNA strand or in transcribed RNA. Overall, these results suggest that longer GC-rich STRs result in more stable secondary structures formed in both DNA and RNA during transcription.

Longer GC-rich repeats were associated with increased expression. Overall, FM eSTRs were equally likely to show increasing vs. decreasing trends between repeat copy number and expression (binomial two-sided p=0.33, n=3,474, 50.9% positive effect sizes). Similarly, CCG repeats did not show a strong bias (binomial two-sided p=0.40, n=23, 60.1% positive) although they had a higher prevalence of positive effects sizes. On the other hand, 67% of G4-forming FM eSTRs had positive effects, significantly more than the 50% expected by chance (binomial two-sided p=0.0017, n=93) (Figure 3F). The effect direction bias was more pronounced for G4 FM eSTRs within 3kb of TSSs (82% positive, binomial two-sided p=0.0015, n=28). When restricting to regions downstream of TSSs, 17/18 (94%) showed positive effect sizes. In many cases changes in expression levels across allele lengths followed an inverse relationship with free energy levels (Figure 3G-I). These results support a model in which longer repeat tracts at promoter G4 STRs form more stable DNA secondary structures which promote transcription, consistent with genome-wide findings of Hansel-Hertsch [30], which found that enhanced G4 formation resulted in increased transcription of nearby genes.

Figure 3.3: CG-rich promoter eSTRs form transcription-inducing DNA secondary structures.
Here, candidate causal or well fine-mapped eSTRs are named “FM eSTRs”. G4 can either inhibit transcription or enhance transcription depending on the strand in which it forms, and the stability of the secondary structure formed. A. Density of POLR2 around STRs - B. Nucleosome occupancy around the transcription start site TSS. In each category, hard lines are all STRs and in dashed lines are FM eSTRs. The plots are centered at the center of STRs. The black lines denote STRs/ and FM eSTRs without distinction of motifs categories. In red are STRs and FM eSTRs with G4 canonical form and blue lines represent STRs and FM eSTRs. Solid lines are STRs without distinction of motifs and dashed lines represent STRs with potential of forming G-quadruplex.) C. FM eSTRs with G-quadruplex potential have lower free energy, especially G-quadruplex of CCG/CGG form. The Y=axis gives the mean free energy. D,E. Correlation between STRs length and free energy. G4 eSTRs free energy and G4 FM eSTRs free energy are negatively correlated with eSTRs length. Free energy was calculated with Mfold and the mean distribution is represented. In each category, Black and grey= all STRs, eSTRs and FM eSTRs. Blue and red are all STRs, and FM eSTRs with G4 potential with red for canonical forms of G4. F. G4 eSTRs and G4 FM eSTRs upstream of TSS drive expression. Fraction of eSTRs for which the calculated effect size is positive in each category of STRs (y-asxis). G-I. Example of causal eSTRs with G4 potential previously linked to diseases. For each example, the top plot represents the expression as a function of STRs length. The bottom plot is the free energy as a function of STR length. In hard line are the Template strand and in dashed line are the energy of reverse strand (non-template strand)


A
RNAPII Coverage (GM12878)
30
25
20
15
10
5
0
All STRs
FM-eSTRs
All CCG/CGG STRs
CCG/CGG FM-eSTRs
All G4 STRs
G4 FM-eSTRs
B
Nucleosome occ. (GM12878)
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−4000
−2000
0
2000
4000
Dist to TSS (bp)
F
Longer STR → Higher Expression
Longer STR → Lower Expression
Fraction of eSTRs with β>0
1.0
0.8
0.6
0.4
0.2
0.0
All
Promoter (TSS +/- 3kb)
Upstream (TSS + 3kb)
eSTRs
FM-eSTRs
CCG/CGG FM-eSTRs
G4 FM-eSTRs
C
Mean free energy (kcal/mol)
0
−10
−20
−30
−40
−50
All STRs
eSTRs
FM-eSTRs
ALL
CCG/CGG
G4
D
Pearson r(length, energy)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
ALL
CCG/CGG
G4
E
Pearson r(length, energy)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
ALL
CCG/CGG
G4
G
CSTB
CGCGGGGCGGGG$_n$
Skeletal muscle
Expression
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
-0.80
Free energy (kcal/mol)
-32.0
-34.0
-36.0
-38.0
-40.0
-12.0
-6.0
0.0
STR length (bp rel. to hg19)
H
ALOX5
CGGGGG$_n$
Esophagus mucosa
0.50
0.00
-0.50
-1.00
-1.50
-22.0
-22.5
-23.0
-23.5
-24.0
-24.5
-12.0
-9.0
-6.0
-3.0
0.0
3.0
STR length (bp rel. to hg19)
I
DYPSL4
CCCGG$_n$
Transformed fibroblasts
1.50
1.00
0.50
0.00
-0.50
-1.00
Mean expr. (+/- 1 s.d.)
-40.0
-45.0
-50.0
-55.0
-60.0
-33.0
-28.0
-23.0
-18.0
-13.0
-8.0
-3.0
STR length (bp rel. to hg19)

3.3.4 eSTRs contribute to GWAs hits

Finally, we investigated the contribution of STRs to trait-associated loci. The vast majority of GWAS studies have been performed using only SNPs or small indels, and do not assay STRs directly. Many highly polymorphic STRs are not well tagged by SNPs and thus are unlikely to be detected by standard SNP-based GWAS. On the other hand, STRs are reputable for being in low linkage with SNPs because their high variability [38], but we reasoned that GWAS may detect a subset of signals driven by causal eSTRs if they are in high LD with common SNPs, although with attenuated effects compared to if the STR was tested directly for association.

We downloaded the GWAS catalog V1.0.1 with hg19 reference coordinates [36] and the catalog of SNPs associated with two psychiatric disorders: Schizophrenia and the Autism Spectrum Disorder (ASD) from the Psychiatric Genomic Consortium [37]. We identified 2,564 FM eSTRs within 50kb of GWAS loci. Of these, 1352 were in moderate LD (r^2>0.1) and181 were in strong LD (r^2>0.8) with the lead SNP. For 11 loci in at least moderate LD, the lead GWAS variant was actually within the STR but was annotated as a bi-allelic indel in dbSNP. For example, the lead variant (rs369552432; -/CGGCGGCGG) at a locus associated with hemoglobin identified by Astle, *et al* [64] falls within a multi-allelic 5'UTR trinucleotide CCG repeat negatively associated with expression of *VLDLR* in whole blood (Figure 3.4A). In another example, the lead variant (rs10709981 -/A) at a locus associated with red blood cell count by the same study falls within a homopolymer A repeat positively associated with expression of *SLC36A1* in 15 tissues (nominal p<0.01). Notably, these examples were predominantly from several studies of blood-related traits [59,60], which is likely explained by the fact that these studies did not explicitly filter STR regions, rather than by a bias for STRs to affect blood phenotypes. We used coloc [39] to determine whether expression and these traits likely share a common causal variant in each region based on comparison of SNP summary statistics from eQTL vs. GWAS associations. A model in which both signals are driven by the same variant had posterior probability of 99.9% for *VLDLR* and 97.7% for *SLC36A1*. eSTR for hemoglobin scoring more that 99% chance for being the causal variant based on coloc.

We next determined whether FM (causal) eSTRs colocalize with genome-wide significant associations identified for schizophrenia [61] and height[62], two example traits with large available sample sizes and publicly available summary statistics (Methods). For each trait, we identified GWAS associations for which (1) the lead SNP was in significant LD ($p<0.05$) with an FM eSTR, (2) the FM eSTR was the most probable causal variant for expression identified by CAVIAR analysis (3) coloc analysis indicated that a model where both expression and the trait are explained by the same causal variant is most probable (posterior probability ≥ 0.50). A total of 7 and 10 such loci were identified for height and schizophrenia, respectively (Table 3.2). For example, a GWAS signal for height (lead SNP rs2336725) and expression of *RFT1* had a 99% posterior probability of colocalization. The lead SNP was in high LD ($r^2=0.85$) with an AC repeat identified as an FM eSTR in heart (CAVIAR p=0.39) and aortic artery (CAVIAR p=0.70) and as a nominally significant eSTR in 17 tissues for *RFT1*, a gene involved in N-glycosylation of proteins (Figure 4B). The STR falls in a cluster of transcription factor binding sites identified by ENCODE near the 3’ end of the gene and exhibits a positive correlation with expression across a range of allele lengths. In a similar example, a GWAS signal for schizophrenia (lead SNP rs9420) and expression of *MED19* had a 90.1% posterior probability of colocalization. The lead SNP was in high LD ($r^2=0.68$) with an AC STR identified as an FM eSTR in subcutaneous adipose (CAVIAR p=0.47) and nominally significant eSTR in 14 tissues for *MED19* which is a component of the mediator complex, and plays a role in silencing of neuronal gene expression [63]. The repeat is located in an intron of *CTNND1* approximately 43kb upstream of *MED19.* Furthermore, a recently developed burden test considering promoter-enhancer identified a significant association between *MED19* expression and schizophrenia risk[70]. Again, the STR showed a linear positive association with expression across a range of allele lengths, strongly suggestive of a causal relationship between STR length and expression. While overall our analysis identified high confidence FM eSTRs potentially underlying dozens of GWAS signals, each association will ultimately have to be verified by testing for association directly with each target phenotype and performing additional experiments to determine the underlying causal variants.

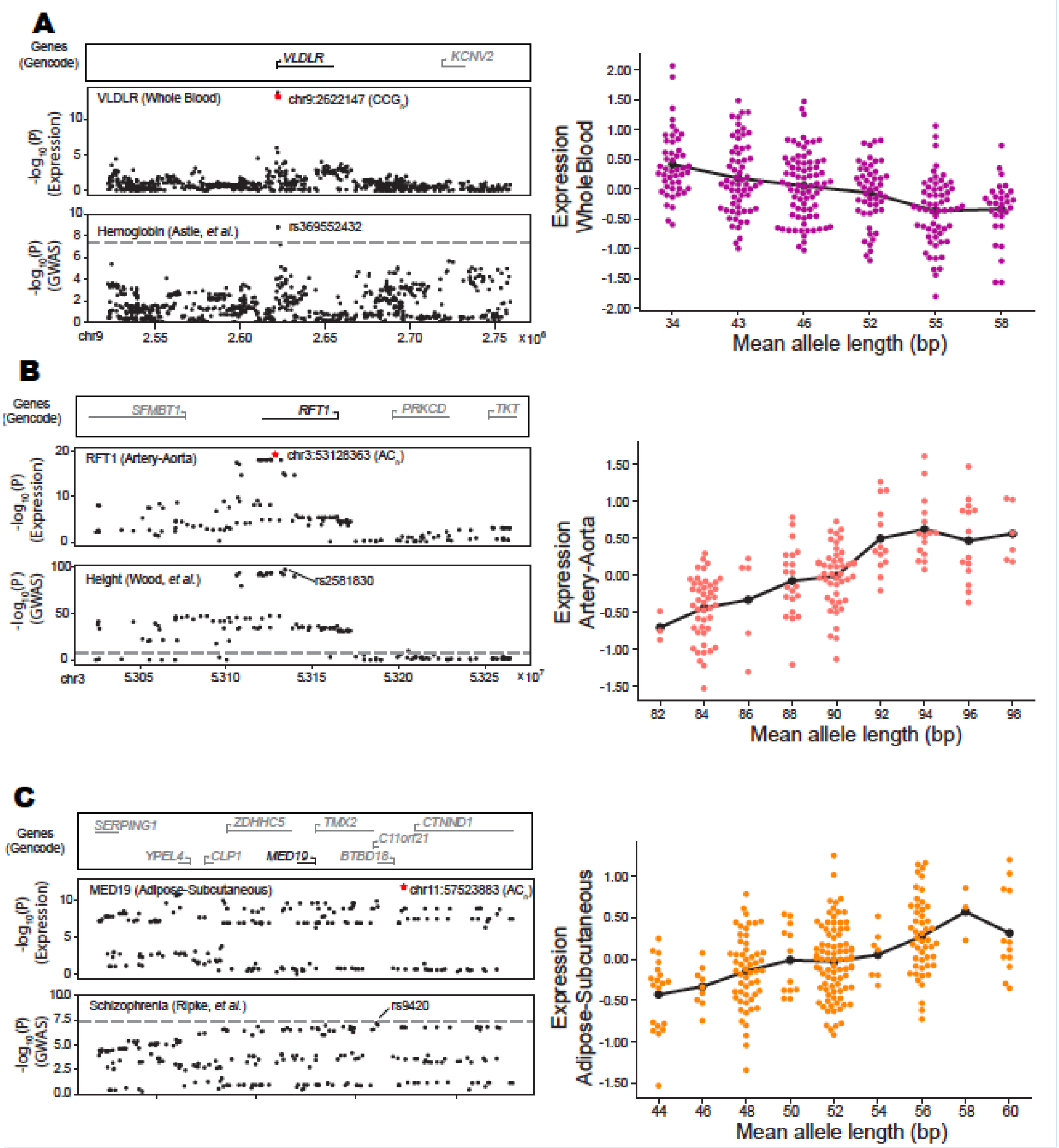


Figure 3.4: Examples of eSTRs more like to contribute to GWAS hits for traits. For each example, the left plot represents: Top rectangle segment location of expressed gene. Middle rectangle is the log10(association pvalue) in GTEx data. Bottom rectangle is log10(pvalue of association) of variant at each location for variant from the GWAS study. For all 3 plots, the shared x-axis represents the chromosome position. The right plot represents the expression as a function of STRs mean allele length in best tissue from GTEx data. A is eSNP and eSTRs for VLDLR and GWAS hit for Height, B is eSNP and best eSTR for RFT1 gene, GWAS hits for Height, and C is eSNP and eSTRs for MED19 gene and GWAS hit for schizophrenia.

Table3.2: FM eSTR that were the most probable causal variant for GWAS hits

Trait	**Gene**	**STR (hg19)**	**Motif**	**Lead SNP**	**SNP-STR LD**	**Top CAVIAR score**	**Coloc**
Height	*RFT1*	3:53128363	AC	rs2336725	0.85	0.70 (Artery-Aorta)	99.0%
Height	*FADS1*	11:61620629	AAC	rs174547	0.56	0.32 (Brain-Cerebellum)	97.7%
Height	*LUZP1*	1:23553614	AAC	rs1738475	0.22	0.48 (Adipose-Subcutaneous)	93.1%
Height	*PTPRCAP*	11:67238583	AAT	rs12795957	0.04	0.11 (Artery-Aorta)	90.5%
Height	*UBE2Z*	17:46965665	AGAT	rs318095	0.37	0.13 (Muscle-Skeletal)	79.0%
Height	*SLCO1C1*	12:20856340	AAC	rs10770705	0.79	0.16 (Skin-NotSunExposed)	76.7%
Height	*ADAMTS3*	4:73471388	AG	rs16848425	0.0058	0.16 (Cells-TransformedFibroblasts)	58.1%
Schizophrenia	*PRR12*	19:50110619	AC	rs56873913	0.14	0.99 (Thyroid)	98.2%
Schizophrenia	*MED19*	11:57523883	AC	rs9420	0.63	0.47 (Adipose-Subcutaneous)	90.1%
Schizophrenia	*CEBPZ*	2:37526655	AC	rs2372993	0.74	0.28 (Lung)	86.7%
Schizophrenia	*FAM134A*	2:220082366	A	rs6707588	0.01	0.12 (Esophagus-Mucosa)	79.8%

Table3.2: FM eSTR that were the most probable causal variant for GWAS hits (continued)

Trait	**Gene**	**STR (hg19)**	**Motif**	**Lead SNP**	**SNP-STR LD**	**Top CAVIAR score**	**Coloc**
Schizophrenia	*TM6SF2*	19:19424949	AAAT	rs2905424	0.82	0.11 (Adipose-Visceral)	77.3%
Schizophrenia	*TMEM81*	1:205076171	AC	rs16937	0.43	0.13 (Nerve-Tibial)	63.3%
Schizophrenia	*CCDC126*	7:23615272	A	rs227932	0.80	0.12 (Thyroid)	60.2%
Schizophrenia	*SPIRE2*	16:89829656	A	rs12449000	0.69	0.69 (Skin-NotSunExposed)	58.7%
Schizophrenia	*VARS2*	6:30884231	AC	rs111782145	0.28	0.12 (Brain-Cerebellum)	54.1%
Schizophrenia	*UBAP2L*	1:154251259	A	rs7521837	0.95	0.11 (Esophagus-Muscularis)	54.1%

3.4 Discussion

In this study, we leveraged high coverage, whole genome sequencing data from 650 samples, and RNA-seq to create a new comprehensive resource of cis-eQTLs containing more than 25 thousand eSTRs affecting nearly 12 thousand genes in 17 tissues, including more than 11 thousand mononucleotide repeats. 3,474 of these eSTRs were potentially causal and thus, less likely to be due to a tagging effect. Although further experiments will be needed to validate the causal relationship, this constitutes the largest to date eSTRs resource that can be useful for sensible candidates screening for disease-causing variants. This map will improve the interpretation and understanding of the biological implication of these variants on risk loci discovered in association studies. For example, eSTRs in literature and previously associated with diseases were largely replicated here. LD analysis showed more than a dozen eSTRs in high LD with GWAS hits, even though STRs are reputable for being in low LD with SNPs. More than half of causal eSTRs in LD with GWAS hits were more likely to be the main driver of the association signals, by the probability of causality estimation.

Our results also suggested that one mechanism by which eSTRs may be regulating gene expression is by forming G-quadruplexes that interact with RNA polymerase pausing and thus enhancing transcription and translation. These conclusions come after testing 3 expected factors required for such assertion, as described in vivo experiments from earlier documentation of the role of G4 secondary structures [31]. We expected G4-eSTRs to not only drive the gene expression up, but also have low free energy, and moreover, be in close proximity to RNA polymerase. All these hypotheses were proven true in this study. While the number of these STRs were low for inferring tissue specificity of this role, we envision this last evidence to be investigated in future studies.

Several findings in this study suggested that there may be other regulatory mechanisms by which eSTRs may affect gene transcription. For example, mononucleotide STRs were the most abundant STRs genotyped and thus had the greatest number of eSTRs identified. Their abundance around the TSS strengthened the notion that these repeats have a regulatory function as previously predicted. The density of poly-A upstream of TSS of affected genes suggested that these mononucleotides may be forming transcription binding sites. This was consistent with previous studies reporting on their abundance around TSS regions of the gene and their role in other organisms. In addition, we observed a significant depletion of eSTR "AC" motif across all tissues, which may be telling of the sensitive role they may be having in promoter regions when present.

This study had some limitations. First, we only studied the linear relationship between STR length and gene expression, ignoring in-sequence variation and alleles. This is because the majority of STR-related diseases have been linked to STR length variation, especially fluctuation by step size increment [40,41]. However, the nucleotide-level variation of STRs has been linked to major phenotypes in the past [41]. Future study may develop appropriate algorithms and methods to account for these types of variation. Second, we only explored cis-association in this study and did not consider the possible effect of STRs in the distant (trans) location from the affected genes. Third, for each tissue, we only considered one strong eSTRs per gene, but it is possible for a gene to be affected by more than one eSTR. Here, we provided statistical estimates for all STRs tested for a gene in supplement files. This will enable reproducibility and more analysis to uncover additional eSTRs. Overall, because of all these limitations, the number of eSTRs and causal eSTRs found in this study may be underestimated.

However, this study has paved the way for future inclusion of STRs in large-scale studies. As research advances in human genetics, it is now possible to integrate short tandem repeats as part of variants screening into clinical practice and cohort studies.

3.5 Material and Methods

3.5.1 Dataset and preprocessing

Next-generation sequencing data was obtained from the Genotype-Tissue Expression (GTEx) through dbGaP under phs000424.v7.p2. This included high coverage (30x) Illumina whole genome sequencing (WGS) data and expression data from 650 unrelated samples with diverse ethnic background (Supplementary Figure 3.5.1). For each sample, we downloaded BAM files containing read alignments to the hg19 reference genome and VCFs containing SNP genotype calls.

STRs were genotyped using HipSTR [42]. Samples were genotyped separately due to computational constraints required for joint calling. VCFs were filtered using the filter_vcf.py script available from HipSTR using recommended settings for high coverage data (min-call-qual 0.9, max-call-flank-indel 0.15, and max-cal-stutter 0.15). VCFs were merged across all samples and further filtered to exclude STRs meeting the following criteria: call rate < 80%; STRs overlapping segmental duplications (UCSC Genome Browser[43] hg19.genomicSuperDups table); penta- and hexamer loci containing long homopolymer runs; and loci whose frequencies did not meet expectation from Hardy-Weinberg Equilibrium ($p<0.05$) as described previously [44]. Additionally, to restrict to polymorphic STRs we filtered loci with heterozygosity < 0.1. Altogether, 175,226 STRs remained for downstream analysis. We additionally obtained gene-level RPKM values for each tissue. We focused on 15 tissues with at least 100 samples which included two brain tissues (Table 3.1). Genes with median expression level of 0 were excluded and expression values for remaining genes were quantile normalized to a standard normal distribution.

3.5.2 eSTR and eSNP identification

For each STR within 100kb upstream and downstream of a gene, we performed a linear regression between STR lengths and expression values: $Y = \beta X + C + \varepsilon$ where X denotes STR genotype lengths, Y denotes expression values, β denotes the effect size, C denotes various covariates, and ε is the error term. Following our previous study [45], we used "STR dosage," defined as the sum of repeat lengths of the two alleles for each sample, to define STR genotypes. All repeat lengths were reported as length difference from the hg19 reference, with 0 representing the reference allele. STR dosages were scaled to have mean 0 and variance 1.

We included sex, population structure, and technical variation in expression as covariates. For population structure, we used the top 15 principal components resulting from perform principal components analysis on the matrix of SNP genotypes from each sample. To control for technical variation in expression, we applied the PEER factor correction [14,46]. As suggested for the PEER method, we used N/4 PEER factors as covariates for each tissue, where N was the sample size. PEER factors altogether accounted for more than 75% of variation in gene expression and were correlated with covariates reported previously for GTEx samples (**Annex Figure 3.2**). We used a gene-level FDR threshold (described previously [45]) of 10% to identify significant STR by gene pairs. eSNPs were identified using the same model and covariates, but using SNP dosages (0, 1, or 2) rather than STR dosage.

3.5.3 Fine-mapping eSTRs

We used model comparison to determine whether the best eSTR for each gene explained variation in gene expression beyond a model consisting of the best eSNP. As described previously[45], for each gene with an eSTR we determined the lead eSNP with the strongest p-value. We then compared two linear models: Y~eSNP (SNP-only model) vs. Y~eSNP+eSTR (SNP+STR model) using the anova_lm function in the python statsmodels.api.stats module. Q-values were obtained using the qvalue package in R [47]. We used CAVIAR v1.0 to further fine-map eSTR signals against the top 100 eSNPs within 100kb of each gene. Pairwise-LD between the eSTR and eSNPs was estimated using the Pearson correlation between SNP dosages (0, 1, or 2) and STR dosages (sum of the two repeat allele lengths).

3.5.4 Enrichment analyses

Enrichments were performed using a two-sided Fisher's exact test as implemented in the python scipy.stats package. Annotated genomic regions were downloaded from the following: Hg19 gene annotation was downloaded from ENCODE V17 and included all genic locations: TSS, TES, introns, exon, and CDS. Chromatin state annotations computed by the ENCODE Project [25] using ChromHMM [48] for GM12878 were downloaded from the UCSC Genome Browser [43] (table hg19. wgEncodeBroadHmmGm12878HMM). Histone modifications from ENCODE were downloaded from the hg19.wgEncodeBroadHistone track. Transcription factor binding sites profiled by the ENCODE Project were downloaded both in the GM12878 lymphoblastoid cell line and Lung tissues. The latter targeted the POLR2A transcription factor.

3.5.5 Biological function of eSTRs

From the GEO repository, accession number GSE7668829 [30], we downloaded G4-Chip-seq from normal skin cells and skin cells treated with Entinostat to induce a stable active chromatin state. We performed enrichment analysis across 17 tissues using Fisher's exact test. To evaluate the potential of G4-eSTRs for regulating gene expression, we used three steps: First, we calculated the density of RNA polymerase around all STRs in the study and restricted to those with potential for G4 secondary structure, and then we compared the overall density around eSTRs, causal eSTRs. Second, we evaluated the percentage of eSTRs with positive direction of effect as well as causal then compared by subgroups (all eSTRs, all causal, and all G4). Finally, we evaluated the stability of these potential secondary structures by calculating the free energies of each for all STRs allele and surrounding context of 50 bps using mfold [34,35], then compared by subgroups (all STRs and all G4 STRs).

3.5.6 The contribution of eSTRs in traits

We compiled a set of SNPs associated with phenotypes from the following sources: we downloaded the GWAS catalog V1.0.1 with hg19 reference coordinates [36]. We obtained the catalog for GWAS hits for Schizophrenia and the Autism Spectrum Disorder (ASD) from the Psychiatric Genomic Consortium [37]. From the compiled set, we selected SNP markers for which genotypes were available in the GTEx data, which led to a set of 2463 GWAS SNPs markers. For each SNP marker, we identified all eSTRs within 50kb of its position and, for each pair (eSNP - eSTR), and we calculated linkage disequilibrium between the two using a the pearsonr function from the scipy.stats python package, recording r-square and significance value (p-value). Causal eSTRs (causal score>10%) in high LD with the GWAS hit (r2>0.5) were considered the likely drivers of GWAS hits.

References

1. Sutherland GR, Richards RI. Simple tandem DNA repeats and human genetic disease. Proc Natl Acad Sci U S A. 1995 Apr 25;92(9):3636–41.

2. Richards RI, Holman K, Yu S, Sutherland GR. Fragile X syndrome unstable element, p (CCG) n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. Hum Mol Genet. 1993;2(9):1429–35.

3. Piccolo G, Cortese A, Tavazzi E, Piccolo L, Sassone J, Ciammola A, Alfonsi E, Colombo I, Moggio M. Late onset oculopharyngeal muscular dystrophy with prominent neurogenic features and short GCG trinucleotide expansion. Muscle Nerve. 2011 Jan;43(1):141–2.

4. Gymrek M. A genomic view of short tandem repeats. Curr Opin Genet Dev. 2017 Jun;44:9–16.

5. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, Erlich Y. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016 Jan;48(1):22–9.

6. Abe H, Gemmell NJ. Abundance, arrangement, and function of sequence motifs in the chicken promoters. BMC Genomics. 2014 Oct 15;15:900.

7. Iolascon A, Faienza MF, Centra M, Storelli S, Zelante L, Savoia A. (TA)8 allele in the UGT1A1 gene promoter of a Caucasian with Gilbert's syndrome. Haematologica. 1999 Jan 1;84(2):106–9.

8. Parvez MK, Goyal A, Kazim N, Hasnain SE, Sarin SK. TA-insertion mutation in bilirubin-UDP glucuronosyltransferase gene (UGT1A1) promoter in Indian patients with Gilbert's syndrome. J Hepatol. 2002 Apr;36:159–60.

9. Wong B, Chen S, Kwon J-A, Rich A. Characterization of Z-DNA as a nucleosome-boundary element in yeast Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 2007 Feb 13;104(7):2229–34.

10. Krynetskiy E. Beyond SNPs and CNV: Pharmacogenomics of Polymorphic Tandem Repeats. J Pharmacogenomics Pharmacoproteomics. 2017 Jun 27;8(2):1–11.

11. Uhlemann A-C, Szlezák NA, Vonthein R, Tomiuk J, Emmer SA, Lell B, Kremsner PG, Kun JFJ. DNA phasing by TA dinucleotide microsatellite length determines in vitro and in vivo expression of the gp91phox subunit of NADPH oxidase and mediates protection against severe malaria. J Infect Dis. 2004 Jun 15;189(12):2227–34.

12. Lalioti MD, Scott HS, Antonarakis SE. Altered spacing of promoter elements due to the dodecamer repeat expansion contributes to reduced expression of the cystatin B gene in EPM1. Hum Mol Genet. 1999 Sep;8(9):1791–8.

13. Tabolacci E, Pietrobono R, Moscato U, Oostra BA, Chiurazzi P, Neri G. Differential epigenetic modifications in the FMR1 gene of the fragile X syndrome after reactivating pharmacological treatments. Eur J Hum Genet. 2005 May;13(5):641–8.

14. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lead analysts:, Laboratory, Data Analysis &Coordinating Center (LDACC):, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:, Battle A, Brown CD, Engelhardt BE, Montgomery SB. Genetic effects on gene expression across human tissues. Nature. 2017 Oct 11;550(7675):204–13.

15. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Methods. 2017 Jun;14(6):590–2.

16. Bahcall OG. Human genetics: GTEx pilot quantifies eQTL variation across tissues and individuals. Nat Rev Genet. 2015 Jul;16(7):375.

17. Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Geuvadis Consortium, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Häsler R, Syvänen A-C, van Ommen G-J, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013 Sep 26;501(7468):506–11.

18. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, Erlich Y. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016 Jan;48(1):22–9.

19. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016 Dec 1;99(6):1245–60.

20. Borel C, Migliavacca E, Letourneau A, Gagnebin M, Béna F, Sailani MR, Dermitzakis ET, Sharp AJ, Antonarakis SE. Tandem repeat sequence variation as causative Cis-eQTLs for protein-coding gene expression variation: The case of CSTB. Hum Mutat. 2012;33(8):1302–9.

21. Kobayashi H, Abe K, Matsuura T, Ikeda Y, Hitomi T, Akechi Y, Habu T, Liu W, Okuda H, Koizumi A. Expansion of intronic GGCCTG hexanucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. Am J Hum Genet. 2011 Jul 15;89(1):121–30.

22. Lee Y-C, Tsai P-C, Guo Y-C, Hsiao C-T, Liu G-T, Liao Y-C, Soong B-W. Spinocerebellar ataxia type 36 in the Han Chinese. Neurol Genet. 2016 Jun;2(3):e68.

23. Vikman S, Brena RM, Armstrong P, Hartiala J, Stephensen CB, Allayee H. Functional analysis of 5-lipoxygenase promoter repeat variants. Hum Mol Genet. 2009 Dec 1;18(23):4521–9.

24. Mougey E, Lang JE, Allayee H, Teague WG, Dozor AJ, Wise RA, Lima JJ. ALOX5 polymorphism associates with increased leukotriene production and reduced lung function and asthma control in children with poorly controlled asthma. Clin Exp Allergy. 2013 May;43(5):512–20.

25. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57–74.

26. Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, Gemmell N. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLoS One. 2013 Feb 6;8(2):e54710.

27. Yáñez-Cuna JO, Arnold CD, Stampfel G, Boryń LM, Gerlach D, Rath M, Stark A. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. Genome Res. 2014 Jul;24(7):1147–56.

28. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012 Feb 28;9(3):215–6.

29. Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, Gemmell N. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLoS One. 2013 Feb 6;8(2):e54710.

30. Hänsel-Hertsch R, Beraldi D, Lensing SV, Marsico G, Zyner K, Parry A, Di Antonio M, Pike J, Kimura H, Narita M, Tannahill D, Balasubramanian S. G-quadruplex structures mark human regulatory chromatin. Nat Genet. 2016 Oct;48(10):1267–72.

31. Hänsel-Hertsch R, Di Antonio M, Balasubramanian S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. Nat Rev Mol Cell Biol. 2017 May;18(5):279–84.

32. Eddy J, Vallur AC, Varma S, Liu H, Reinhold WC, Pommier Y, Maizels N. G4 motifs correlate with promoter-proximal transcriptional pausing in human genes. Nucleic Acids Res. 2011 Jul;39(12):4975–83.

33. Szlachta K, Thys RG, Atkin ND, Pierce LCT, Bekiranov S, Wang Y-H. Alternative DNA secondary structure formation affects RNA polymerase II promoter-proximal pausing in human. Genome Biol. 2018 Jul 12;19(1):89.

34. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003 Jul 1;31(13):3406–15.

35. Andronescu M, Aguirre-Hernández R, Condon A, Hoos HH. RNAsoft: A suite of RNA secondary structure prediction and design software tools. Nucleic Acids Res. 2003 Jul 1;31(13):3416–22.

36. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014 Jan;42(Database issue):D1001–6.

37. Downloads — Psychiatric Genomics Consortium [Internet]. 2017 [cited 2018 Sep 2]. Available from: https://www.med.unc.edu/pgc/results-and-downloads

38. Willems T, Gymrek M, Highnam G, 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014 Nov;24(11):1894–904.

39. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014 May;10(5):e1004383.

40. Schmidt MHM, Pearson CE. Disease-associated repeat instability and mismatch repair. DNA Repair. 2016 Feb;38:117–26.

41. Tang H, Kirkness EF, Lippert C, Biggs WH, Fabani M, Guzman E, Ramakrishnan S, Lavrenko V, Kakaradov B, Hou C, Hicks B, Heckerman D, Och FJ, Caskey CT, Venter JC, Telenti A. Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. Am J Hum Genet. 2017 Nov 2;101(5):700–15.

42. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Methods [Internet]. 2017 Apr 24; Available from: http://dx.doi.org/10.1038/nmeth.4267

43. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996–1006.

44. Saini S, Mitra I, Gymrek M. A reference haplotype panel for genome-wide imputation of short tandem repeats [Internet]. 2018. Available from: http://dx.doi.org/10.1101/277673

45. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, Erlich Y. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016 Jan;48(1):22–9.

46. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012 Feb 16;7(3):500–7.

47. Robinson D, Storey MJD, Bass AJ. Package “qvalue.” Available from: https://bioc.ism.ac.jp/packages/3.3/bioc/manuals/qvalue/man/qvalue.pdf

48. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. 2017 Dec;12(12):2478–92.

49. Køllgaard T, Kornblit B, Petersen J, Klausen TW, Mortensen BK, Brændstrup P, Sengeløv H, Høgdall E, Müller K, Vindeløv L, Andersen MH, Straten PT. (GT)n Repeat Polymorphism in Heme Oxygenase-1 (HO-1) Correlates with Clinical Outcome after Myeloablative or Nonmyeloablative Allogeneic Hematopoietic Cell Transplantation. PLoS One [Internet]. 2016 [cited 2018 Aug 21];11(12). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5172582/

50. Borrmann L, Seebeck B, Rogalla P, Bullerdiek J. Human HMGA2 promoter is coregulated by a polymorphic dinucleotide (TC)-repeat. Oncogene. 2003 Feb 6;22(5):756–60.

51. Borel C, Migliavacca E, Letourneau A, Gagnebin M, Béna F, Sailani MR, Dermitzakis ET, Sharp AJ, Antonarakis SE. Tandem repeat sequence variation as causative cis-eQTLs for protein-coding gene expression variation: the case of CSTB. Hum Mutat. 2012 Aug;33(8):1302–9.

52. Matsuzono K, Imamura K, Murakami N, Tsukita K, Yamamoto T, Izumi Y, Kaji R, Ohta Y, Yamashita T, Abe K, Inoue H. Antisense Oligonucleotides Reduce RNA Foci in Spinocerebellar Ataxia 36 Patient iPSCs. Mol Ther Nucleic Acids. 2017 Sep 15;8:211–9.

53. Shimajiri S, Arima N, Tanimoto A, Murata Y, Hamada T, Wang KY, Sasaguri Y. Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. FEBS Lett. 1999 Jul 16;455(1-2):70–4.

54. Gebhardt F, Zänker KS, Brandt B. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. J Biol Chem. 1999 May 7;274(19):13176–80.

55. Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M. A polymorphic microsatellite that mediates induction of PIG3 by p53. Nat Genet. 2002 Mar;30(3):315–20.

56. Vikman S, Brena RM, Armstrong P, Hartiala J, Stephensen CB, Allayee H. Functional analysis of 5-lipoxygenase promoter repeat variants. Hum Mol Genet. 2009 Dec 1;18(23):4521–9.

57. Saha A, Dhir A, Ranjan A, Gupta V, Bairwa N, Bamezai R. Functional IFNG polymorphism in intron 1 in association with an increased risk to promote sporadic breast cancer. Immunogenetics. 2005 May;57(3-4):165–71.

58. Johnson AD, Kavousi M, Smith AV, Chen M-H, Dehghan A, Aspelund T, Lin J-P, van Duijn CM, Harris TB, Cupples LA, Uitterlinden AG, Launer L, Hofman A, Rivadeneira F, Stricker B, Yang Q, O'Donnell CJ, Gudnason V, Witteman JC. Genome-wide association meta-analysis for total serum bilirubin levels. Hum Mol Genet. 2009 Jul 15;18(14):2700–10

59. William J. Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L. Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A. Kostadima, John J. Lambourne, Suthesh Sivapalaratnam, Kate Downes, Kousik Kundu, Lorenzo Bomba, Kim Berentsen, John R. Bradley, Louise C. Daugherty, Olivier Delaneau,Kathleen Freson, Stephen F. Garner, Luigi Grassi, Jose Guerrero, Matthias Haimel, Eva M. Janssen-Megens, Anita Kaan,Mihir Kamat, Bowon Kim, Amit Mandoli, Jonathan Marchini, Joost H.A. Martens, Stuart Meacham, Karyn Megy,Jared O'Connell, Romina Petersen, Nilofar Sharifi, Simon M. Sheard, James R. Staley, Salih Tuna, Martijn van der Ent,Klaudia Walter, Shuang-Yin Wang, Eleanor Wheeler, Steven P. Wilder, Valentina Iotchkova,Carmel Moore, Jennifer Sambrook, Hendrik G. Stunnenberg, Emanuele Di Angelantonio, Stephen Kaptoge, Taco W. Kuijpers, Enrique Carrillo-de-Santa-Pau, David Juan, Daniel Rico, Alfonso Valencia, Lu Chen, Bing Ge, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yang, Roderic Guigo, Stephan Beck, Dirk S. Paul, Tomi Pastinen, David Bujold, Guillaume Bourque, Mattia Frontini, John Danesh, David J. Roberts, Willem H. Ouwehand, Adam S. Butterworth, Nicole Soranzo. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167,** 1415–1429.e19 (2016).

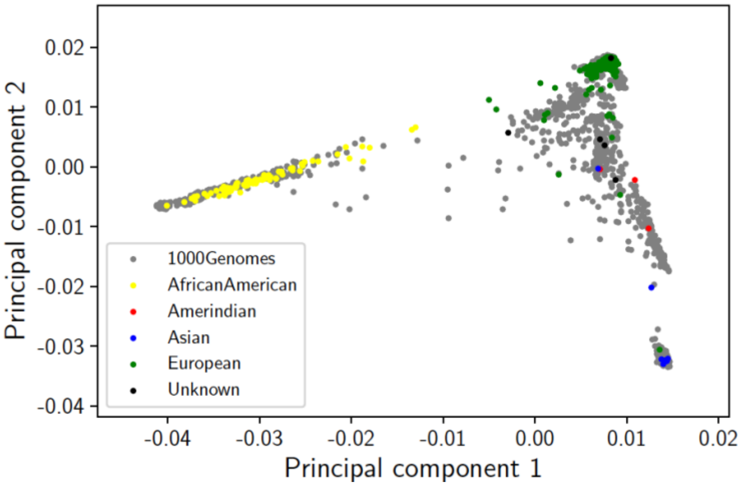
60. Benjamin B. Sun, Joseph C. Maranville, James E. Peters, David Stacey, James R. Staley, James Blackshaw, Stephen Burgess, Tao Jiang, Ellie Paige, Praveen Surendran, Clare Oliver-Williams, Mihir A. Kamat, Bram P. Prins, Sheri K. Wilcox, Erik S. Zimmerman, An Chi, Narinder Bansal, Sarah L. Spain, Angela M. Wood, Nicholas W. Morrell, John R. Bradley, Nebojsa Janjic, David J. Roberts, Willem H. Ouwehand, John A. Todd, Nicole Soranzo, Karsten Suhre, Dirk S. Paul, Caroline S. Fox, Robert M. Plenge, John Danesh, Heiko Runz, Adam S. Butterworth. Genomic atlas of the human plasma proteome. *Nature* **558,** 73–79 (2018).

61. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).

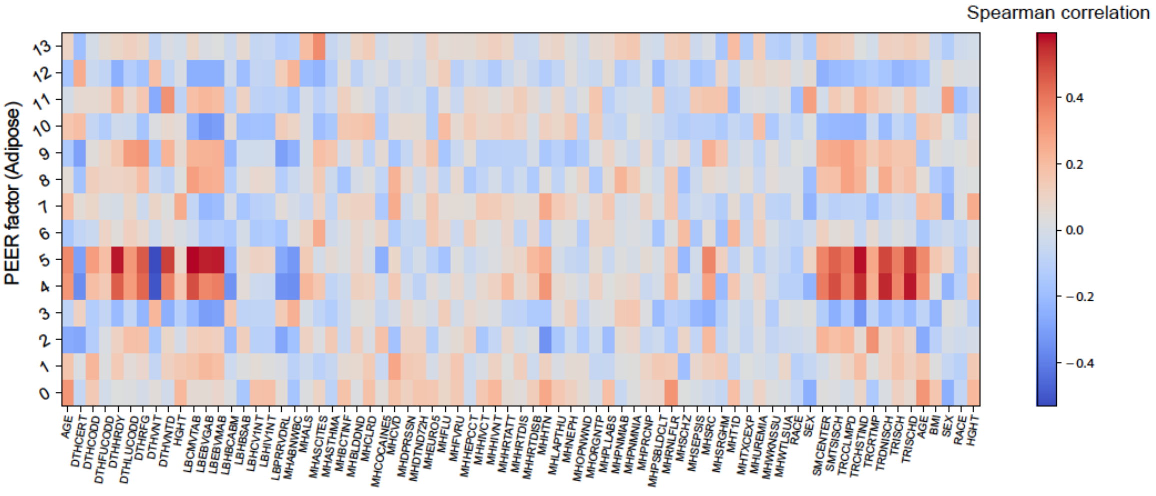
62. Loic Yengo, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, Andrew R Wood, Michael N Weedon, Timothy M Frayling, Joel Hirschhorn, Jian Yang, Peter M Visscher, GIANT Consortium doi: https://doi.org/10.1101/274654 Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry. *bioRxiv* 274654 (2018). doi:10.1101/274654

63. Ding, N. Tomomori-Sato C, Sato S, Conaway RC, Conaway JW, Boyer TG MED19 and MED26 are synergistic functional targets of the RE1 silencing transcription factor in epigenetic silencing of neuronal gene expression. *J. Biol. Chem.* **284,** 2648–2656 (2009).

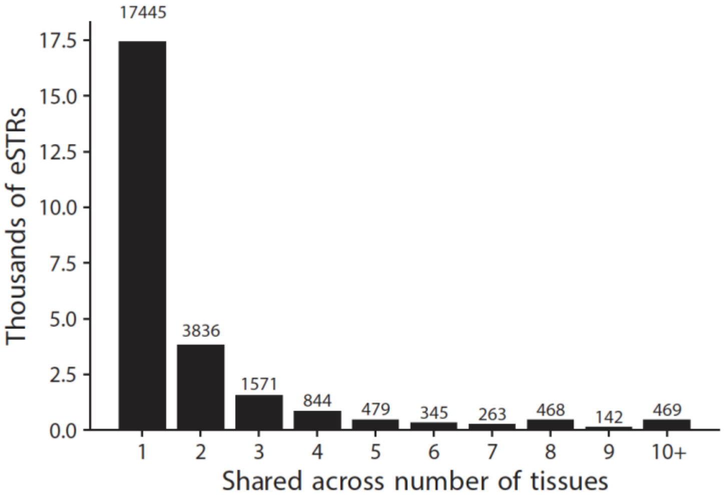
Appendix of Chapter3



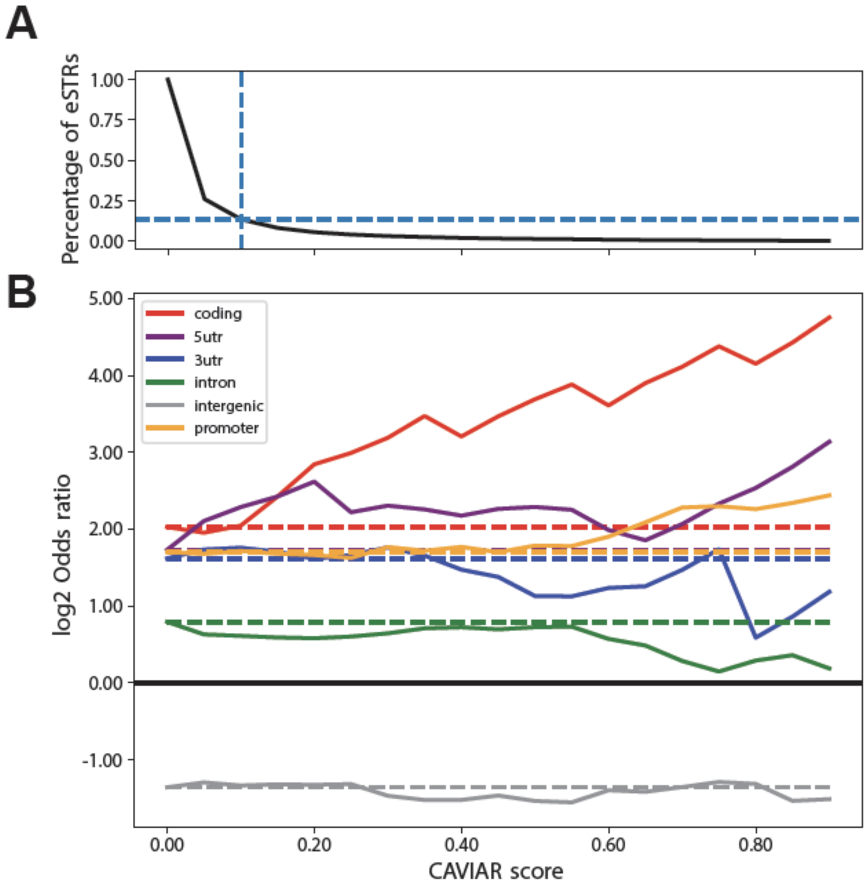
Supplementary Figure S3.1.1: Analysis of GTEx population structure. Principal component analysis was performed using SNP genotypes from the GTEx and 1000G cohorts. Samples from the 1000 Genomes project are shown in gray and GTEx samples are shown as colored dots based on ethnicity provided for each sample (yellow=African American; red=Amerindian; blue=Asian; green=European, black=Unknown).



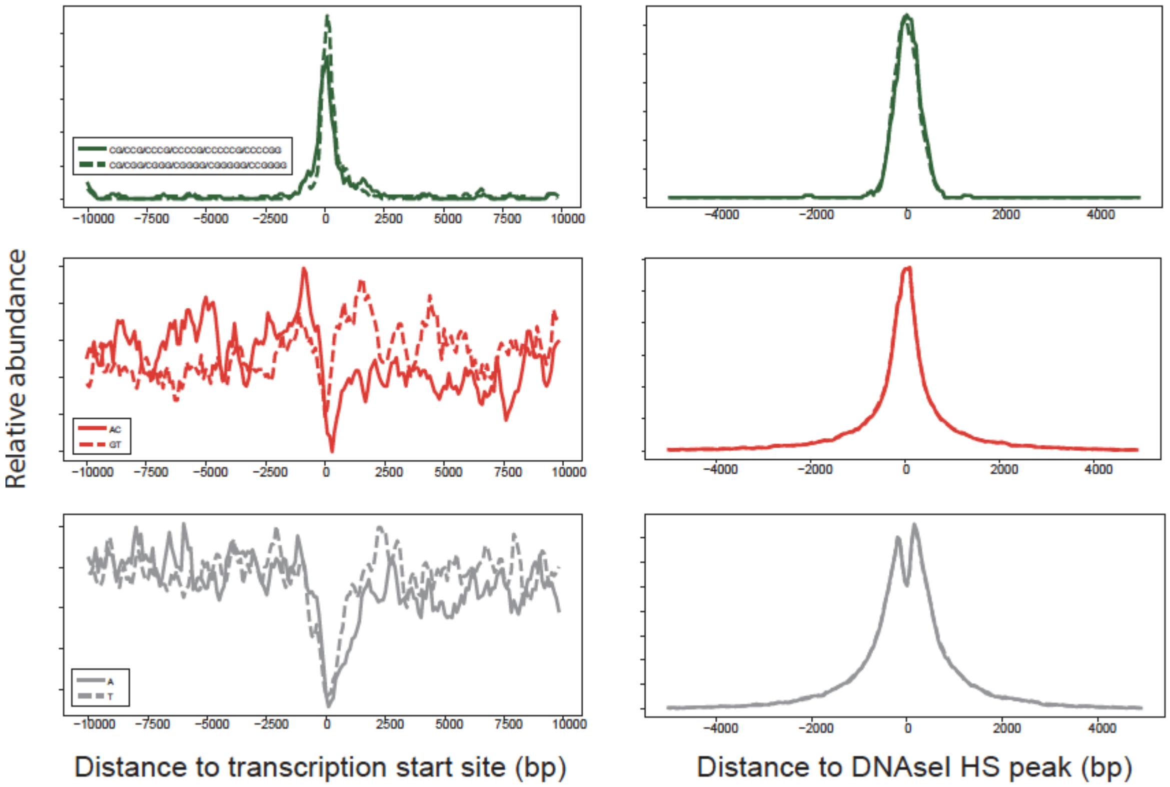
Supplementary Figure S3.1.2: Correlation of sample metadata with PEER factors. Each cell gives the squared spearman correlation of PEER factor with data processing covariates. The x-axis gives each variable as defined for dbGaP study phs000424.v7.p2. For example, covariates most strongly associated with PEER factors included DTHHRDY (Hardy scale for death classification) and TRISCHD (ischemic time). The y-axis gives PEER factors obtained from PEER analysis of gene expression from Adipose-subcutaneous tissue. Similar correlations were observed for other tissues.



Supplementary Figure S3.1.3: Sharing of eSTRs across tissues. The x-axis gives the number of tissues that share a given STR. The y-axis and annotated values for each bar give the number of eSTRs shared across that many tissues.

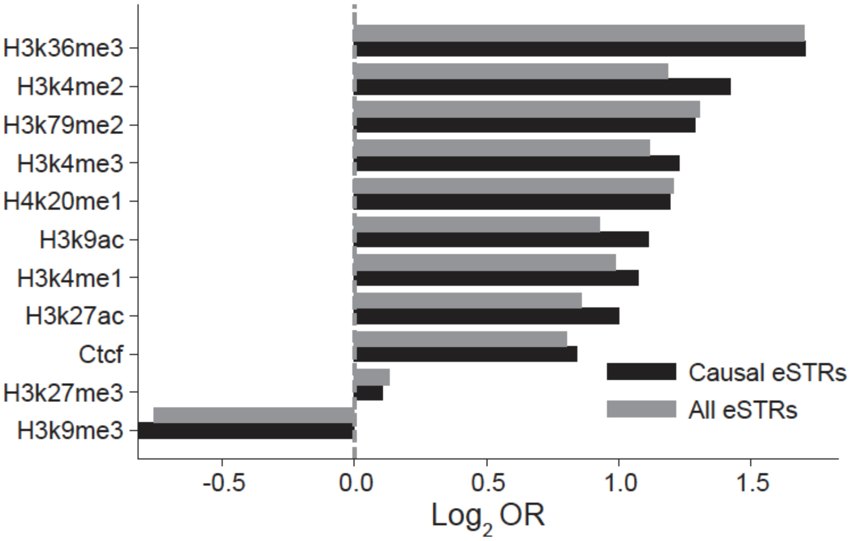


Supplementary Figure S3.1.4: Causal eSTRs are enriched in coding and regulatory regions. A. Percent of eSTRs that are annotated as causal as a function of CAVIAR threshold. The x-axis gives CAVIAR posterior probability and the y-axis gives the percentage of eSTRs with CAVIAR scores above each threshold. The dashed horizontal line gives the percent of eSTRs with CAVIAR scores of at least 10%. **B. Enrichment of gene annotation categories as a function of CAVIAR threshold.** The y-axis gives the $\log_2$ odds ratio for enrichment of each category in eSTRs passing each threshold (solid line). The dashed line gives the odds ratio when considering all eSTRs regardless of CAVIAR score. Red=coding, purple=5' UTR, blue=3' UTR, green=intron, gray=intergenic, orange= promoter (within 3kb upstream of each gene).



Supplementary Figure S3.2.1: Localization of STRs around putative regulatory regions. Left and right plots give localization around transcription start sites and DNAseI HS clusters, respectively. The y-axis gives the relative number of STRs of each type in each bin. For promoters, the x-axis is divided into 100bp bins. For DNAseI HS sites, the x-axis is divided into 50bp bins. In each plot, values were smoothed by taking a sliding average of each four consecutive bins. Only STR-gene pairs passing all filters are considered.

Each plot compares localization of the two possible sequences of a given repeat motif on the coding strand. *I.e.* top plots compare motifs of the form C_nG_m vs. their reverse complement on the opposite strand, middle plots compare AC vs. GT repeats. And bottom plots compare A vs. T repeats. The strand of each STR was determined based on the coding strand of each target gene.



Supplementary Figure S3.2.2: Enrichment of eSTRs in peaks of histone modifications. Bars give the $\log_2$ odds ratio from performing Fisher's exact test comparing eSTRs to all STRs (gray bars) or all causal eSTRs to all STRs (black bars). Histone modification peaks were obtained from ENCODE (**Methods**).

Supplementary Table S3.1.1 STRs previously associated with expression that were analyzed in the GTEx cohort. The right-hand column reports STR by gene associations in the GTEx cohort with nominal p<0.01. Nominal p-values and effect sizes (β) are given for each nominal eSTR. Bolded results in the right column indicate eSTRs that agree with the direction of effect reported previously. Note several previously reported eSTRs (*e.g.* a TG repeat in the promoter of *HMOX1*[49] and a TC repeat in the promoter of *HMGA2*[50]) did not pass our quality filters and thus were not included.

Gene	STR (hg19)	Reference effect	eSTR evidence (p<0.01)
CSTB[51]	21:45196326 (CGGGGCGGGGCG$_n$)	Expansion implicated in myoclonus epilepsy. Normal alleles show increasing effect in LCLs (β>0);	**Adipose-Visceral; p=2.5e-6; β=0.34** **Transformedfibroblasts; p=1.5e-18; β=0.55** **Skin-SunExposed; p=1.89e-14; β=0.43 Heart-LeftVentricle; p=1.6e-11; β=0.46** **Esophagus-Muscularis; p=2.8e-9; β=0.40 Artery-Tibial; p= 2.3e-15; β=0.46** **Artery-Aorta; p=2.0e-5; β=0.31** **Adipose-Subcutaneous; p=1.9e-4; β=0.23 Skin-NotSunExposed; p=6.3e-7; β=0.34** **Thyroid; p=4.1e-16; β=0.48** **WholeBlood; p=2.1e-14; β=0.41** **Nerve-Tibial; p=7.3e-14; p=0.45** **Muscle-Skeletal; p=1.3e-21; β=0.49**
NOP56[52]	20:2633379 (GGCCTG$_n$)	Expansion implicated in spinocerebellar ataxia 36. Large expansions show decreasing effect in SCA36 patient iPSC cell lines (β<0). No data on normal alleles.	Skin-SunExposed; p= 6.0e-05; β=0.23 Artery-Tibial; p=9.5e-5; β=0/24 Adipose-Subcutaneous; p=0.0011; β=0.20 Skin-NotSunExposed; p=0.0039; β=0.21 Thyroid; p=2.4e-8; β=0.34 Muscle-Skeletal; p=7.8e-12; β=0.36
MMP9[53]	20:44637413 (AC$_n$)	Increasing effect in esophageal carcinoma cell lines (β>0)	**Skin-SunExposed; p=0.0039; β=0.18**
EGFR[54]	7:55088254 (AC$_n$)	*In vitro* decreasing effect; decreasing but non-linear effect across various cell lines (β<0)	**Heart-LeftVentricle; p=0.0073; β=-0.21** Esophagus-Mucosa; p=0.0018; β=0.21 Muscle-Skeletal; p=0.0070; β=0.16
TP53I3[55]	2:24307211 (TGYCC$_n$)	Increasing effect in presence of p53 in H1299 cells (β>0)	**Esophagus-Mucosa; p=0.0081; β=0.18** **Thyroid; p=0.00060; β=0.22** **Muscle-Skeletal; p=0.0068; β=0.15**
ALOX5[56]	10:45869549 (CGGGGG$_n$)	Increasing effect in monocytes (β>0), decreasing effect in lymphocytes (β<0)	**Esophagus-Mucosa; p=4.35e-7; β=0.32** **Brain-Cerebellum; p=0.0027; β=0.29 Skin-SunExposed; p=9.0e-4; β=-0.19**

IFNG[57]	12:68552495 (AC$_n$)	In vitro reporter assay showed increasing effect for CA12 vs. CA15 ($\beta>0$)	No GTEx eSTRs found
UGT1A1[58]	2:234668880 (AT$_n$)	Decreasing effect in human hepatoma ($\beta<0$)	No GTEx eSTRs found

Supplementary Table S3.2.1 Enrichment of eSTRs for genomic annotations. For each annotation, a Fisher's exact test was performed to test whether eSTRs showed significantly more or less overlap compared to all STRs analyzed. Causal STRs represent all STRs with CAVIAR score of at least 10% as defined in the main text. OR=odds ratio and p-values are two-sided.

Annotation	# eSTRs			eSTR enrichment		Causal eSTR enrichment	
	Total	eSTRs	Causal	P-val	OR	P-val	OR
Coding	217	75	16	4.0e-19	4.1	5.0e-6	4.1
5' UTR	657	197	56	1.9e-37	3.3	2.8e-20	4.9
3' UTR	1,660	468	100	3.1e-77	3.1	2.6e-23	3.4
Intron	74,501	10,998	1,759	1.4e-294	1.7	1.1e-33	1.5
Intergenic	91,325	6,458	1,016	p<1e-300	0.39	2.5e-142	0.40
Promoter (within 3kb upstream of TSS)	6,866	1,941	380	p<1e-300	3.2	7.9e-77	3.2

Supplementary Table S3.2.2 Causal eSTRs overlapping protein-coding regions.

Chrom	**STR position (hg19)**	**Gene**	**Motif**	**Top CAVIAR score**	**Top tissue**
chr14	24769851	*DHRS1*	CCT	1.00	Thyroid (β=-0.42)
chr9	88356816	*AGTPBP1*	CCG	1.00	Cells-Transformedfibroblasts (β=0.36)
chr3	40503522	*RPL14*	CTG	0.99	Nerve-Tibial (β=-0.32)
chr11	6411932	*SMPD1*	CGCTGG	0.79	Thyroid (β=0.27)
chr2	25384461	*POMC*	AGC	0.59	Muscle-Skeletal (β=0.35)
chr19	55790888	*HSPBP1*	CGG	0.58	Cells-Transformedfibroblasts (β=-0.37)
chr11	124750441	*ROBO3*	AGCCGG	0.39	Thyroid (β=0.33)
chr3	184429135	*MAGEF1*	AGG	0.38	Thyroid (β=-0.33)
chr19	50093219	*PRR12A*	ACCCCC	0.36	Skin-SunExposed (β=0.51)
chr19	11558342	*EPOR*	CCT	0.29	Heart-LeftVentricle (β=0.28)
chr17	17697095	*TOM1L2*	CTG	0.22	Lung (β=0.28)
chr16	71956508	*IST1*	ATGCCC	0.20	Artery-Aorta (β=0.27)
chr20	3026347	*MRPS26*	CCCCG	0.17	Nerve-Tibial (β=0.32)
chr14	21560753	*ZNF219*	AGCCTC	0.14	Adipose-Subcutaneous (β=0.22)
chr7	96635365	*DLX5*	CTG	0.12	Muscle-Skeletal (β=-0.21)
chr21	47721987	*MCM3AP*	ACC	0.11	Artery-Aorta (β=0.46)

Supplementary Table S3.2.3: Enrichment of eSTRs in ChromHMM states. For each state, a Fisher's exact test was performed to test whether eSTRs showed significantly more or less overlap compared to all STRs analyzed. Causal STRs represent all STRs with CAVIAR score of at least 10% as defined in the main text. OR=odds ratio and p-values are two-sided. State annotations are combined across various ENCODE cell lines (**Methods**).

ChromHMM state	**# eSTRs**			**eSTR enrichment**		**Causal eSTR enrichment**	
	Total	**eSTRs**	**Causal**	**P-val**	**OR**	**P-val**	**OR**
Active Promoter	1,542	499	132	4.4e-106	3.8	4.8e-46	5.0
Weak Promoter	2,374	725	168	4.7e-139	3.5	3.7e-47	4.1
Txn Elongation	12,191	3,632	654	p<1e-300	3.8	1.7e-130	3.4
Txn Transition	3,420	875	179	1.2e-117	2.7	4.3e-33	3.0
Weak Txn	53,913	10,386	1,721	p<1e-300	2.7	7.7e-142	2.5
Strong Enhancer	5,261	1,088	210	4.7e-84	2.1	3.4e-23	2.2
Poised Promoter	239	239	48	4.0e-12	1.7	4.4e-5	1.9
Weak Enhancer	15,863	2,774	513	1.6e-121	1.7	1.5e-32	1.9
Repetitive	548	93	17	1.2e-4	1.6	5.6e-2	1.7
Repressed	20,807	3,001	514	7.9e-43	1.4	5.4e-10	1.4
Insulator	1,724	262	43	3.5e-6	1.4	7.5e-2	1.3
Heterochromatin	150,499	13,021	1,970	p<1e-300	0.2	4.6e-318	0.2

Supplementary Table 3.2.4: Enrichment of eSTRs in transcription factor binding sites annotated by ENCODE. For each factor, a Fisher's exact test was performed to test whether eSTRs showed significantly more or less overlap compared to all STRs analyzed. Causal STRs represent all STRs with CAVIAR score of at least 10% as defined in the main text. OR=odds ratio and p-values are two-sided. Transcription factor annotations are described in **Methods**.

ChromHMM state	# eSTRs			eSTR enrichment		Causal eSTR enrichment	
	Total	eSTRs	Causal	P-val	OR	P-val	OR
KDM5A	13	8	6	2.3E-05	12.3	7.1E-08	44.4
RDBP	9	5	3	1.7E-03	9.6	5.3E-04	25.9
THAP1	32	19	6	1.1E-10	11.3	2.8E-05	11.9
SAP30	170	72	27	2.0E-24	5.7	3.8E-17	9.8
SIX5	99	42	15	6.3E-15	5.7	7.0E-10	9.3
PHF8	462	198	68	2.4E-65	5.8	1.0E-38	9.1
ELK4	111	46	15	9.9E-16	5.5	3.6E-09	8.1
NRF1	127	47	17	8.0E-14	4.5	3.9E-10	8.0
KDM5B	321	137	42	1.8E-45	5.8	1.8E-22	7.9
SREBP1	38	17	5	2.8E-07	6.2	7.3E-04	7.8
CTCFL	62	16	8	1.9E-03	2.7	2.3E-05	7.7
NFYA	81	29	10	1.0E-08	4.3	3.3E-06	7.3
HMGN3	310	126	38	2.9E-39	5.3	1.7E-19	7.3
E2F6	416	159	50	5.0E-45	4.8	9.2E-25	7.2
CREB1	241	93	29	2.2E-27	4.9	5.1E-15	7.1
SP4	117	54	14	8.1E-21	6.6	5.7E-08	7.0
ZBTB7A	315	113	36	1.3E-29	4.3	1.5E-17	6.7
CEBPD	175	66	20	2.8E-19	4.7	2.1E-10	6.7
GTF2B	44	17	5	3.3E-06	4.9	1.4E-03	6.6
CCNT2	401	143	45	1.1E-36	4.3	3.1E-21	6.6
HDAC1	244	92	27	3.1E-26	4.7	3.3E-13	6.5

ELK1	141	50	15	9.1E-14	4.2	9.5E-08	6.2
TAF7	196	79	20	6.0E-25	5.2	1.5E-09	5.9
IRF1	430	147	43	2.7E-35	4.0	1.9E-18	5.8
SIN3A	585	185	58	1.1E-38	3.6	3.6E-24	5.8
E2F4	295	96	29	7.2E-22	3.7	9.3E-13	5.7
ZKSCAN1	72	16	7	8.5E-03	2.2	4.4E-04	5.6
SRF	159	41	15	5.6E-07	2.7	4.6E-07	5.4
TAF1	937	298	85	5.7E-62	3.6	6.4E-32	5.3
GABPA	586	180	53	7.4E-36	3.4	2.0E-20	5.2
E2F1	604	190	54	2.4E-39	3.6	1.5E-20	5.1
MAZ	992	314	85	1.1E-64	3.6	3.7E-30	4.9
RBBP5	710	216	61	5.6E-42	3.4	4.0E-22	4.9
NFYB	200	57	17	7.7E-11	3.1	3.6E-07	4.8
UBTF	273	88	23	7.0E-20	3.7	4.0E-09	4.8
ETS1	250	88	21	8.9E-23	4.2	2.0E-08	4.8
ZEB1	60	17	5	3.1E-04	3.0	5.7E-03	4.7
EGR1	658	186	54	1.3E-31	3.1	7.0E-19	4.7
SMARCB1	257	89	21	1.9E-22	4.1	3.2E-08	4.6
SP2	37	12	3	6.1E-04	3.7	3.3E-02	4.6
CHD1	512	144	41	9.9E-25	3.0	2.4E-14	4.5
CHD2	624	177	49	1.6E-30	3.1	1.6E-16	4.5
BRCA1	116	38	9	1.1E-09	3.8	3.9E-04	4.4
FOXP2	572	141	44	2.0E-18	2.5	1.1E-14	4.3
PML	642	175	49	7.3E-28	2.9	5.0E-16	4.3
HDAC2	472	132	36	1.4E-22	3.0	3.6E-12	4.3
TCF3	342	93	26	2.2E-15	2.9	3.6E-09	4.3
MXI1	786	218	59	2.0E-35	3.0	1.2E-18	4.3

ELF1	665	200	50	3.4E-38	3.3	4.6E-16	4.2
POU2F2	492	125	37	9.8E-18	2.6	2.8E-12	4.2
SIN3AK20	900	263	67	3.8E-47	3.2	8.9E-21	4.2
MBD4	165	46	12	9.3E-09	3.0	8.4E-05	4.1
TBP	834	218	59	1.6E-31	2.7	1.9E-17	4.0
PAX5	594	161	41	1.8E-25	2.9	2.9E-12	3.9
MEF2C	159	32	11	1.6E-03	1.9	2.5E-04	3.8
STAT1	392	107	27	1.1E-17	2.9	1.5E-08	3.8
STAT2	58	17	4	2.0E-04	3.2	2.4E-02	3.8
SMARCC2	60	15	4	3.4E-03	2.6	2.7E-02	3.7
GTF2F1	301	79	19	1.4E-12	2.7	6.7E-06	3.5
RFX5	430	97	27	9.4E-11	2.2	9.6E-08	3.5
BCLAF1	225	53	14	3.2E-07	2.4	1.2E-04	3.4
RELA	897	216	55	4.0E-26	2.5	8.4E-14	3.4
TFAP2A	390	83	24	2.8E-08	2.1	6.9E-07	3.4
SP1	758	191	46	9.3E-26	2.6	1.2E-11	3.4
YY1	1,187	278	71	6.5E-31	2.4	8.0E-17	3.3
REST	952	209	57	5.0E-20	2.2	8.0E-14	3.3
TBL1XR1	449	110	27	1.7E-14	2.5	2.3E-07	3.3
ZBTB33	234	64	14	3.7E-11	2.9	1.8E-04	3.3
MYBL2	438	109	26	7.2E-15	2.6	4.8E-07	3.3
POLR2A	4,645	1,118	262	1.1E-130	2.5	4.1E-54	3.3
FOSL1	170	35	10	6.4E-04	2.0	1.6E-03	3.2
BHLHE40	757	170	44	1.3E-17	2.2	1.3E-10	3.2
MTA3	353	83	20	2.7E-10	2.4	1.9E-05	3.1
MYC	1,984	460	110	4.0E-49	2.4	1.1E-22	3.1
RXRA	249	63	14	1.5E-09	2.6	3.4E-04	3.1

WRNIP1	254	63	14	4.5E-09	2.5	4.2E-04	3.0
NR2C2	145	30	8	1.5E-03	2.0	6.8E-03	3.0
ZNF143	620	139	34	1.3E-14	2.2	6.2E-08	3.0
MAX	1,718	429	92	6.9E-55	2.6	3.0E-18	3.0
USF2	279	64	15	8.0E-08	2.3	3.4E-04	2.9
NR3C1	467	94	25	8.1E-08	1.9	4.9E-06	2.9
STAT5A	474	96	25	4.4E-08	2.0	6.4E-06	2.9
USF1	665	153	35	5.2E-17	2.3	1.1E-07	2.9
TFAP2C	504	112	26	8.5E-12	2.2	6.2E-06	2.8
MEF2A	340	64	17	7.9E-05	1.8	3.3E-04	2.7
ARID3A	542	116	27	5.0E-11	2.1	7.8E-06	2.7
SMARCA4	120	24	6	6.1E-03	1.9	2.7E-02	2.7
RCOR1	805	177	40	3.1E-17	2.2	6.5E-08	2.7
NFIC	1,161	222	57	3.3E-14	1.8	1.9E-10	2.7
HNF4A	266	60	13	3.4E-07	2.2	1.9E-03	2.7
BACH1	226	54	11	1.7E-07	2.4	4.3E-03	2.6
CBX3	392	100	19	1.6E-14	2.6	2.3E-04	2.6
CTBP2	207	44	10	6.7E-05	2.1	6.6E-03	2.6
JUNB	228	51	11	3.2E-06	2.2	4.6E-03	2.6
EZH2	858	153	41	4.9E-08	1.7	2.0E-07	2.6
TCF12	813	167	38	1.2E-13	2.0	7.8E-07	2.6
RUNX3	1,193	238	55	2.6E-17	1.9	3.6E-09	2.5
SMARCC1	240	51	11	1.6E-05	2.1	6.7E-03	2.5
GTF3C2	111	29	5	1.7E-05	2.7	6.1E-02	2.4
ATF3	313	82	14	8.7E-13	2.7	3.0E-03	2.4
EBF1	682	129	30	1.7E-08	1.8	2.9E-05	2.4
FOXM1	548	123	24	4.6E-13	2.2	2.0E-04	2.4

TRIM28	228	56	10	3.2E-08	2.5	1.3E-02	2.4
BATF	390	64	17	4.0E-03	1.5	2.1E-03	2.4
GATA1	668	165	29	1.6E-21	2.5	4.9E-05	2.4
ATF1	161	43	7	9.1E-08	2.8	3.5E-02	2.4
FOSL2	719	145	31	2.1E-11	2.0	4.5E-05	2.3
ZNF263	814	167	35	1.2E-13	2.0	1.3E-05	2.3
FOXA2	584	108	25	6.7E-07	1.8	2.0E-04	2.3
PBX3	213	52	9	1.3E-07	2.5	2.1E-02	2.3
JUND	1,707	312	71	1.6E-16	1.7	1.9E-09	2.3
TCF7L2	1,256	224	52	3.8E-11	1.7	3.6E-07	2.3
HNF4G	225	47	9	5.5E-05	2.0	4.3E-02	2.2
TEAD4	987	187	39	9.5E-12	1.8	3.0E-05	2.1
ZZZ3	76	10	3	5.9E-01	1.2	1.8E-01	2.1
BCL3	948	182	37	5.1E-12	1.8	6.0E-05	2.1
SPI1	795	167	30	1.6E-14	2.1	5.6E-04	2.0
FOXA1	1,309	216	49	6.2E-08	1.5	1.5E-05	2.0
NR2F2	244	52	9	1.2E-05	2.1	5.4E-02	2.0
ATF2	697	147	25	4.5E-13	2.1	3.0E-03	1.9
SMC3	737	144	26	2.8E-10	1.9	2.7E-03	1.9
SETDB1	709	118	25	4.4E-05	1.5	3.4E-03	1.9
ESR1	284	43	10	6.2E-02	1.4	7.5E-02	1.9
EP300	2,558	413	88	1.4E-12	1.5	1.7E-07	1.9
NFATC1	551	109	18	2.0E-08	1.9	2.7E-02	1.7
IKZF1	252	46	8	1.5E-03	1.7	1.6E-01	1.7
CTCF	2,746	522	86	7.3E-31	1.8	1.1E-05	1.7
IRF4	317	49	10	3.4E-02	1.4	9.9E-02	1.7
GATA2	1,628	240	49	6.6E-05	1.3	1.9E-03	1.6

ZNF217	299	37	9	6.5E-01	1.1	1.9E-01	1.6
BCL11A	271	38	8	1.8E-01	1.3	1.8E-01	1.6
TAL1	309	64	9	2.8E-06	2.0	2.0E-01	1.6
SUZ12	421	73	12	4.0E-04	1.6	1.5E-01	1.5
JUN	923	157	26	6.1E-07	1.6	5.1E-02	1.5
CEBPB	2,329	405	64	2.9E-17	1.6	4.5E-03	1.5
KAP1	1,024	172	28	4.8E-07	1.6	6.4E-02	1.5
RPC155	117	23	3	8.7E-03	1.9	4.9E-01	1.4
GATA3	1,134	139	29	4.3E-01	1.1	1.0E-01	1.4
STAT3	1,069	159	27	7.5E-04	1.3	1.4E-01	1.3
RAD21	1,247	205	30	1.7E-07	1.5	1.8E-01	1.3
MAFF	646	94	13	1.9E-02	1.3	7.7E-01	1.1
FOS	2,045	278	39	3.4E-03	1.2	9.4E-01	1.0
MAFK	1,176	156	21	6.0E-02	1.2	9.1E-01	0.9

Supplementary Table S3.2.4: Enrichment of STR motifs in eSTRs. For each motif, a Fisher's exact test was performed to test whether eSTRs were significantly more or less likely to have each motif sequence compared to all STRs analyzed. Causal STRs represent all STRs with CAVIAR score of at least 10% as defined in the main text. OR=odds ratio and p-values are two-sided. Motifs give the canonicalized repeat sequence (**Methods**).

Motifs	# eSTRs			eSTR enrichment		Causal eSTR enrichment	
	Total	eSTRs	Causal	P-val	OR	P-val	OR
CCCGG	13	7	3	2.4E-04	9.0	1.7E-03	15.5
CCCCG	41	18	9	1.8E-07	6.0	6.4E-08	14.6
CCCCCG	24	11	4	2.7E-05	6.5	1.0E-03	10.3
AGGCGG	18	8	3	4.6E-04	6.2	4.5E-03	10.3
AGCCCC	35	11	5	1.4E-03	3.5	5.0E-04	8.6
CCCCGG	23	11	3	1.6E-05	7.1	9.1E-03	7.8
CCG	210	66	22	1.0E-14	3.5	1.4E-10	6.1
AAAAG	213	39	11	3.4E-03	1.7	2.8E-03	2.8
AGC	293	55	10	3.1E-04	1.8	7.9E-02	1.8
AATC	271	48	9	2.2E-03	1.7	1.1E-01	1.8
C	1,163	165	38	4.8E-03	1.3	1.6E-03	1.8
AATT	268	50	8	5.3E-04	1.8	1.8E-01	1.6
ACC	250	32	7	4.9E-01	1.1	2.5E-01	1.5
AAAG	1,854	261	49	6.5E-04	1.3	2.6E-02	1.4
AAAT	8,806	1,306	218	2.5E-22	1.4	8.5E-05	1.3
AAAAC	3,289	485	77	1.2E-08	1.3	7.0E-02	1.2
AGAGGG	220	46	5	7.2E-05	2.0	6.2E-01	1.2
AAAC	7,779	1,158	173	2.1E-20	1.4	3.4E-02	1.2
AAAAG	491	67	11	1.4E-01	1.2	5.1E-01	1.2
ACAG	187	23	4	7.3E-01	1.1	7.8E-01	1.1
AGG	481	83	10	1.8E-04	1.6	7.4E-01	1.1

A	64,118	7,922	1,276	1.0E-17	1.1	3.4E-02	1.1
AAAAT	1,738	267	33	1.0E-06	1.4	1.0E+00	1.0
AAC	3,347	501	61	8.3E-10	1.4	8.0E-01	1.0
AC	48,732	4,389	841	4.2E-95	0.7	9.6E-04	0.9
AAGG	1,227	103	20	4.5E-04	0.7	6.0E-01	0.9
AATG	1,274	170	20	4.2E-02	1.2	4.7E-01	0.8
ATCC	930	79	14	3.3E-03	0.7	4.7E-01	0.8
AAAAAC	1,018	156	15	2.1E-04	1.4	4.2E-01	0.8
AAT	2,915	362	43	1.1E-01	1.1	1.0E-01	0.8
AGGG	348	39	5	9.3E-01	1.0	6.9E-01	0.8
AAAAAT	498	71	7	5.7E-02	1.3	5.1E-01	0.7
ACAT	656	91	9	5.7E-02	1.2	3.9E-01	0.7
AT	7,950	703	103	5.2E-15	0.7	2.5E-05	0.7
ATC	725	69	9	1.0E-01	0.8	2.2E-01	0.6
AGAT	3,461	187	42	1.4E-35	0.4	2.0E-03	0.6
AG	6,552	565	80	1.5E-14	0.7	1.7E-05	0.6
AAG	395	47	4	8.1E-01	1.0	2.6E-01	0.5

Supplementary Table S3.3.1: STR motifs with the ability to form G4 quadruplexes. STR motifs with the ability to form G4 quadruplexes. Motifs with average overlap fraction with G4 structures at least 0.7 were selected from Table 4 of Sawaya *et al.*[26]. For each motif, the first sequence gives the canonical form of the motif and the second gives the reverse complement of the canonical form.

Motif
AGGG/CCCT
ACCC/GGGT
AGGGG/CCCCT
C/G
ACCCC/GGGGT
CCCG/CGGG
CCCCG/CGGGG
AAGGG/CCCTT
AGCCC/GGGCT
AGGGC/GCCCT
ACCCCC/GGGGGT
AGCCCC/GGGGCT
CCCCCG/CGGGGG
CCCGG/CCGGG
AGAGGG/CCCTCT

Supplementary Table S3.4.1: STR in LD with GWAS SNPs more likely to drive association signal. Causal eSTRs in LD higher than 0.3 with the trait risk loci. Table containing all GWAS tested is provided as supplementary file.

chrom	**STR start**	**SNP POS**	**rsid**	**r2 (LD)**	**pval (LD)**	**Caviar score**	**eSTR qvalue**	**Best tissue**	**Traits**	**Gene name**
chr11	57523883	57510293		0.6278	5.62E-135	0.4726	1.01E-03	Adipose-Subcutaneous	Schizophrenia	MED19
chr11	124621018	124613956		0.3409	9.34E-60	0.3941	1.01E-03	Adipose-Subcutaneous	Schizophrenia	VSIG2
chr3	36835922	36858582		0.3954	5.61E-70	0.2856	4.31E-03	Skin-NotSunExposed	Schizophrenia	DCLK3
chr3	52801574	52845104		0.4274	6.45E-71	0.1224	2.78E-03	Nerve-Tibial	Schizophrenia	GNL3
chr1	16252880	16299312	rs10927875	0.3534	1.99E-58	0.1343	1.42E-02	Thyroid	Dilated_cardiomyopathy	SPEN
chr3	52801574	52815905	rs2710323	0.4398	1.25E-73	0.1224	2.78E-03	Nerve-Tibial	Schizophrenia,_schizoaffective_disorder_or_bipolar_disorder	GNL3
chr3	52801574	52831701	rs2240919	0.6006	1.74E-115	0.1224	2.78E-03	Nerve-Tibial	Height	GNL3
chr3	52801574	52833219	rs2535629	0.4861	2.66E-84	0.1224	2.78E-03	Nerve-Tibial	Autism_spectrum_disorder, Attention_deficithyperactivity, bipolar_disorder, major_depressive and_schizophrenia(combined)	GNL3
chr3	52801574	52838402	rs4687552	0.6277	3.51E-124	0.1224	2.78E-03	Nerve-Tibial	Schizophrenia	GNL3
chr6	110523881	110530560	rs2817782	0.8284	1.02E-246	0.563825	1.15E-03	Esophagus-Mucosa	Electroencephalogram traits	WASF1
chr12	69654100	69661130	rs1373453	0.8279	1.41E-201	0.57338	1.57E-03	Artery-Aorta	Orofacial clefts	LYZ

Supplementary Table S3.4.1: STR in LD with GWAS SNPs more likely to drive association signal. (continued)

chrom	STR start	SNP POS	rsid	r2 (LD)	pval (LD)	Caviar score	eSTR qvalue	Best tissue	Traits	Gene name
chr7	44188112	44196069	rs4607517	0.6040	2.85E-131	0.410155	8.14E-04	Nerve-Tibial	Fasting plasma glucose	GCK
chr7	44188112	44196069	rs4607517	0.6040	2.85E-131	0.410155	8.14E-04	Nerve-Tibial	Fasting glucose-related traits	GCK
chr7	44188112	44196069	rs4607517	0.6040	2.85E-131	0.410155	8.14E-04	Nerve-Tibial	Fasting glucose-related traits (interaction with BMI)	GCK
chr11	953546	941941	rs10751667	0.5955	4.26E-116	0.114603	1.29E-02	Thyroid	Alzheimer's disease (late onset)	AP2A2
chr11	953546	915764	rs7107977	0.4355	4.86E-74	0.114603	1.29E-02	Thyroid	Cannabis use	AP2A2
chr11	810012	762791	rs4963124	0.4316	2.01E-81	0.471544	8.14E-04	Nerve-Tibial	Proteinuria and onic kidney disease	RPLP2
chr11	810012	762791	rs4963124	0.4316	2.01E-81	0.138485	3.42E-03	Brain-Cerebellum	Proteinuria and onic kidney disease	PNPLA2

CHAPTER 4: SUMMARY AND CONCLUSION

In this study, we characterized short tandem repeats (STRs) that contribute to gene expression variation in different contexts, thus enabling their future integration in various studies and facilitating their clinical application. We created a catalog of more than 25 thousand eSTRs (10% FDR) in more than 11 thousand genes (eGenes) across 17 tissues. We identified more than 3400 high confidence causal eSTRs based on their CAVIAR causality score above 10%. These causal eSTRs included several STRs that were previously linked to diseases and more than a dozen eSTRs that may be leading previous GWAS hits.

In Chapter 2, we presented the importance of including STRs in large-scale studies and causal variant studies. We highlighted that mononucleotide STRs also contribute to various processes and mechanisms in the cell, including gene regulation. Out of all eSTRs identified in this study, more than 10 thousand were mononucleotide eSTRs (e-homopolymers). While poly-A were the most abundant eSTRs and STRs overall, their enrichment was not necessarily significant in most predicted regulatory regions when taken together with other eSTRs. e-homopolymers tended to accumulate in the promoter regions of the affected genes, suggesting their implication in gene regulation.

In Chapter 3, we confirmed the contribution of STRs to gene expression variation and characterized these contributions. More than 32% of eSTRs were shared by two or more tissues.
To assure the correctness of our association tests, control tests were performed using permuted samples expression and the results followed a uniform distribution of p-values, as expected. The majority of previously identified eSTRs were replicable. We performed both ANOVA tests, comparing a model where expression variation is explained by eSNPs only on one hand and by both the top eSTR and top eSNP together. In addition, we used the CAVIAR tool to estimate the causality scores of top eSTRs and a set of top 100 e-SNPs. We defined causal eSTRs as all eSTRs for which the addition to the model improved the model and had causality exceeding 10%. More than 14% of

identified eSTRs were causal according to these criteria. The candidate causal eSTRs will need to be validated by other experiments and analysis in the future, but they already included in the set, were some STRs that have already been confirmed as causal to disease phenotypes by previous studies (See Table S3.1.1).

We found that eSTRs were enriched in regulatory regions including promoters, enhancers and chromatin states. Enrichment was even stronger for causal eSTRs. eSTRs were strongly enriched for repeats with CG-rich motifs across all tissues. More specifically, G-rich eSTRs had the potential of forming G-quadruplex secondary structures, which suggested one mechanism by which STRs may be controlling gene expression.

We tested for G-quadruplexes regulatory characteristics in our set of eSTRs with the potential of forming G-quadruplexes, herein G4 eSTRs. We found that: (1)- Both G4 eSTRs and G4 causal/FM eSTRs mainly increased gene expression. (2)- Both G4 eSTRs had lower free energy compared to other eSTRs or causal eSTRs; and, (3)- their free energy decreased, the longer the sequence indicating stability of the structure. These finding suggested that eSTRs may be regulating gene by forming G-quadruplex secondary structures. Our results also suggested that there may be other mechanisms of gene transcription regulation by eSTRs, including transcription factor binding sites formation.

The future vision of STRs study involves improving methods for identification of longer STRs, the use of our catalog in different contexts of research, especially in a clinical application and understanding the mechanisms underlying some association results. For example, we found more than half a dozen high confidence FM eSTRs likely to be the main driver of GWAS hits for Schizophrenia, Height and blood traits.

A few limitations to this study represent in majority unexplored territories and thus are great topics for discussion and further research in themselves. Even though the majority of STRs variation reported to be linked to diseases and other phenotypes in previous studies and cited throughout this study, have been caused by STRs length changes, there's also been evidence of nucleotide-level variation of STRs causing diseases. For that reason, it would be important to consider this aspect on future studies in order to characterize the effect of in-sequence variation of STRs. The future of research includes: exploring the non-linear relationship between STRs length and gene expression, exploring the effects of STR alleles and in-sequence mutation of STRs on gene expression and other phenotypes, and the development of algorithms well suited for such exploration. We also envision the identification of STRs contributing to gene expression in a trans manner.

Overall, a clear understanding of the underlying processes and mechanisms by which eSTRs operate in regulatory regions has the potential of uncovering the molecular machinery affected by causal variants in diseases and thus suggest therapeutics. Research in this direction (like the current study analysis) helps hypothesize and test, explore the roles of variants in today’s less understood complex traits.