Title

Towards Leveraging Short Tandem Repeats for Large Scale Genome-Wide Association Studies

Author

Saini, Shubham

Publication Date

2021

UNIVERSITY OF CALIFORNIA SAN DIEGO

Towards Leveraging Short Tandem Repeats for Large Scale Genome-Wide Association Studies

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

in

Computer Science

by

Shubham Saini

Committee in charge:

Professor Melissa Gymrek, Chair
Professor Vineet Bafna
Professor Yoav Freud
Professor Graham McVicker
Professor Abraham Palmer

2021

The dissertation of Shubham Saini is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To Family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I thank my doctoral advisor Dr. Melissa Gymrek for training me as a scientist and for being a constant source of support and encouragement throughout graduate school. I thank her for having faith in me and welcoming me as one of her early students. Her enthusiasm towards science even during the difficult times proved to be a great source of motivation and helped me to successfully overcome obstacles. This thesis would not be possible without her presence and guidance. I also thank her for introducing me to Team Gymrek, a highly passionate and smart team of scientists, who helped foster collaborative spirit in me.

I would also like to express my gratitude to Dr. Vineet Bafna, who took the time to introduce me to the field of genetics. It was in his class that I first got exposure to this field, and his guidance through my master's thesis is a big factor in my accomplishments. I am thankful to my dissertation committee, Dr. Yoav Freud, Dr. Abraham Palmer, and Dr. Graham McVicker, for their advice and recommendations on the direction of my research.

Most importantly, I thank my family for their unconditional faith and support during all the phases of my life. I would like to make a special mention to my parents, Dr. Meera Saini and Dr. Subhash Saini, for their service towards their patients during this challenging COVID-19 pandemic. I additionally thank my sister, Dr. Divya Saini, for keeping up with me during our childhood years, and for being my confidant in all my mischiefs. I thank my brother-in-law Kartik Shanbhag for his endless support and encouragement during my graduate studies and for sharing a common love of Bacardi Limon. My grandmother deserves a special mention for her constant desire to keep me well-fed. I am also forever indebted to my grandfather for providing a stepping stone and pathway for everyone in my family to pursue their dreams. And last, but not least, I thank my girlfriend Jinneva Santiesteban for her love and support during my graduate school career, and for shaping me as a better human. I would also like to thank Jax, my furry pupper for being the good boy he is.

I finally thank my friends from all phases of my life. I am grateful for Bhavesh Kasliwal,

Osho Bajpai, Vivek Varia, and Shraey Bhatia from my undergrad studies for being a part of all my crazy adventures. I am also extremely thankful to my Ph.D. peers Dr. Saurabh Mogre and Dr. Anant Dhayal who went through similar experiences of moving to a new country for a fresh start, and for coming together in each others' lives when we needed friends the most.

Chapter 2, in full, contains 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.” Nature Communications (2018). I was a primary investigator and author of this paper.

Chapter 3, in full, contains material from Shubham Saini, Brittany S Leger, Jonghun Park, PGC Schizophrenia Working Group, Vineet Bafna, Alon Goren, Melissa Gymrek. ”Genome-wide analysis of the contributions of short tandem repeat variants to schizophrenia risk”, which is currently being prepared for submission for publication of the material. I was the primary investigator and author of this material.

VITA

2014	B. Tech in Computer Science and Engineering, Vellore Institute of Techonology, India
2014-2015	Research Associate, Indraprastha Institute of Information Technology, India
2015-2017	Master of Science in Computer Science, University of California San Diego
2017-2021	Doctor of Philosophy in Computer Science, University of California San Diego

PUBLICATIONS

Nima Mousavi, Jonathan Margoliash, Neha Pusarla, Shubham Saini, Richard Yanicky, Melissa Gymrek, TRTools: a toolkit for genome-wide analysis of tandem repeats, Bioinformatics (2020).

Stephanie Feupe Fotsing, Jonathan Margoliash, Catherine Wang, Shubham Saini, Richard Yanicky, Sharona Shleizer-Burko, Alon Goren, Melissa Gymrek. ”Multi-tissue analysis reveals short tandem repeats as ubiquitous regulators of gene expression and complex traits.” Nature Genetics (2019).

Shubham Saini, Ileena Mitra, Nima Mousavi, Stephanie Feupe Fotsing, Melissa Gymrek. ”A reference haplotype panel for genome-wide imputation of short tandem repeats.” Nature Communications (2018).

Tsering Stobdan, Ali Akbari, Priti Azad, Dan Zhou, Orit Poulsen, Otto Appenzeller, Gustavo F Gonzales, Amalio Telenti, Emily HM Wong, Shubham Saini, Ewen F Kirkness, J Craig Venter, Vineet Bafna, Gabriel G Haddad. ”New insights into the genetic basis of Monge's disease and adaptation to high-altitude.” Molecular Biology and Evolution (2017): msx239.

Afreen Ferdoash, Shubham Saini, Jitesh Khurana, Amarjeet Singh, ”Analytics Driven Operational Efficiency in HVAC Systems” at The 2nd ACM International Conference on Embedded Systems For Energy-Efficient Built Environments, BuildSys 2015, Seoul, South Korea.

Shubham Saini, Pandarasamy Arjunan, Amarjeet Singh, Ullas Nambiar, ”E-Adivino: A Novel Framework for Electricity Consumption Prediction based on Historical Trends” at The 6th ACM International Conference on Future Energy Systems, eEnergy 2015, Bangalore, India.

Amarjeet Singh, Shubham Saini, Sanchit Sharma, Priyank Trivedi, ”Energy Optimization in Commercial Buildings: From Monitoring to Savings Realization” at The 6th ACM International Conference on Future Energy Systems, e-Energy 2015, Bangalore, India.

Shubham Saini, Shraey Bhatia, I. Sumaiya Thaseen, ”sv(M)kmeans - A Hybrid Feature Selection Technique for Reducing False Positives in Network Anomaly Detection” at The 20th International Conference on Management of Data, COMAD 2014, Hyderabad, India.

Bhavesh Kasliwal*, Shraey Bhatia, Shubham Saini*, I.Sumaiya Thaseen, Ch.Aswani Kumar, ”A Hybrid Anomaly Detection Model using G-LDA” at The 4th IEEE International Advance Computing Conference, IACC 2014, Gurgaon, India.

ABSTRACT OF THE DISSERTATION

Towards Leveraging Short Tandem Repeats for Large Scale Genome-Wide Association Studies

by

Shubham Saini

Doctor of Philosophy in Computer Science

University of California San Diego, 2021

Professor Melissa Gymrek, Chair

Most of the efforts in human genetics are directed towards identifying and characterizing genetic variants that impact human traits, achieved by examining relationships between traits and variants. A Genome Wide Association Study (GWAS) quantifies statistical association between genetic variation and phenotypes. These statistical associations can tell us about the biological mechanisms affecting the phenotype and can allow us to predict the phenotype from genetic information in a clinical setting. However, the majority of GWAS datasets have been generated with commodity genotype arrays of single-nucleotide polymorphism (SNP) that fail to explain the majority of heritability for many complex traits even with large sample sizes.

One compelling hypothesis explaining the missing heritability dilemma is that complex variants, such as multi-allelic repeats not in strong linkage with common SNPs, are important drivers of complex traits but are largely invisible to current analyses. Short tandem repeats (STRs), consisting of repeated motifs of 1–6bp in tandem, comprise more than 3% of the human genome. Multiple lines of evidence support a role of STRs in complex traits, particularly in neurological and psychiatric phenotypes. However, existing technologies have not allowed for systematic STR association studies.

To overcome these challenges, we recently generated a reference STR+SNP haplotype panel that enables imputation of STR genotypes into SNP genotypes available for most GWAS cohorts. Our imputation pipeline achieves a high concordance and can be used to impute nearly 500,000 STRs genome-wide. Next, we leveraged our reference haplotype panel to impute STRs into GWAS data for more than 50,000 samples from the Psychiatric Genomics Consortium (PGC) to perform a genome-wide analysis of associations between STR lengths and schizophrenia.

In this dissertation, I demonstrate an end-to-end pipeline for conducting large biobank scale GWAS using STRs that serves as one of the initial studies which researchers can find useful for incorporating complex variants into their analysis.

Chapter 1

Introduction

1.1 The Human DNA

The human genome is a 3 billion base pairs long nucleic acid sequence, encoded as 23 pairs of chromosome. Humans are diploid, meaning they carry two copies of each chromosome, with one inherited each from the mother and father. The Human Genome Project [1, 2] produced the first complete reference of the human genome in 2001. Now, nearly two decades later, hundreds of thousands of human genomes have been completely sequenced, with resulting data being used for biomedicine, forensics, anthropology and other sciences.

1.1.1 Classes of Genetic Variation

A **variant** refers to a specific region of the genome that differs between two individuals. Different versions of the same variant are known as **alleles**. For example one individual may have the allele C at a position of the genome, while another individual may have allele G. Figure 1.1 gives cartoon representation of these concepts. Mutations are the original source of genetic variation. A mutation in an individual is said to have occurred when there is an allele change (a

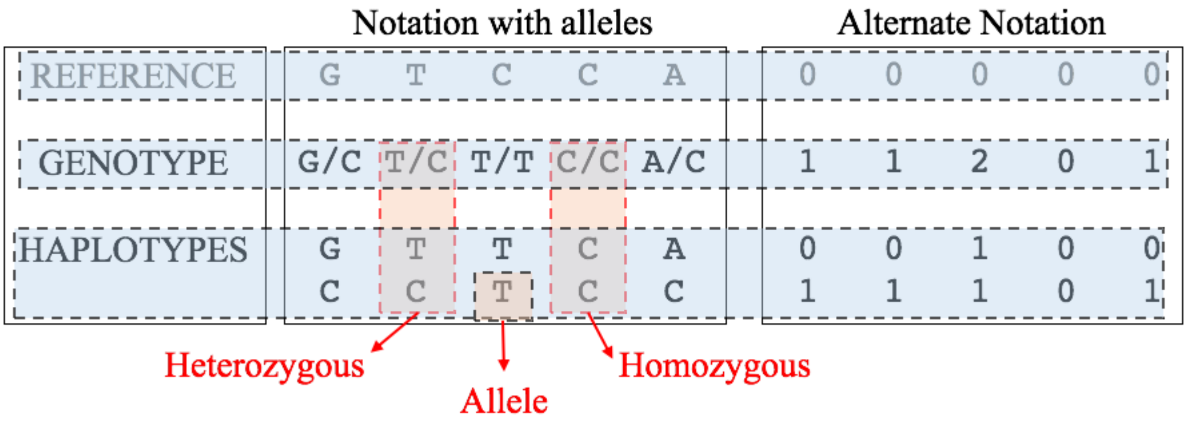

Notation with alleles
Alternate Notation
REFERENCE G T C C A 0 0 0 0 0
GENOTYPE G/C T/C T/T C/C A/C 1 1 2 0 1
HAPLOTYPES G T T C A 0 0 1 0 0
C C T C C 1 1 1 0 1
Heterozygous
Allele
Homozygous

Figure 1.1: A cartoon representation of a genotype, a haplotype, an allele, a heterozygous variant, and a homozygous variant. The alternate notation assumes the reference allele as 0 and the alternate allele as 1. The genotype can be expressed as the sum of the two alleles.

change at a site in the genome) while inheriting DNA from the parents.

Human genomes have different types of genetic variants. The simplest of them are:

1. Single base pair mutations, also called Single Nucleotide Polymorphisms (**SNPs**) result from mutations of a single base pair.

2. Insertion or Deletion, also called InDels, result from insertion of deletion of a sequence of bair pairs as compared to the reference human genome.

3. Structural Variations (SV) occur over larger DNA regions. This category includes both Copy Number Variations (CNVs) and Rearrangements of regions of the genome that may span thousands of base pairs.

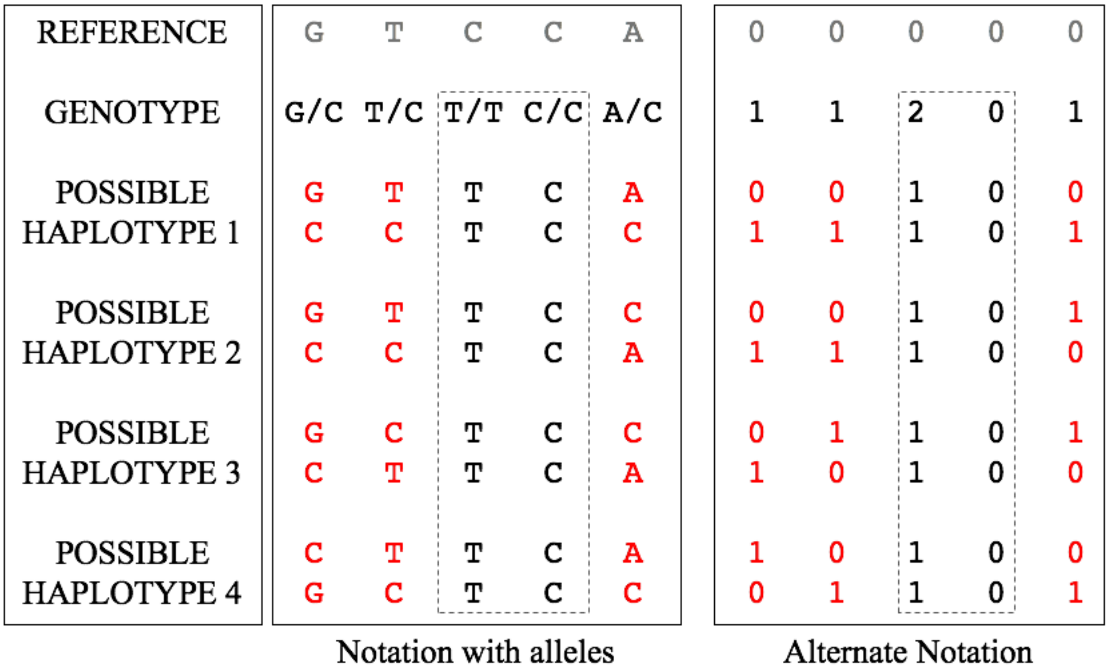
The probability of a mutation happening at a site is referred to as the average mutation rate. Mutation rates are different for different genetic variant types. On an average, 50 SNP mutations, 3 InDel mutations,and 0.2 structural variant mutations are observed genome-wide per person [3]. These mutation rates imply that most human genomes are highly (around 99.5%) similar, and differ at a small fraction of sites. With a total of 3 billion base pairs across all 23 chromosomes combined, the number of differences expected between two human genomes is around 20 million base pairs [3], or 1 SNP every 150 base pairs.

1.1.2 Short Tandem Repeats

Short tandem repeats (STRs), consisting of repeated motifs of 1–6 bp in tandem, comprise more than 3% of the human genome. Multiple lines of evidence support a role of STRs in complex traits, particularly in neurological and psychiatric phenotypes. Due to their rapid mutation rates, STRs exhibit high rates of heterozygosity and likely contribute at least as many de novo mutations per generation as SNPs. Furthermore, STRs have been shown to play a significant role in regulating gene expression, splicing, and DNA methylation. Intriguingly, more than 30 Mendelian disorders are caused by STR expansions via a range of mechanisms, including polyglutamine aggregation (Huntington's Disease, ataxias), hypermethylation (Fragile-X Syndrome), and RNA toxicity (ALS/FTD). Furthermore, causal STRs driving existing GWAS signals have already been identified.

1.2 Genotype Phasing and Imputation

The human genome is diploid in nature, with one copy of each chromosome inherited from the mother and one copy from the father. A **genotype** refers to an unordered, or **unphased** pair of alleles at a single position, one from each chromosome, with no information about the parental origin of the alleles. A **haplotype** refers to a **phased** sequence along a single chromosome with the same parental origin. More precisely, a haplotype is a string of length k with each character an element of $\{A,C,G,T\}$. We represent a haplotype as $\{A,C,G,T\}^k$. Because the vast majority of positions are bi-alleic, i.e. we have a maximum of two possible alleles, a haplotype can also be represented as $\{0,1\}^k$. Due to limitations of current genome sequencing technologies, it is difficult to accurately separate the pair of chromosomes to obtain phased genotypes.



	Notation with alleles					Alternate Notation				
REFERENCE	G	T	C	C	A	0	0	0	0	0
GENOTYPE	G/C	T/C	T/T	C/C	A/C	1	1	2	0	1
POSSIBLE HAPLOTYPE 1	G	T	T	C	A	0	0	1	0	0
	C	C	T	C	C	1	1	1	0	1
POSSIBLE HAPLOTYPE 2	G	T	T	C	C	0	0	1	0	1
	C	C	T	C	A	1	1	1	0	0
POSSIBLE HAPLOTYPE 3	G	C	T	C	C	0	1	1	0	1
	C	T	T	C	A	1	0	1	0	0
POSSIBLE HAPLOTYPE 4	C	T	T	C	A	1	0	1	0	0
	G	C	T	C	C	0	1	1	0	1

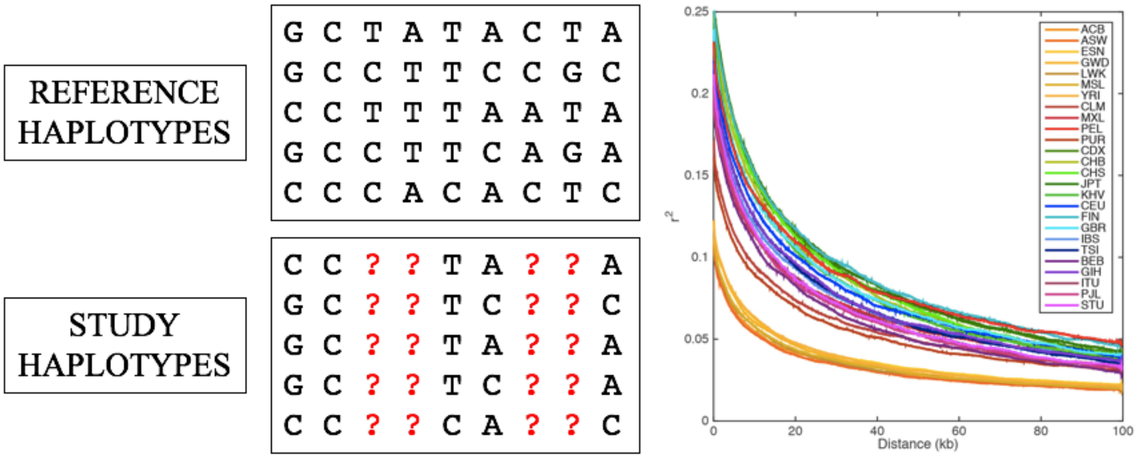
Figure 1.2: Genotype phasing: showing some of the possible haplotypes from the observed genotypes. The variants within the dashed lines are homozygous, hence we do not need to phase them. The alternate notation assumes the reference allele as 0 and the alternate allele as 1. The genotype can be expressed as the sum of the two alleles.

We define the **genotype phasing** problem as follows: For a length k unphased genotype string, there are $2^{(k-1)}$ unique haplotype pairs that may explain the genotype. The objective of the genotype phasing problem is to recover the two haplotypes out of $2^{(k-1)}$ possible haplotypes given an unphased genotype string. The problem can be formulated as follows:

$$\textbf{Input}: \text{Genotype } G = (g_1, g_2, g_3, ..., g_k),\ where\ g_i \in \{0,1,2\}\ for\ 1 <= i <= k$$

$$\textbf{Output}: \text{Pair of haplotypes } H = \{h_1, h_2\},\ where\ h_1, h_2 \in \{0,1\}^k$$

For many applications, haplotypes are more informative than unphased genotypes. Haplotypes contain the history of a variant. We can use haplotypes for detecting natural selection,



```
G C T A T A C T A
G C C T T C C G C
C C T T T A A T A
G C C T T C A G A
C C C A C A C T C
```

```
C C ? ? T A ? ? A
G C ? ? T C ? ? C
G C ? ? T A ? ? A
G C ? ? T C ? ? A
C C ? ? C A ? ? C
```

Figure 1.3: Genotype imputation: process of inferring the missing variants in the study samples using the reference haplotypes. Genotype imputation makes use of correlation between nearby variants, as a result of Linkage Disequilibrium (LD). The figure on the right shows the decaying correlation among different population groups as the genetic distance between two variants increase.

i.e. finding how long until a variant becomes common in a population. This requires long range haplotypes across large samples. Another important application of haplotypes of great medical significance is analyzing effect of mutations. Gene function is determined by, among other things, mutations on the two copies of the chromosome. If multiple mutations occur on the same copy of the chromosome (also called cis mutations) only one gene is altered, whereas if mutations occur on different copies of the chromosome (trans mutations), both the genes are altered. If at least one copy of the gene is required, then multiple mutations in cis may be harmless, whereas mutations in trans (known as compound heterozygosity) may lead to disease. Without phase information, we cannot distinguish the difference.

Finally, a closely related problem to genotype phasing is **genotype imputation** that require phased genotypes as an input. Early days of the Human Genome Project [1, 2] showed a strong correlation between nearby SNPs, as a result of Linkage Disequilibrium (LD). Genotype imputation is used to fill in the missing gaps in the genome by utilizing the Linkage Disequilibrium (LD) structure between nearby sites. Since Whole Genome Sequencing (WGS), the process of determining the complete human genome at a single time is expensive when done across hundreds of thousands of samples, geneticists make use of small subset of markers as this subset provides information on the missing markers as well. The idea that small subset of variants provide useful information about other variants forms the basis of genetic linkage studies that typically use $<$ 10000 markers to survey entire human genome. We formulate the genotype imputation problem as follows:

$$\textbf{Input} : \text{Reference Haplotypes } H_i = \{H_{i1}, H_{i2}, ..., H_{il}\}$$

$$\text{Study Haplotypes } h_j = \{h_{jm}\}$$

$$\textbf{Output} :\text{Study Haplotypes } h'_j = \{h_{j1}, h_{j2}, ..., h_{jl}\}$$

$$\text{where } H_{il}, h_{jm} \in \{0,1\}, l \in L, m \in M, M \subset L$$

$$\text{i.e. we infer } h_{jk},\ k \in L - M$$

1.3 Genome-Wide Association Studies (GWAS)

Most of the efforts in human genetics are directed towards identifying and characterizing genetic variants that impact human traits, achieved by examining relationships between traits and variants. A phenotype, also referred to as a trait, is a measured property of an individual. A Genome Wide Association Studies (GWAS) quantifies statistical association between genetic variation and phenotypes. These statistical associations can tell us about the biological mechanisms affecting the phenotype and can allow us to predict the phenotype from genetic information in a clinical setting.

The two main varieties of GWAS are studying quantitative traits or disease phenotypes. Quantitative trait GWAS involves stuyding the association between genetic variants and quantitative traits, for example height or BMI. Almost all quantitative trait GWAS are performed using an additive model that relies on the assumption that the means of the phenotypes depend additively on the number of minor alleles in the genotype of the individuals. This can be achieved by fitting a linear model $y = \mu + x\beta + \varepsilon$ where y is the phenotype, x is the genotype (0,1 or 2), μ is the mean of genotype 0 and β is the effect of each copy of minor allele on the mean phenotype. On the other hand, disease trait (or case-control traits) GWAS involve studying the association between genetic variants and binary phenotypes like disease status. A generalized model like logistic regression is fit when the phenotype is binary. The model explains the logarithm of the odds of the disease by the genotype. Like a quantitative GWAS, we use an additive model and estimate the effect size β that gives us the odds-ratio (or relative increase in odds between two genotypes) on a log scale. $log\left(\frac{Pr(Y=1|X=x)}{Pr(Y=0|X=x)}\right) = \mu + x\beta$. Here μ is the logarithm of odds for genotype 0 and β is the log of odds ratio (logOR) between genotype 1 and 0 (and $exp(\beta)$ is the corresponding odds ratio)

GWAS based on linkage disequilibrium (LD) uses a small number of SNPs that characterize 80% of the genetic variation in a given population [4]. While these studies can identify SNPs in correlation with the traits under study, they are just proxies for causal SNPs. In order to identify the causal variants for a given trait, genotype imputation may be used to infer the missing genotype markers using subset of available markers. Imputation works by using haplotype patterns in reference data of comprehensively typed samples to predict the missing variants in the study samples.

1.4 Towards Leveraging Short Tandem Repeats for Large-Scale Genome-Wide Association Studies

While Genome-wide association studies (GWAS) have become increasingly successful at identifying genetic loci significantly associated with complex traits in humans, common SNPs still fail to explain the majority of heritability for most complex traits. One possible reason for this is that complex variants, such as multi-allelic repeats not in strong LD with common SNPs are important drivers of complex traits.

Short tandem repeats (STRs), are 1-6bp repeats that comprise more than 3% of the human genome. Multiple lines of evidence support a role of STRs in complex traits[5, 6, 7], particularly in neurological and psychiatric phenotypes. Due to their rapid mutation rates[8], STRs exhibit high rates of heterozygosity[9] and likely contribute at least as many de novo mutations per generation as SNPs[10, 11]. Furthermore, STRs have been shown to play a significant role in regulating gene expression[12, 13] , splicing[14, 15, 16], and DNA methylation[13]. Intriguingly, more than 30 Mendelian disorders are caused by STR expansions via a range of mechanisms, including polyglutamine aggregation (Huntington's Disease, ataxias[17]), hypermethylation (Fragile-X Syndrome[18]), and RNA toxicity (ALS/FTD[19]). Furthermore, causal STRs driving existing GWAS signals have already been identified[20].

Although next-generation sequencing (NGS) can be used to directly genotype short STRs, it is too expensive to perform on sufficiently large sample sizes. An alternative approach is to impute STRs into existing SNP array datasets. Previous studies have demonstrated that STRs are often in significant LD with nearby SNPs[21, 22, 23] and found that STRs and SNPs provide complementary information about the evolutionary history of a genomic region. Despite widespread SNP-STR LD, statistical phasing of STRs and SNPs is challenging for several reasons: SNP-STR LD is notably weaker than SNP-SNP LD[23] due to the rapid mutation rates[8, 24] and high prevalence of recurrent mutations in STRs. As a result, the relationship between STR repeat number and SNP haplotype can be complex: the same STR allele may be present on multiple SNP haplotypes. On the other hand, a single SNP haplotype may harbor multiple distinct STR alleles. Furthermore, LD patterns at STRs vary widely as a function of properties of the repeat, such as the repeat unit length, mutation rate, and mutation step size[23]. Finally, STRs are prone to genotyping errors induced during PCR amplification[25, 26], further ambiguating phase information.

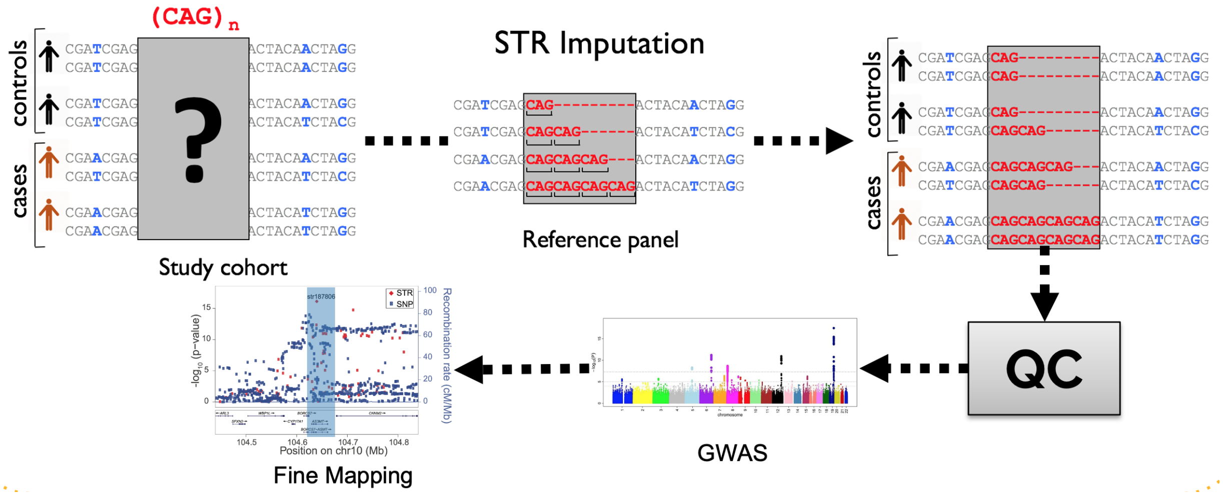


Figure 1.4: End-to-end pipeline for conducting large biobank scale GWAS using STRs.

In this dissertation, I demonstrate an end-to-end pipeline for conducting large biobank scale GWAS using STRs that serves as one of the initial studies which researchers can find useful for incorporating complex variants into their analysis (Figure 1.4). In Chapter 2, I present our recently generated reference STR+SNP haplotype panel that enables imputation of STR genotypes into SNP genotypes available for most GWAS cohorts. Our imputation pipeline achieves an average imputed genotype concordance of 97% on European samples and can be used to impute nearly 500,000 STRs genome-wide. Chapter 2, in full, contains 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.” Nature Communications (2018). I was a primary investigator and author of this paper. Finally, in Chapter 3, I present how we leveraged our reference haplotype panel to impute STRs into GWAS data for more than 50,000 samples from the Psychiatric Genomics Consortium (PGC) to perform a genome-wide analysis of associations between STR lengths and schizophrenia. Chapter 3, in full, contains material from Shubham Saini, Brittany S Leger, Jonghun Park, PGC Schizophrenia Working Group, Vineet Bafna, Alon Goren, Melissa Gymrek. ”Genome-wide analysis of the contributions of short tandem repeat variants to schizophrenia risk”, which is currently being prepared for submission for publication of the material. I was the primary investigator and author of this material.

Bibliography

[1] International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001.

[2] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[3] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[4] Michael Krawczak. *Genotype Imputation*, pages 1–7. American Cancer Society, 2015.

[5] A. J. Hannan. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. *Trends Genet*, 26(2):59–65, 2010.

[6] A. J. Hannan. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet*, 19(5):286–298, 2018.

[7] M. O. Press, K. D. Carlson, and C. Queitsch. The overdue promise of short tandem repeat variation for heritability. *Trends Genet*, 30(11):504–12, 2014.

[8] J. X. Sun, A. Helgason, G. Masson, S. S. Ebenesersdottir, H. Li, S. Mallick, S. Gnerre, N. Patterson, A. Kong, D. Reich, and K. Stefansson. A direct characterization of human mutation based on microsatellites. *Nat Genet*, 44(10):1161–5, 2012.

[9] T. Willems, M. Gymrek, G. Highnam, Consortium Genomes Project, D. Mittelman, and Y. Erlich. The landscape of human str variation. *Genome Res*, 24(11):1894–904, 2014.

[10] R. Acuna-Hidalgo, J. A. Veltman, and A. Hoischen. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol*, 17(1):241, 2016.

[11] T. Willems, M. Gymrek, G. D. Poznik, C. Tyler-Smith, Y. Group Genomes Project Chromosome, and Y. Erlich. Population-scale sequencing data enable precise estimates of y-str mutation rates. *Am J Hum Genet*, 98(5):919–933, 2016.

[12] M. Gymrek, T. Willems, A. Guilmatre, H. Zeng, B. Markus, S. Georgiev, M. J. Daly, A. L. Price, J. K. Pritchard, A. J. Sharp, and Y. Erlich. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet*, 48(1):22–9, 2016.

[13] J. Quilez, A. Guilmatre, P. Garg, G. Highnam, M. Gymrek, Y. Erlich, R. S. Joshi, D. Mittelman, and A. J. Sharp. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and dna methylation in humans. *Nucleic Acids Res*, 44(8):3750–62, 2016.

[14] J. R. Tollervey, T. Curk, B. Rogelj, M. Briese, M. Cereda, M. Kayikci, J. Konig, T. Hortobagyi, A. L. Nishimura, V. Zupunski, R. Patani, S. Chandran, G. Rot, B. Zupan, C. E. Shaw, and J. Ule. Characterizing the rna targets and position-dependent splicing regulation by tdp-43. *Nat Neurosci*, 14(4):452–8, 2011.

[15] J. Hui, L. H. Hung, M. Heiner, S. Schreiner, N. Neumuller, G. Reither, S. A. Haas, and A. Bindereif. Intronic ca-repeat and ca-rich elements: a new class of regulators of mammalian alternative splicing. *EMBO J*, 24(11):1988–98, 2005.

[16] T. W. Hefferon, J. D. Groman, C. E. Yurk, and G. R. Cutting. 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*, 101(10):3504–9, 2004.

[17] S. M. Mirkin. Expandable dna repeats and human disease. *Nature*, 447(7147):932–40, 2007.

[18] J. S. Sutcliffe, D. L. Nelson, F. Zhang, M. Pieretti, C. T. Caskey, D. Saxe, and S. T. Warren. Dna methylation represses fmr-1 transcription in fragile x syndrome. *Hum Mol Genet*, 1(6):397–400, 1992.

[19] M. van Blitterswijk, M. DeJesus-Hernandez, and R. Rademakers. How do c9orf72 repeat expansions cause amyotrophic lateral sclerosis and frontotemporal dementia: can we learn from other noncoding repeat expansion disorders? *Curr Opin Neurol*, 25(6):689–700, 2012.

[20] T. G. Grunewald, V. Bernard, P. Gilardi-Hebenstreit, V. Raynal, D. Surdez, M. M. Aynaud, O. Mirabeau, F. Cidre-Aranaz, F. Tirode, S. Zaidi, G. Perot, A. H. Jonker, C. Lucchesi, M. C. Le Deley, O. Oberlin, P. Marec-Berard, A. S. Veron, S. Reynaud, E. Lapouble, V. Boeva, T. Rio Frio, J. Alonso, S. Bhatia, G. Pierron, G. Cancel-Tassin, O. Cussenot, D. G. Cox, L. M. Morton, M. J. Machiela, S. J. Chanock, P. Charnay, and O. Delattre. Chimeric ewsr1-fli1 regulates the ewing sarcoma susceptibility gene egr2 via a ggaa microsatellite. *Nat Genet*, 47(9):1073–8, 2015.

[21] J. L. Mountain, A. Knight, M. Jobin, C. Gignoux, A. Miller, A. A. Lin, and P. A. Underhill. Snpstrs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome Res*, 12(11):1766–72, 2002.

[22] S. A. Tishkoff, E. Dietzsch, W. Speed, A. J. Pakstis, J. R. Kidd, K. Cheung, B. Bonne-Tamir, A. S. Santachiara-Benerecetti, P. Moral, and M. Krings. Global patterns of linkage disequilibrium at the cd4 locus and modern human origins. *Science*, 271(5254):1380–7, 1996.

[23] B. A. Payseur, M. Place, and J. L. Weber. Linkage disequilibrium between strps and snps across the human genome. *Am J Hum Genet*, 82(5):1039–50, 2008.

[24] M. Gymrek, T. Willems, D. Reich, and Y. Erlich. Interpreting short tandem repeat variations in humans using mutational constraint. *Nat Genet*, 49(10):1495–1501, 2017.

[25] Y. Lai, D. Shinde, N. Arnheim, and F. Sun. The mutation process of microsatellites during the polymerase chain reaction. *J Comput Biol*, 10(2):143–55, 2003.

[26] Y. Lai and F. Sun. Microsatellite mutations during the polymerase chain reaction: mean field approximations and their applications. *J Theor Biol*, 224(1):127–37, 2003.

Chapter 2

A reference haplotype panel for genome-wide imputation of short tandem repeats

2.1 Introduction

Genome-wide association studies (GWAS) have become increasingly successful at identifying genetic loci significantly associated with complex traits in humans, largely due to the enormous growth in available sample sizes[1, 2, 3]. Hundreds of thousands of individuals have been genotyped using commodity genotyping arrays. These arrays take advantage of the correlation structure between nearby variants induced by linkage disequilibrium (LD), which allows genome-wide imputation based on genotypes of only a small subset of loci[4]. However, GWAS based on single-nucleotide polymorphism (SNP) associations face important limitations. Even with sample sizes of up to 100,000 individuals, common SNPs still fail to explain the majority of heritability for many complex traits[2, 5].

One compelling hypothesis explaining the missing heritability dilemma is that complex variants, such as multi-allelic repeats not in strong LD with common SNPs, are important drivers of complex traits but are largely invisible to current analyses. Indeed, dissection of the strongest schizophrenia association, located in the major histocompatibility complex, revealed a poorly tagged polymorphic copy number variant (CNV) to be the causal variant[6]. The signal could not be localized to a single SNP and could only be explained after deep characterization of the underlying CNV. This and subsequent discoveries[7, 8] highlight the importance of considering alternative variant classes.

Short tandem repeats (STRs), consisting of repeated motifs of 1–6bp in tandem, comprise more than 3% of the human genome[9]. Multiple lines of evidence support a role of STRs in complex traits[10, 11, 12], particularly in neurological and psychiatric phenotypes. Due to their rapid mutation rates[13], STRs exhibit high rates of heterozygosity[14] and likely contribute at least as many de novo mutations per generation as SNPs[15, 16]. Furthermore, STRs have been shown to play a significant role in regulating gene expression[17, 18] , splicing[19, 20, 21], and DNA methylation[18]. Intriguingly, more than 30 Mendelian disorders are caused by STR expansions via a range of mechanisms, including polyglutamine aggregation (Huntington's Disease, ataxias[22]), hypermethylation (Fragile-X Syndrome[23]), and RNA toxicity (ALS/FTD[24]). Furthermore, causal STRs driving existing GWAS signals have already been identified[25].

Existing technologies have not allowed for systematic STR association studies. Next-generation sequencing (NGS) can be used to directly genotype short STRs, but NGS is still too expensive to perform on sufficiently large cohorts for GWAS of most complex traits. An alternative approach is to impute STRs into existing SNP array datasets. Previous studies have demonstrated that STRs are often in significant LD with nearby SNPs[26, 27, 28] and found that STRs and SNPs provide complementary information about the evolutionary history of a genomic region. Despite widespread SNP-STR LD, statistical phasing of STRs and SNPs is challenging for several reasons: SNP-STR LD is notably weaker than SNP-SNP LD[28] due to the rapid mutation rates[13, 29] and high prevalence of recurrent mutations in STRs. As a result, the relationship between STR repeat number and SNP haplotype can be complex: the same STR allele may be present on multiple SNP haplotypes. On the other hand, a single SNP haplotype may harbor multiple distinct STR alleles. Furthermore, LD patterns at STRs vary widely as a function of properties of the repeat, such as the repeat unit length, mutation rate, and mutation step size[28]. Finally, STRs are prone to genotyping errors induced during PCR amplification[30, 31], further ambiguating phase information.

Sequencing related samples allows haplotype resolution by directly tracing inheritance patterns. The recent generation of deep NGS using PCR-free protocols for hundreds of nuclear families in combination with accurate tools for genotyping STRs from NGS[32] now enables applying this technique genome-wide. Here, we profile STRs in 479 families and use pedigree information to phase STR genotypes onto SNP haplotypes to create a genome-wide reference for imputation. We use this panel to impute STRs into an external dataset of similar ethnic background with average 97% concordance with observed STR genotypes. Imputation accuracy varies across STRs, ranging from nearly perfect concordance at bi-allelic STRs to around 70% for highly polymorphic forensic markers. We show that STR imputation achieves greater power than individual SNPs to detect underlying STR associations and demonstrate the utility of our panel by detecting STRs not previously known to be associated with gene expression. Finally, we impute genotypes at STRs previously implicated in human disorders and show that we could accurately identify specific SNP haplotypes associated with long normal alleles most at risk for expansion.

To facilitate use by the community, we release a phased SNP+STR haplotype panel for samples genotyped as part of the 1000 Genomes Project (see Data availability). This resource will enable large-scale studies of STR associations in hundreds of thousands of available SNP datasets, and will likely yield significant new insights into complex traits.

2.2 Results

2.2.1 A catalog of STR variation in 479 families

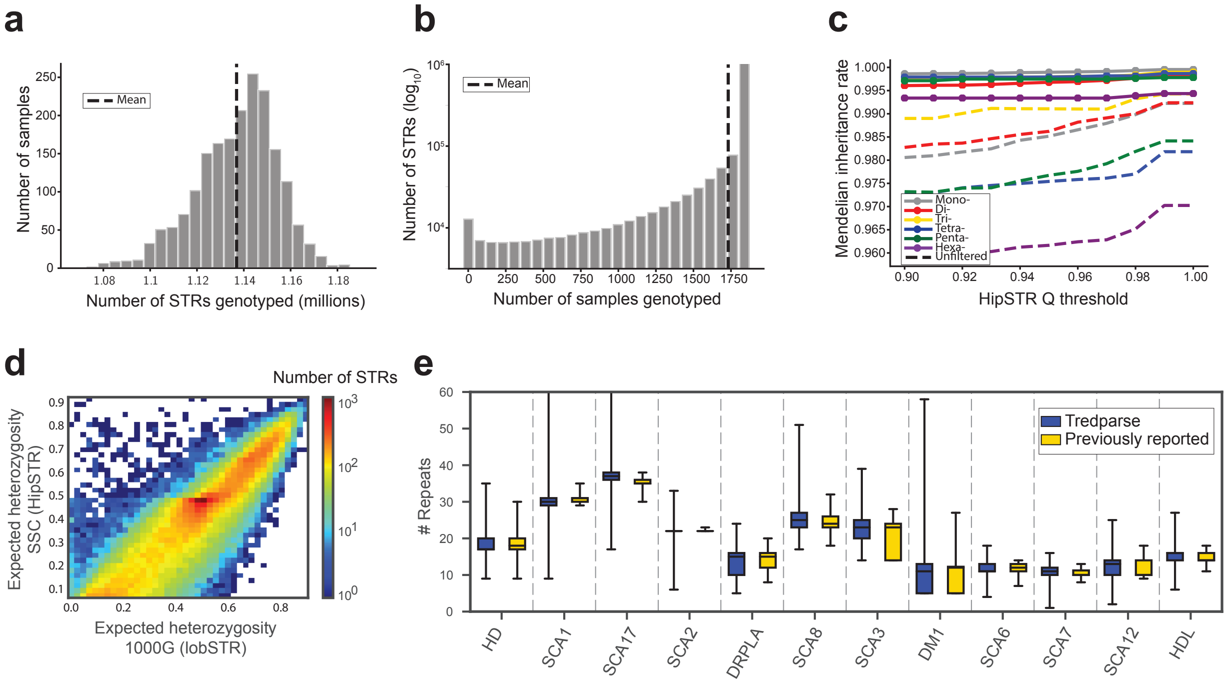


Figure 2.1: **A deep catalog of STR variation in the SSC cohort. a.** Number of STRs called per sample. Dashed line represents the mean of 1.14 million STRs per sample. **b.** Call rate per locus. Dashed line represents the mean call rate of 90%. **c.** Mendelian inheritance rate at filtered vs. unfiltered STRs. The x-axis gives the posterior genotype score (Q) returned by HipSTR. The y-axis gives the average Mendelian inheritance rate for each bin across all calls on chromosome 21. STRs that were homozygous for the reference allele in all members of a family were removed. Colors represent different motif lengths. **d.** Per-STR expected heterozygosity in SSC vs. 1000 Genomes. Only STRs with expected heterozygosity ¿0.095 in SSC are included. Color scale gives the log10 number of STRs represented in each bin. **e.** Allele frequency distributions at pathogenic STRs obtained in SSC samples vs. previously reported normal alleles. Blue=SSC, Gold=Previously reported. Boxes span the interquartile range and horizontal lines give the medians. Whiskers extend to the minimum and maximum data points. The y-axis gives the number of repeat units. Sources of previously reported allele frequencies are described in detail in Methods. HD Huntington's disease, SCA spinocerebellar ataxia, DRPLA Dentatorubral-pallidoluysian atrophy, DM1 myotonic dystrophy type 1, HDL Huntington's disease-like 2

We first generated a genome-wide catalog of STR variation in a cohort of families included in the Simons Simplex Collection (SSC) (see URLs). We focused on 1916 individuals from

479 family quads (parents and two children) that were sequenced to an average depth of 30x using Illumina's PCR-free protocol. Based on comparison to 1000 Genomes Project samples, we estimated the cohort to consist primarily of Europeans (83%), with 2.0%, 9.0%, and 3.6% of East Asian, South Asian, and African ancestry, respectively (Supplementary Fig. S2.1). We used HipSTR[32] to profile autosomal STRs in each sample. HipSTR takes aligned reads and a reference set of STRs as input and outputs maximum likelihood diploid genotypes for each STR in the genome. While HipSTR infers the entire sequence of each STR allele, we focus here on differences in repeat copy number rather than sequence variation within the repeat itself. To maximize the quality of genotype calls, individuals were genotyped jointly with HipSTR's multi-sample calling mode using phased SNP genotypes and aligned reads as input (Methods). Multi-sample calling allows HipSTR to leverage information on haplotypes discovered across all samples in the dataset to estimate per-locus error parameters and output genotype likelihoods for each possible diploid genotype. Notably, our HipSTR catalog excluded most known STRs implicated in expansion disorders such as Huntington's Disease and hereditary ataxias, since even the normal allele range for these STRs is above or near the length of Illumina reads[33, 34, 35, 36]. To supplement our panel, we applied a second STR genotyper, Tredparse[37], to genotype a targeted set of known pathogenic STRs in our cohort (Supplementary Table S2.1). Tredparse incorporates multiple features of paired-end reads to estimate the size of repeats longer than the read length. For seven STRs called by both Tredparse and HipSTR, Tredparse genotypes were used for downstream analyses.

An average of 1.14 million STRs passed HipSTR's default filtering settings in each sample (Fig. 2.1a). We obtained at least one call for 97% of all STRs in the HipSTR reference of 1.6 million STRs and for 15 of 25 STRs in the Tredparse reference with an average overall call rate of 90% (Fig. 2.1b). We applied additional stringent genotype quality filters to ensure accurate calls for downstream phasing and imputation analysis. STRs overlapping segmental duplications, with call rates ¡80%, or with genotype frequencies unexpected under Hardy-Weinberg Equilibrium were removed (Methods). We further removed STRs with low expected heterozygosity (¡0.095) to restrict analysis to polymorphic STRs. We found that these filters increased the quality of our calls, as evidenced by the average Mendelian inheritance rate of 99.8% and 97.9% at STRs that passed and failed quality filters, respectively (Fig. 2.1c). After filtering, 453,671 and 9 STRs from the HipSTR and Tredparse panels, respectively, remained in our catalog.

We further assessed the quality of our STR genotypes by comparing patterns of variation from SSC to previous catalogs of STR variation obtained using a distinct set of samples and STR genotyping methods. We found that per-locus heterozygosities (Methods) were highly concordant with a catalog generated from the 1000 Genomes[38] Project data using lobSTR[39]. (Pearson r=0.96; $p < 10^{200}$; n=386,100) (Fig. 2.1d). Allele length distributions at known pathogenic STRs observed in SSC matched closely to previously reported normal allele frequencies at each STR (Fig. 2.1e). For STRs genotyped both by HipSTR and Tredparse, estimated repeat lengths were highly concordant (average concordance 99.4%, Supplementary Table S2.1). Overall, these results show that our catalog consists of robust STR genotypes suitable for downstream phasing and imputation analysis.

2.2.2 A genome-wide SNP+STR haplotype reference panel

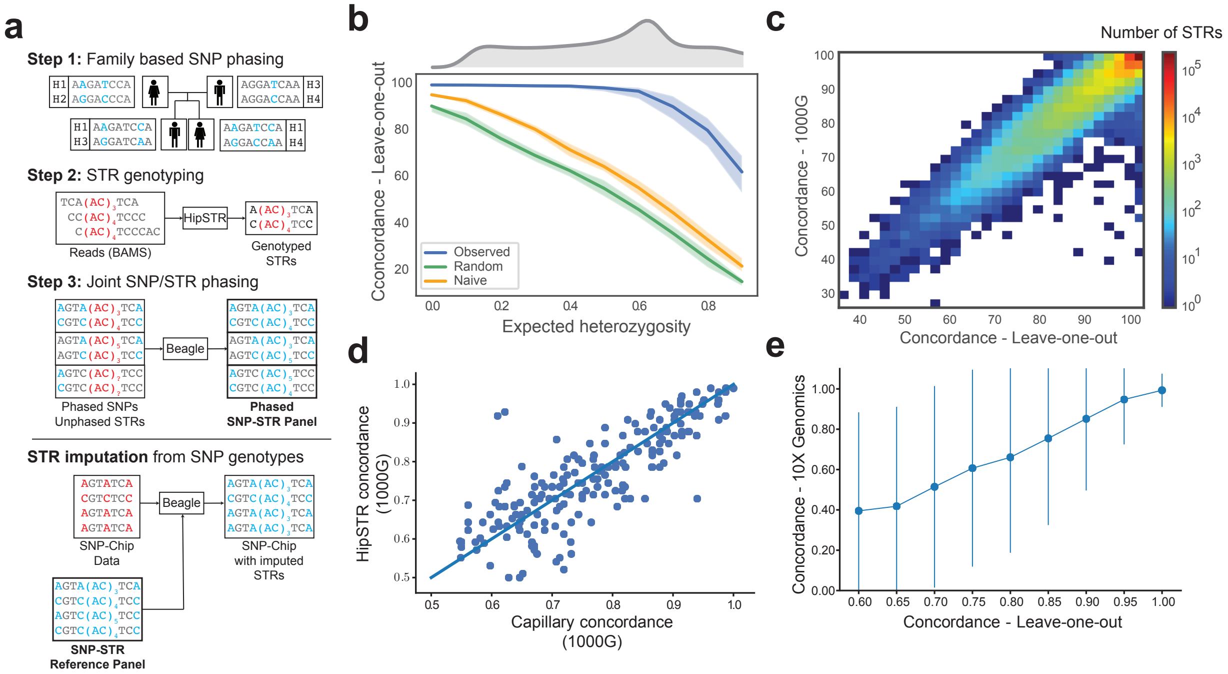


Figure 2.2: **Creating a reference SNP-STR haplotype panel. a.** Schematic of phasing pipeline in the SSC cohort. To create the phased panel, STR genotypes were placed onto phased SNP haplotypes using Beagle. Any missing STR genotypes were imputed. The resulting panel was then used for downstream imputation from orthogonal SNP genotypes. Blue and red denote phased and unphased variants, respectively. Positions in gray are homozygous. **b.** Concordance of imputed STR genotypes vs. expected heterozygosity. Blue denotes observed per-locus values, green denotes values expected under a random model and orange denotes values expected under a naive model. Solid lines give median values for each bin and filled areas span the 25th to 75th percentile of values in each bin. x-axis values were binned by 0.1. Upper gray plot gives the distribution of expected heterozygosity values in our panel. Concordance values are based on the leave-one-out analysis in the SSC cohort. **c.** Per-locus imputation concordance in SSC vs. 1000 Genomes cohorts. Color scale gives the $\log_{10}$ number of STRs represented in each bin. Concordance values are based on the subset of samples from the 1000 Genomes deep WGS cohort with European ancestry. **d.** Per-locus imputation concordance using HipSTR vs. capillary electrophoresis genotypes. Each dot represents one STR. The x-axis and y-axis give imputation concordance using capillary electrophoresis or HipSTR genotypes as a ground truth, respectively. Concordance was measured in separate sets of 1000 Genomes European samples for each technology. **e.** Concordance of imputed vs. 10X STR genotypes in NA12878 stratified by concordance in SSC. STRs were binned by concordance value based on the leave-one-out analysis. Concordance in NA12878 was measured across all STRs in each bin. Dots give mean values for each bin and lines denote ±1 s.d. In all cases leave-one-out refers to analyses performed in the SSC cohort

We examined the extent of linkage disequilibrium between STRs and nearby SNPs using two metrics. The first, termed length r^2, is defined as the squared Pearson correlation between STR allele length and the SNP genotype. The second, termed allelic r^2, treats each STR allele as a separate bi-allelic locus and is computed similar to traditional SNP-SNP LD (Methods). Similar to previous studies[28], SNP-STR LD was dramatically weaker than SNP-SNP LD by both metrics (Supplementary Fig. S2.2a) with length r^2 generally stronger than allelic r^2. We additionally determined the best tag SNP (Methods) for each STR, which was on average 5.5kb away (Supplementary Fig. S2.2b). Nearly all STRs were in significant LD (length r^2 p¡0.05) with the best tag SNP, suggesting that phasing would result in informative haplotypes.

We developed a pipeline to phase STRs onto SNP haplotypes leveraging the quad family structure (Fig. 2.2a). Based on our LD analysis, we used a window size of ±50kb to phase each STR separately using Beagle[40], which was recently demonstrated to perform well in phasing multi-allelic STRs[41] and can incorporate pedigree information. Resulting phased haplotypes from the parent samples were merged into a single genome-wide reference panel for downstream imputation.

We first evaluated the utility of our phased panel for imputation using a leave-one-out analysis in the SSC samples. For each sample, we constructed a modified reference panel with that sample's haplotypes removed and then performed genome-wide imputation. We measured concordance, length r^2, and allelic r^2 between imputed vs. observed genotypes at each STR, where observed refers to genotypes obtained by HipSTR or Tredparse. We additionally evaluated imputation performance under two null models where genotypes were either imputed randomly (random model) or always imputed as the most frequent diploid genotype (naive model) (Methods). Imputed genotypes showed an average of 96.7% concordance with observed genotypes, compared to 61.0% or 71.7% expected under the random and naive models, respectively (Table 2.1). As expected, concordance was strongest at the least polymorphic STRs (Fig. 2.2b, Supplementary Fig. S2.3a) and allelic r^2 was highest for the most common alleles (Supplementary Fig. S2.3b).

Length r^2 was not strongly associated with expected heterozygosity, although the least and most heterozygous STRs tended to have lower length r^2 (Supplementary Fig. S2.3c). Imputation metrics were weakly negatively correlated with distance to the best tag SNP (Pearson r=-0.06; p=0.06, Pearson r=-0.04; p=0.27; and Pearson r=-0.06, $p = 7.5 \times 10^{-5}$ between distance to the best tag SNP and concordance, length r^2, and allelic r^2, respectively). To further evaluate imputation performance at highly polymorphic STRs, we examined the CODIS STRs used in forensic analysis (Supplementary Table S2.2). Per-STR concordances were highly correlated with imputation results recently reported by Edge et al.[41] (Pearson r^2=0.93; $p = 6.3 \times 10^{-6}$; n=10), but were on average 8.8% higher (average concordance 69.1% vs. 60.3% using our panel vs. in Edge et al.[41] restricting to STRs imputed in both studies), likely as a result of our larger and more homogenous cohort. Per-locus imputation statistics for all STRs are reported in Supplementary Data 1 and 2).

We next evaluated our ability to impute STR genotypes into external datasets. For this, we focused on samples from the 1000 Genomes Project[38] with high quality SNP genotypes obtained from low coverage whole-genome sequencing (WGS) (n=2504) or genotyping arrays (n=2486 for Affy 6.0, and n=2318 for Omni 2.5). We validated imputed genotypes for subsets of 1000 Genomes samples using data obtained from three pipelines: (1) Illumina WGS+HipSTR, (2) capillary electrophoresis, and (3) 10X Genomics+HipSTR, in each case using the orthogonal data as the truth set. Each of these datasets evaluates a different aspect of our imputation pipeline. The first tests whether a pipeline identical to that used to create our reference panel can achieve similar performance on datasets collected by different groups using different protocols. Additionally, since it consists of both Europeans and non-Europeans, it allows us to evaluate imputation across a variety of population groups. The second tests whether our results are robust across STR genotyping technologies and allows us to compare imputed STRs based on statistically inferred HipSTR genotypes to those obtained experimentally using capillary electrophoresis. The third returns phased genotypes, allowing us to directly compare inferred haplotypes and phase

Table 2.1: **Imputation performance summary.** Results indicate mean across all STRs analyzed. Allelic r^2 values include all common alleles (frequency at least 5%). Multi-allelic refers to STRs with three or more common alleles. Naive and random denote the two null imputation models as defined in the Methods

Panel (*n*=number of samples)	Observed concordance	Naive concordance	Random concordance	Observed length r^2	Random length r^2	Observed allelic r^2	Random allelic r^2
SSC—LOO (*n*=1916)	96.7%	71.7%	61.0%	0.906	0.605	0.861	0.552
SSC—LOO (multi-allelic)	94.3%	62.2%	48.5%	0.888	0.334	0.800	0.333
1000 Genomes—EUR (*n*=49)	97.0%	75.1%	63.2%	0.921	0.678	0.892	0.543
1000 Genomes—EUR (multi-allelic)	94.8%	66.6%	50.0%	0.900	0.334	0.828	0.314
1000 Genomes—AFR (*n*=46)	90.6%	70.2%	57.9%	0.746	0.619	0.706	0.493
1000 Genomes—AFR (multi-allelic)	85.6%	61.1%	44.4%	0.708	0.336	0.653	0.310
1000 Genomes—EAS (*n*=45)	93.8%	77.2%	66.0%	0.823	0.690	0.781	0.557
1000 Genomes—EAS (multi-allelic)	89.4%	69.7%	53.7%	0.780	0.336	0.663	0.313

information.

First, we used HipSTR to genotype STRs in separate high-coverage (30×) WGS datasets available for 150 of the samples (see URLs) from European (*n*=50), African (*n*=50), and East Asian (*n*=50) backgrounds. Per-locus concordance, length r^2, and allelic r^2 were highly concordant between the SSC panel and 1000 Genomes samples of European origin (Pearson *r*=0.94, 0.63, and 0.85, respectively) (Fig. 2.2c; Supplementary Fig. S2.5; Table 2.1). Overall imputation performance did not vary when using phased genotypes obtained from WGS vs. Omni2.5 for imputation (Supplementary Table S2.3). Concordance was noticeably weaker in African and East Asian samples, likely due to different population background compared to the SSC samples and lower LD in African populations[42].

Next, we compared imputed genotypes to capillary electrophoresis data[43] (see URLs) available for a subset of samples in our panel at highly polymorphic STRs. After filtering non-European samples and STRs that could not be reliably mapped to HipSTR notation (Methods), 41 samples and 206 STRs remained for comparison. We obtained an average overall concordance of 76.9% with capillary genotypes compared with 76.4% expected based on HipSTR analysis. Per-locus concordances based on HipSTR vs. capillary genotypes were strongly correlated (Pearson r=0.83; $p = 1.05 \times 10^{-53}$; n=206) (Fig. 2.2d).

Finally, we compared imputed genotypes from the highly characterized NA12878 genome to phased data available from 10X Genomics (see URLs), a synthetic long read technology. We constructed a phased validation panel by calling HipSTR separately on reads from each phase and combining with phased SNP genotypes (Methods, Supplementary Fig. S2.6). We could obtain phased 10X calls for 116,764 of the STRs in our panel. We used the nearest heterozygous SNP to each STR to match phase order between our panel and the 10X data, which allowed us to directly compare imputed alleles and evaluate phase accuracy. Overall, imputed STR alleles showed 96% concordance with those obtained from 10X and per-locus genotype concordance was consistent with concordance metrics measured in SSC (Fig. 2.2e). Taken together, validation of imputed STR genotypes against three separate truth sets demonstrates the accuracy of our original SNP+STR haplotype panel and shows that our quality metrics are reliable indicators of per-STR imputation performance across datasets.

2.2.3 Imputation increases power to detect STR associations

We sought to determine whether our SNP+STR haplotype panel could increase power to detect underlying STR associations over standard GWAS. First, we simulated phenotypes based on a single causal STR and examined the power of the imputed STR genotypes vs. nearby SNPs to detect associations. We focused primarily on a linear additive model relating STR dosage, defined as the average allele length, to quantitative phenotypes (Fig. 2.3a), since the majority of known functional STRs follow similar models (e.g., refs. [17, 21, 44, 45]). Association testing simulations were performed 100 times for each STR on chromosome 21 in our dataset (Methods). As expected, the strength of association for each variant as measured by the negative log10 p-value was linearly related with its length r^2 with the causal variant (Fig. 2.3b). On average, imputed STR genotypes explained 17.7% more variation in STR allele length compared to the best tag SNP (mean length r^2=0.92 and 0.74 for imputed STRs vs. SNPs, respectively). The advantage from STR imputation grew as a function of the number of common STR alleles (Supplementary Fig. S2.7). Imputed genotypes showed a corresponding increase in power to detect associations at a given p-value threshold (Fig. 2.3c). Similar trends were observed for case–control traits (Supplementary Fig. S2.8). We additionally tested the ability of imputed STR genotypes to identify associations due to non-linear models relating STR genotype to phenotype (Supplementary Fig. S2.9). While both STR and SNP-based tests had limited power to detect non-linear associations, per-allele STR association tests had higher power than the best tag SNP in 60% of simulations. Importantly, testing for complex models relating repeat length to phenotype will only be possible when allele lengths are available, thus demonstrating an additional need for STR imputation over SNP-based tests to detect these associations.

We next determined whether STR imputation could identify STR associations using real phenotypes. We focused on gene expression, given the large number of reported associations between STR length and expression of nearby genes in cis[17, 18] (termed eSTRs). To this end, we analyzed eSTRs from samples in the Genotype-Tissue Expression [46] (GTEx) dataset

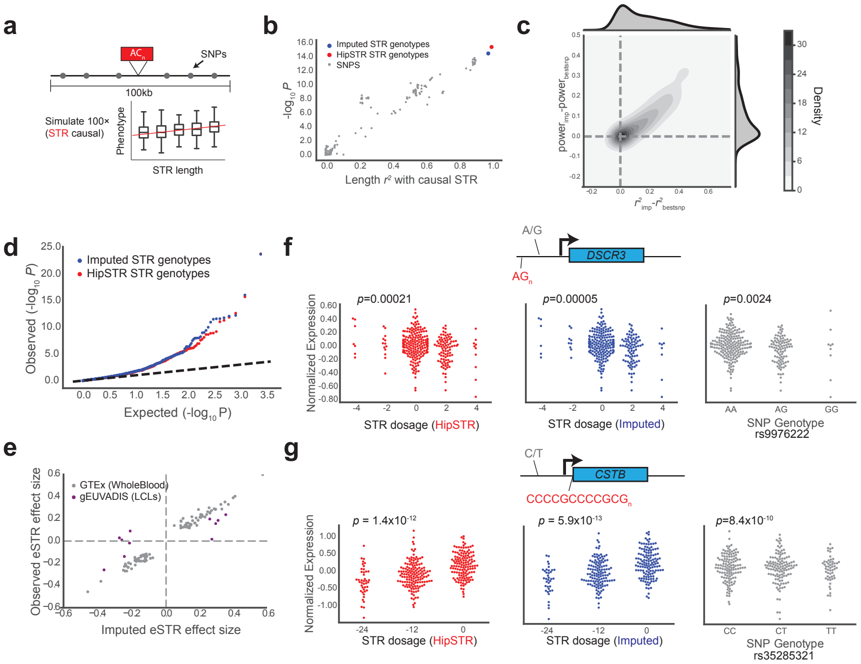


Figure 2.3: **STR imputation improves power to detect STR associations. a.** Example simulated quantitative phenotype based on SSC genotypes. A quantitative phenotype was simulated assuming a causal STR (red). Power to detect the association was compared between the causal STR, imputed STR genotypes, and all common SNPs (MAF¿0.05) within a 50kb window of the STR (gray). **b.** Strength of association (-$\log_{10}$ p) is linearly related with LD with the causal variant. For SNPs, the x-axis gives the length r^2 calculated using observed genotypes. For the imputed STR (blue), the x-axis gives the length r^2 from leave-one-out analysis. **c.** The gain in power using imputed genotypes is linearly related to the gain in length r^2 compared to the best tag SNP. Gray contours give the bivariate kernel density estimate. Top and right gray area gives the distribution of points along the x- and y-axes, respectively. Power was calculated based on the number of simulations out of 100 with nominal p¡0.05. **d.** Quantile-quantile plot for eSTR association tests. Each dot represents a single STR×gene test. The x-axis gives the expected $\log_{10}$ p-value distribution under a null model of no eSTR associations. Red and blue dots give $\log_{10}$ p-values for association tests using HipSTR genotypes and imputed STR genotypes, respectively. Black dashed line gives the diagonal. **e.** 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. **f, g** 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

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 HipSTR (Methods). A total of 2452 STR × 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 eSTR associations (Fig. 2.3d). For all nominally significant associations (*p*¡0.05), effect sizes were strongly correlated when using imputed vs. HipSTR genotypes (Pearson *r*=0.99; $p = 1.01 \times 10^{-79}$, *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[17]) (Pearson *r*=0.79; *p*=0.0042, *n*=11) (Fig. 2.3e).

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 (Methods). We additionally applied CAVIAR[47] to fine-map associations using the most strongly associated STR and the top 100 associated SNPs for each gene (Methods). We identified three genes with ANOVA *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[48] and expansions of this repeat are implicated in myoclonus epilepsy[49]. In each case, imputed STR genotypes were more strongly associated with gene expression compared to the best tag SNP (Fig. 2.3f–g, Supplementary Table S2.4).

2.2.4 Imputing normal alleles at known pathogenic STRs

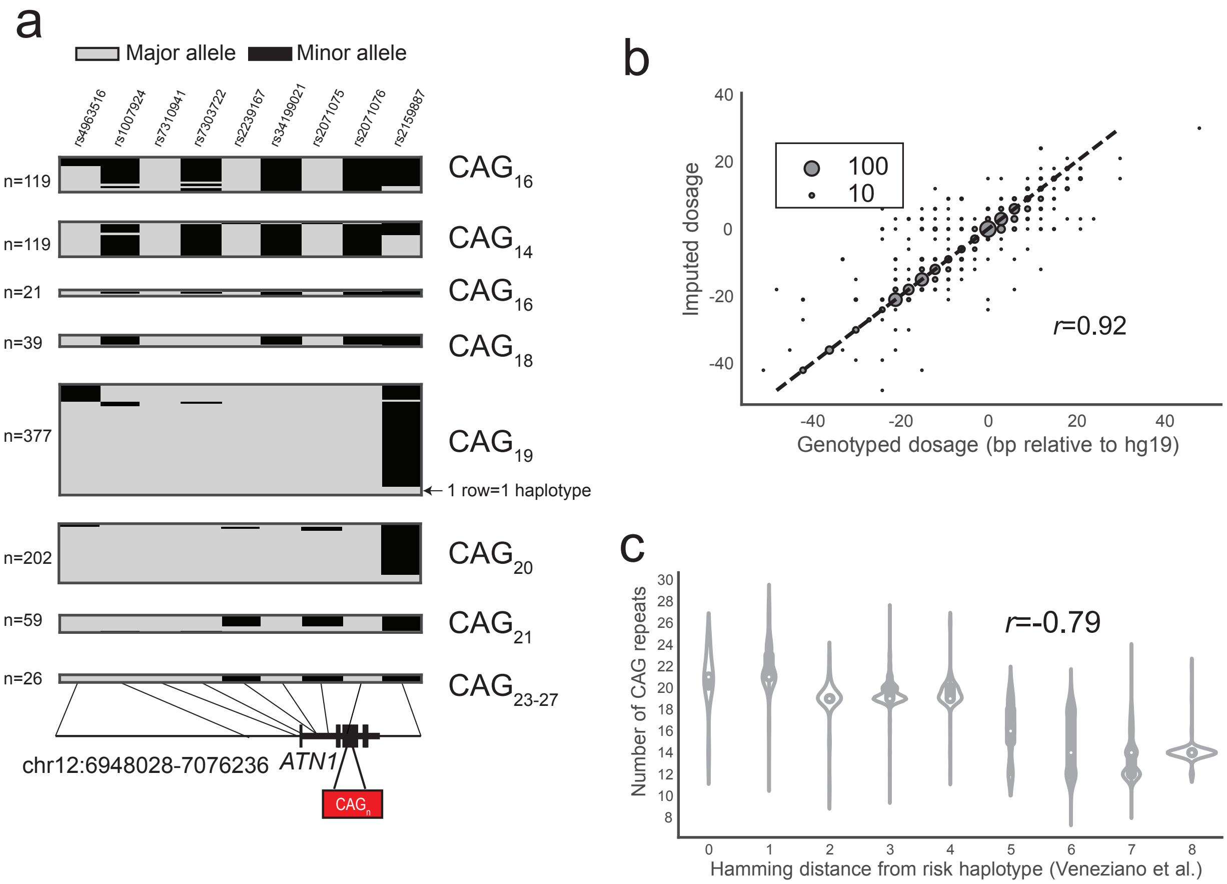


Figure 2.4: **SNP haplotypes distinguish allele lengths at known pathogenic STRs.** **a.** Example SNP-STR haplotypes inferred in European samples at a polyglutamine repeat in *ATN1* implicated in DRPLA. Each column represents a SNP from the founder haplotype reported by Veneziano et al. Each row represents a single haplotype inferred in 1000 Genomes Project phase 3 European samples, with gray and black boxes denoting major and minor alleles, respectively. Haplotypes are grouped by the corresponding STR allele. The number of SNP haplotypes for each group of STR alleles is annotated to the left of each box. Alleles seen fewer than 10 times in 1000 Genomes samples were excluded from the visualization. **b.** Comparison of imputed vs. observed STR genotypes in SSC samples at the DRPLA locus. The *x*-axis gives the maximum likelihood genotype dosage returned by HipSTR and the *y*-axis gives the imputed dosage. Dosage is defined as the sum of the two allele lengths of each genotype relative to the hg19 reference genome. The bubble size represents the number of samples summarized by each data point. **c.** Distribution of DRPLA repeat length vs. similarity to the pathogenic founder haplotype. The founder haplotype refers to the SNP haplotype reported by Veneziano, et al. on which a pathogenic expansion in *ATN1* implicated in DRPLA likely originated. The *x*-axis gives the Hamming distance between observed haplotypes and the founder haplotype, computed as the number of positions with discordant alleles. White dots represent the median length

Finally, to determine whether alleles at known pathogenic STRs could be accurately imputed, we examined results of our imputation pipeline at 12 STRs previously implicated in expansion disorders that were included in our panel (Table 2.2). Our analysis focused on alleles in the normal repeat range for each STR, since pathogenic repeat expansions at these STRs are unlikely to be present in the SSC cohort. Notably, accurate imputation of non-pathogenic allele ranges is still informative as (1) long normal or intermediate size alleles may result in mild symptoms in some expansion disorders[50, 51, 52] (2) longer alleles are more at risk for expansion[53] and (3) allele lengths below the pathogenic range could potentially be associated with more complex phenotypes[51].

Similar to the CODIS markers, these STRs are highly polymorphic with 10 or more alleles per locus. In all cases, imputed genotypes were more strongly correlated with observed genotypes compared to the best tag SNP. Where both HipSTR and Tredparse genotypes were available, concordance results were nearly identical across all STRs (Supplementary Table S2.5). Visualization of SNP-STR haplotypes at the CAG repeat implicated in dentatorubral-pallidoluysian atrophy (DRPLA)[54] reveals a typical complex relationship between STR allele length and local SNP haplotype (Fig. 2.4a), with the same STR allele often present on multiple SNP haplotype backgrounds. Still, for most STRs there is a clear association of specific haplotypes with different allele length ranges allowing accurate imputation across a large range of allele sizes (Fig. 2.4b, Supplementary Fig. S2.10).

Resolution of SNP-STR haplotypes can be used to infer the mutation history of a specific STR locus[26, 27]. Notably, for many STR expansion orders it has been shown that pathogenic expansion alleles originated from a founder haplotype[55, 56, 57, 58] associated with a long allele. We compared SNP haplotypes at the DRPLA locus in our dataset to a previously reported founder haplotype[55]. In concordance with the hypothesis of a single founder haplotype, we found that SNP haplotypes with smaller Hamming distance to the known founder haplotype had longer CAG tracts (Pearson r=-0.79; $p < 10^{-200}$). This finding demonstrates that while we were

Table 2.2: **Imputation performance at known pathogenic repeats.** *HD* Huntington's disease; SCA spinocerebellar ataxia, DRPLA Dentatorubral-pallidoluysian atrophy, DM1 myotonic dystrophy type 1, HDL huntington's disease-like 2. The best tag SNP for an STR is defined as the SNP within 50kb with the highest length r^2. LOO refers to leave-one-out analysis in the SSC cohort $r^2_{bestSNP}$ gives the length r^2 between STR genotype length and the genotype of the best tagging SNP within 50kb of the STR

Locus	Motif	Disorder	Length r^2 LOO	Observed concordance	Naive concordance	Random concordance	Best tag SNP	r^2 bestSNP
3:63898362	CAG	SCA7	0.75	92.0%	75.6%	63.9%	rs58676857	0.57
4:3076604	CAG	HD	0.47	64.3%	39.4%	27.5%	rs762855	0.11
5:146258292	CAG	SCA12	0.88	93.8%	59.9%	46.3%	rs2082405	0.64
6:16327867	CAG	SCA1	0.72	85.3%	55.0%	33.8%	rs17860797	0.04
6:170870996	CAG	SCA17	0.51	80.0%	39.8%	31.5%	rs9472489	0.15
12:112036755	CAG	SCA2	0.49	96.2%	88.2%	80.2%	rs148019457	0.28
12:7045892	CAG	DRPLA	0.86	81.2%	38.8%	24.9%	rs34199021	0.69
13:70713516	CTG / CAG	SCA8	0.87	84.7%	27.0%	24.0%	rs9564660	0.39
14:92537355	CAG	SCA3	0.88	86.4%	33.8%	27.5%	rs7144492	0.27
16:87637894	CAG	HDL	0.55	88.2%	55.2%	46.5%	rs2434850	0.34
19:46273463	CTG	DM1	0.87	86.9%	39.4%	30.8%	rs7254351	0.44
19:13318673	CAG	SCA6	0.81	92.0%	44.1%	39.2%	rs2070737	0.63

unable to directly impute pathogenic expansion alleles, STR imputation can accurately identify which individuals are at risk for carrying expansions or pre-pathogenic mutations and the inferred haplotypes can reveal the history by which such mutations arise.

2.3 Discussion

Our study combines available whole-genome sequencing datasets with existing bioinformatics tools to generate the first phased SNP+STR haplotype panel allowing genome-wide imputation of STRs into SNP data. Despite their exceptionally high rates of polymorphism, 92% of STRs in our panel could be imputed with at least 90% concordance, and 38% achieved greater than 99% concordance. Imputation performance varied widely across STRs, primarily due to differences in polymorphism levels across loci. Bi-allelic STRs could be imputed nearly perfectly (average concordance ¿99%, compared to 80% expected under a naive model), whereas STRs with the highest heterozygosity, including forensic markers and known pathogenic repeats, could be imputed to around 70% concordance (compared to approximately 50% expected under a naive model). We additionally show that imputation improves power to detect STR associations over standard SNP-based GWAS and could detect both known and previously unknown associations between STR lengths and expression of nearby genes.

A widely recognized limitation of GWAS is the fact that common SNP associations still explain only a small fraction of heritability of most traits. Multiple explanations for this have been proposed, including minute effect sizes of individual variants and a potential role for high-impact rare variation[59]. However, studies in large cohorts reaching hundreds of thousands of samples[3, 2, 1], as well as deep sequencing studies to detect rare variants[60], have so far not confirmed these hypotheses. An increasingly supported idea is that complex variants not well tagged by SNPs may comprise an important component of the missing heritability[10, 11, 12]. GWAS is essentially blind to contributions from highly polymorphic STRs and other repeats, despite their known importance to human disease and molecular phenotypes. Thus, STR association studies will undoubtedly uncover additional heritability that is so far unaccounted for. Notably, while autism phenotypes are available for the SSC families, this cohort is too small to perform a GWAS and was specifically ascertained for families enriched for de novo, rather than inherited, pathogenic mutations. In future work our panel can be applied to impute STRs into larger cohorts for autism and other complex traits for which tens of thousands of SNP array datasets are available.

Our initial haplotype panel faces several important limitations. First, the majority of samples are of European origin, limiting imputation accuracy in other population groups. Second, imputation accuracy is mediocre for the most highly polymorphic STRs, some of which will ultimately have to be directly genotyped to adequately test for associations. Notably, our work relied on existing tools originally designed for SNP imputation. Further work on computational methods specifically for imputing repeats may be able to improve performance. Finally, thousands of long STRs are filtered from our panel due to the limitation imposed by short read lengths. While we have included target STRs implicated in STR expansion disorders, many long STRs are still inaccessible using current tools. New methods are now being developed for genome-wide genotyping of more complex STRs[37, 61] and longer variable number tandem repeats (VNTRs)[62] from short reads and can be used to expand our panel in the future. Overall, our STR imputation framework will enable an entire new class of variation to be interrogated by reanalyzing hundreds of thousands of existing datasets, with the potential to lead to novel genetic discoveries across a broad range of phenotypes.

2.4 Methods

2.4.1 SSC Dataset

The SSC Phase 1 dataset consists of 1916 individuals from 479 quad families. Access to SSC data was approved for this project under SFARI Base project ID 2405.1. This study was certified as exempt from institutional review board (IRB) review by the University of California San Diego IRB (Project #161286XX) since only de-identified data was accessed. Informed consents were obtained for each participating family by SSC recruitment sites in accordance with their local IRBs.

Aligned BAM and gVCF files for whole-genome sequencing data of individuals were obtained through SFARI base (see URLs) and processed on Amazon Web Services (AWS). SNP genotypes were called from gVCF files using the GATK version 3 joint calling pipeline[63]. A total of 27,185,239 variants that passed the default GATK filters and overlapped with sites reported in the 1000 Genomes Project phase 3[38] data were retained for downstream analysis.

We performed principal components analysis (PCA) using SNPs from 2504 samples from Phase 3 of the 1000 Genomes Project[38] and projected SSC samples onto the resulting PCs to infer sample ancestry (Supplementary Fig. S2.1). We estimated that the SSC cohort consists of 1585 Europeans, 39 East Asian, 172 South Asian, 69 African samples, and 51 individuals that did not clearly belong to any single population group.

2.4.2 Genome-wide multi-sample STR genotyping

STRs were jointly genotyped on the AWS EC2 platform in batches of 500 STRs. We streamed the corresponding region of each BAM file and of the phased SNP VCF files to a local EBS volume attached to each EC2 instance using samtools[64] version 1.4 and tabix[65] version 1.2, respectively. HipSTR[32] version v0.5 was called individually per STR with default parameters. Phased SNPs were provided as input to allow HipSTR to perform physical phasing when possible. Resulting VCF files from each batch were merged to create a genome-wide callset in VCF format.

HipSTR calls were filtered using the filter_vcf.py script in the HipSTR package with suggested parameters (–min-call-qual 0.9 –max-call-flank-indel 0.15 –max-call-stutter 0.15). We used the following criteria to remove problematic STRs from the callset: (i) STRs overlapping segmental duplications (UCSC Table Browser[66] hg19.genomicSuperDups table) were removed from the callset using intersectBed[67] v2.25.0; (ii) Pentanucleotides and hexanucleotides containing homopolymer runs of at least 5 or 6 nucleotides, respectively, in the hg19 reference genome were removed as they were found to contain an excess of indels in the homopolymer regions; (iii) STRs with call rate ¡80%; (iv) STRs with expected heterozygosity ¡0.095, corresponding to a minor allele frequency of 5% for bi-allelic markers, were removed to restrict to polymorphic STRs; (v) STRs with significantly more or fewer heterozygous genotypes compared to expectation under Hardy-Weinberg equilibrium (p¡0.01) as suggested previously[68]. After filtering, 453,671 STRs remained in our panel.

2.4.3 Genotyping clinically relevant STRs

A total of 25 clinically relevant STRs were called using Tredparse[37] v0.75 from the aligned BAM files obtained through SFARI base on Amazon EC2. Default profiles containing information about the genomic position, reference repeat length, and repeat motif supplied with the software were used. We filtered STRs with call rate less than 80% or for which only a single allele was identified (Supplementary Table S2.1). Nine STRs remained after filtering.

2.4.4 Computing expected STR heterozygosity

For an STR with alleles $\{1...n\}$, let p_i be the frequency of the ith allele computed from observed genotypes. Expected STR heterozygosity is defined as: $H = 1 - \sum_{i=1}^{n} p_i^2$. For this study all alleles with identical length are treated as the same allele. On average each length-based allele corresponded to 1.8 sequence-based alleles.

2.4.5 Comparison to 1000G catalog

STR genotypes for 1000 Genomes samples generated by Willems et al.[14] were downloaded from the strcat site (see URLs). Expected heterozygosity was computed using the PyVCF package (see URLs) for the 1000 Genomes calls and using a custom script for the SSC data to collapse alleles of identical length into a single allele. STRs passing all filters described above included in the comparison. Analysis was restricted to STRs with at least 500 calls in the 1000 Genomes dataset.

2.4.6 Normal allele frequency distributions at pathogenic STRs

Control distributions for Fig. 2.1e were obtained from previous studies of normal alleles at known pathogenic STRs. Allele frequencies for SCA1, SCA2, SCA3, SCA6, SCA12, SCA8, SCA17, and DRPLA were obtained from Fig. 1 of Majounie et al.[36] and are based on 307 controls of Welsh origin. Frequencies for DM1 were obtained from Fig. 1 of Ambrose et al.[35] and are based on 254 controls of Chinese origin. Frequencies for HDL were obtained from Fig. 1 of Figley et al.[34] and are based on 352 controls of North American Caucasian origin. Frequencies for SCA7 were obtained from Fig. 1 of Gouw et al.[33] and are based on 180 controls of European origin. Frequencies for HTT are based on data in the phv00173896.v1.p1 variable of dbGaP study phs000371.v1.p1 (Genetic modifiers of Huntington's Disease) based on the shorter allele of 2802 patients with Huntington's Disease.

2.4.7 Phasing SNPs in the SSC

SNP genotypes were phased using SHAPEIT[69] version 2.r837 with 1000 Genomes Phase 3 genotypes as a reference panel and ignoring pedigree information. SHAPEIT's duoHMM [70] version 0.1.7 method was used to refine phased haplotypes using pedigree structure and correcting for Mendelian errors.

2.4.8 Phasing STRs

Beagle[40] version 4.0 was used to phase each STR separately using phased SNP genotypes, pedigree information, and unphased STR genotypes as input. In order to leverage the HipSTR genotype likelihoods (GL field), Beagle requires all samples to have GL information. To accommodate this, phasing was performed in two steps. First, samples with missing data were removed and the remaining samples were phased using the -gl Beagle flag. Next, missing samples were added back to the VCF and all samples were jointly phased in a second Beagle round using default parameters. In this step Beagle additionally imputed any calls with missing genotypes. Genotype values (GT field) were used for the STRs genotyped using Tredparse as it does not report genotype likelihoods, and phasing and imputation of STRs was done in a single step. Phased STRs and SNPs for only the unrelated parent samples from each locus were then merged into a single genome-wide reference panel in VCF format.

2.4.9 Imputation performance metrics

Let $X = \{x1, x2, ...xn\}$ be the true STR genotypes for samples $1..n$ and $Y = \{y_1, y_2, ...y_n\}$ be the imputed STR genotypes. Each genotype x_i is defined as $x_i = (x_{i1}, x_{i2})$ where x_{i1} and x_{i2} give the (unordered) lengths of the two STR alleles for a diploid sample and similarly for Y. We then define the following metrics:

Genotype concordance c_i was defined as: 1 if both genotypes match $(x_{i1} = y_{i1} and x_{i2} =$

$y_{i2} or x_{i2} = y_{i1} and x_{i1} = y_{i2})$; 0 if neither imputed allele matched a true allele; else 0.5 if one but not both imputed alleles matched the true alleles. Genotype concordance for an STR is the average over all the samples $C = \frac{1}{n}\sum_{i=1}^{n} c_i$.

Define the STR genotype dosage as the sum of the lengths of the two alleles at a given site: $d_i = x_{i1} + x_{i2}$ and $X_d = \{d_1, d_2, ..., d_n\}$. Length r^2 is computed as:

$$cov^2(X_d, Y_d)/(Var(X_d)Var(Y_d))$$

For a given allele length a, define $X_a = \{a_1, a_2, ..., a_n\}$ where $a_i = \sum_{j=1}^{2} 1_{(xij=a)}$. Allelic r^2 is computed as $cov^2(X_a, Y_a)/(Var(X_a)Var(Y_a))$.

The best tag SNP for an STR is defined as the SNP within 50kb with the highest length r^2.

For all concordance metrics, outlier genotypes containing alleles seen less than three times in the entire cohort were removed from the analysis.

For each STR, we additionally computed the expected value of each metric under a random model where genotypes are imputed randomly based on the frequency of underlying alleles and a naive model where genotypes are imputed to be the most common diploid genotype. Expected genotype concordance under the random model was calculated as

$$\sum_{i,j} f_i f_j (\sum_{k,l} C(i,j,k,l))$$

, where $(i,j) \in \{1,...,n\}^2$ and $(k,l) \in \{1,...,n\}^2$, n is the number of alleles, f_x gives the frequency of allele x, and $C(i,j,k,l)$ gives the concordance between genotypes (i,j) and (k,l) as defined above. For example, for a bi-allelic marker with allele frequencies f_1 and f_2 expected genotype concordance under the random model is given by $f_1^2(f_1^2 + (0.5)2(f_1)(f_2)) + 2f_1f_2((0.5)f_1^2 + 2f_1f_2 + (0.5)f_2^2) + f_2^2(f_2^2 + (0.5)2f_1f_2)$. Random model values for length r_2 and allelic r_2 were computed by comparing genotypes imputed randomly based on population allele frequencies to true genotypes at each STR. Concordance under the naive model was computed by comparing each sample's genotype to the most frequent diploid genotype. Length r_2 and allelic r_2 are not defined under the naive model since all imputed genotypes are identical.

2.4.10 Evaluating imputation performance in the 1000 Genomes data

STRs were imputed into SNP data downloaded from the 1000 Genomes Project site from three sources (WGS, phased SNPs from Affy6.0 array; and phased SNPs from Omni2.5 array; see URLs and Supplementary Table S2.3) with Beagle version 4.1 using the SSC SNP-STR haplotype panel. For comparison to WGS, STRs were jointly genotyped in high-coverage WGS datasets for 150 of the 1000 Genomes Project samples (see URLs) using HipSTR version 0.6 followed by the filtering steps described above for the SSC cohort.

Capillary electrophoresis genotypes for 209 samples at 721 Marshfield STRs were downloaded from the Payseur Lab website (see URLs). PCR product sizes were converted to length differences in bp from the reference genome using product size annotations[71] available from the Rosenberg Lab website (see URLs). Prior to comparing genotypes, offsets were calculated to match HipSTR lengths to the length of Marshfield STRs as previously described[14]. STRs with imperfect repeat structures were removed. Capillary genotypes were rounded down to the nearest number of repeat units.

10X Genomics data for NA12878 was obtained from the NA12878 Gemline Genome v2 available on the 10X Genomics website (see URLs). We extracted reads belonging to phase 1 or 2 from the phased, barcoded BAM based on the HP tag into separate BAM files. HipSTR v0.6.1 was called separately on each BAM with non-default parameters –def-stutter-model –min-reads 5 –use-unpaired and with –haploid-chrs containing a list of all autosomal chromosomes to force a haploid genotyping model. Haploid STR calls were obtained for both phases at 118,353 STRs. We identified the nearest heterozygous SNP to each STR that was genotyped in both the 10X data and in our phased panel. STRs for which the nearest SNP had discordant genotypes in the two datasets were discarded leaving 116,764 STRs for analysis.

2.4.11 Simulations for power analysis

We analyzed parental genotypes for 5838 STRs across chromosome 21 that passed filtering and quality control as described above. For each STR, we simulated quantitative phenotype datasets under the model: $P = \beta G + E$, where P is a vector of standard normalized phenotypes, β gives the effect size, E gives the error term drawn from a normal distribution $N(0, 1-\beta)$, and G is a vector of the sum of genotype lengths for each individual scaled to have mean 0 and variance 1. For each simulated phenotype dataset, we tested the causal STR, the imputed STR genotypes, and the best tag SNP (strongest length r^2) within 50kb of the STR for association. Association tests were performed using the Python statsmodels library OLS method (see URLs).

We performed additional simulations under a case–control model shown in Supplementary Fig. S2.8. Phenotypes (0=control, 1=case) were drawn for each sample according to the model logit (p_i) = βX_i where p_i is the probability that sample i is a case and X_i is the scaled genotype for individual i as described above. Association tests were performed using the Python statsmodels Logit method.

For the non-additive phenotype example (Supplementary Fig. S2.9), we performed simulations under a quadratic model: $P = \beta G^2 + E$ where G is a vector of the squared sum of allele lengths scaled by the mean allele length, and P, β, E are as described above. Two sets of association tests were performed: the first tested for association between STR length and phenotype (Supplementary Fig. S2.9b) and the second set performed a separate association test for each STR allele treating the allele as a bi-allelic locus (Supplementary Fig. S2.9c).

In all cases 100 separate simulations were performed and power was defined as the percent of simulations for which the nominal association p-value was ¡ 0.05. Figures show results for all simulations with β set to 0.1.

2.4.12 eSTR analysis

Data for eSTR analysis was obtained from the Genotype-Tissue Expression (GTEx) through dbGaP under phs000424.v7.p2. This included high-coverage (30×) Illumina whole-genome sequencing (WGS) data from 650 unrelated samples, Omni 2.5 SNP genotypes for 450 samples, and gene-level RPKM values for whole blood in 336 samples. STRs were genotyped from WGS data using HipSTR v0.5 and subject to the same quality filtering as SSC samples. STRs were additionally imputed to Omni2.5 data with Beagle as described above. Downstream analyses were restricted to the 336 samples with available whole blood expression data. These samples consisted of 284 European, 45 African American, 3 Asian, and 3 Amerindian samples and 2 samples with no population label available.

We performed separate eSTR analyses using HipSTR and imputed genotypes. In each case, we performed a separate association test between gene expression and each STR within 100kb of the gene using a model $Y = \beta X + C + \varepsilon$, 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[17], 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 are reported as length differences from the hg19 reference, with 0 representing the reference allele. STR dosages were scaled to have mean 0 and variance 1. Genes with median expression of 0 were excluded and expression values for remaining genes were quantile normalized to a standard normal distribution. 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 PEER factor correction[72, 73] using 83 PEER factors.

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. 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. 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).

2.4.13 Comparison to DRPLA founder haplotypes

The founder haplotype for the expansion allele in ATN1 implicated in DRPLA was taken from Table 1 of Veneziano et al.[55] and consists of rs4963516, rs1007924, rs7310941, rs7303722, rs2239167, rs34199021, rs2071075, rs2071076, and rs2159887 with hg19 alleles G, A, G, T, A, A, T, C, and C, respectively. Distance from the founder haplotype was calculated as the number of mismatches.

URLs

For Simons Simplex Collection, see https://base.sfari.org/.

For HipSTR, see https://github.com/tfwillems/HipSTR.

For Beagle, see https://faculty.washington.edu/browning/beagle/b4_0.html.

For 1000 Genomes phased Affy6.0 and Omni2.5 SNP data, see ftps.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/shapeit2_scaffolds/hd_chip_scaffolds/.

For 1000 Genomes Phase 3, see http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

For 1000 Genomes STR data, see http://strcat.teamerlich.org/download.

For Marshfield Capillary electrophoresis data, see https://payseur.genetics.wisc.edu/strpData.htm.

For Marshfield marker annotations, see https://web.stanford.edu/group/rosenberglab/data/pembertonEtAl2009/Pemberton_AdditionalFile1_11242009.txt.

For NA12878 10X Genomics data, see https://support.10xgenomics.com/genome-exome/datasets/2.2.1/NA12878_WGS_v2.

For High-coverage Illumina sequencing for 1000 Genomes samples, see https://www.ebi.ac.uk/ena/data/view/PRJEB20654.

For PyVCF, see https://github.com/jamescasbon/PyVCF.

For Python statsmodels, see http://www.statsmodels.org/stable/index.html.

Code availability

Analysis scripts and Jupyter notebooks for reproducing the figures in this study are provided in the Github repository https://github.com/gymreklab/snpstr-imputation.

Data availability

Phased SNP-STR haplotypes for 1000 Genomes Project phase 3 samples and example commands for imputation are available from Gymrek Laboratory webpage [https://gymreklab.github.io/2018/03/05/snpstr_imputation.html].

Phased SNP-STR haplotypes for the SSC samples are available through SFARI base Accession Code: SFARI_SSC_WGS_1c. 1000 Genomes phased Affy6.0 and Omni2.5 SNP data are available through the 1000 Genomes FTP server [ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/shapeit2_scaffolds/hd_chip_scaffolds/].

1000 Genomes phase 3 Whole-Genome Sequencing data is available through the 1000 Genomes FTP server [http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/].

1000 Genomes STR data is available from strcat [http://strcat.teamerlich.org/download].

Marshfield Capillary electrophoresis data is available from the Payseur Laboratory webpage [https://payseur.genetics.wisc.edu/strpData.htm].

Marshfield marker annotations are available from the Rosenberg Laboratory webpage [https://web.stanford.edu/group/rosenberglab/data/pembertonEtAl2009/Pemberton_AdditionalFile1_11242009.txt].

NA12878 10X Genomics data is available at the 10X Genomics Datasets Repository [https://support.10xgenomics.com/genome-exome/datasets/2.2.1/NA12878_WGS_v2].

High-coverage Illumina sequencing for 1000 Genomes samples is available from the European Nucleotide Archive Accession Code PRJEB20654

Bibliography

[1] R. A. Scott, L. J. Scott, R. Magi, L. Marullo, K. J. Gaulton, M. Kaakinen, N. Pervjakova, T. H. Pers, A. D. Johnson, J. D. Eicher, A. U. Jackson, T. Ferreira, Y. Lee, C. Ma, V. Steinthorsdottir, G. Thorleifsson, L. Qi, N. R. Van Zuydam, A. Mahajan, H. Chen, P. Almgren, B. F. Voight, H. Grallert, M. Muller-Nurasyid, J. S. Ried, N. W. Rayner, N. Robertson, L. C. Karssen, E. M. van Leeuwen, S. M. Willems, C. Fuchsberger, P. Kwan, T. M. Teslovich, P. Chanda, M. Li, Y. Lu, C. Dina, D. Thuillier, L. Yengo, L. Jiang, T. Sparso, H. A. Kestler, H. Chheda, L. Eisele, S. Gustafsson, M. Franberg, R. J. Strawbridge, R. Benediktsson, A. B. Hreidarsson, A. Kong, G. Sigurethsson, N. D. Kerrison, J. Luan, L. Liang, T. Meitinger, M. Roden, B. Thorand, T. Esko, E. Mihailov, C. Fox, C. T. Liu, D. Rybin, B. Isomaa, V. Lyssenko, T. Tuomi, D. J. Couper, J. S. Pankow, N. Grarup, C. T. Have, M. E. Jorgensen, T. Jorgensen, A. Linneberg, M. C. Cornelis, R. M. van Dam, D. J. Hunter, P. Kraft, Q. Sun, S. Edkins, K. R. Owen, J. R. B. Perry, A. R. Wood, E. Zeggini, J. Tajes-Fernandes, G. R. Abecasis, L. L. Bonnycastle, P. S. Chines, H. M. Stringham, H. A. Koistinen, L. Kinnunen, B. Sennblad, T. W. Muhleisen, M. M. Nothen, S. Pechlivanis, D. Baldassarre, K. Gertow, S. E. Humphries, E. Tremoli, N. Klopp, J. Meyer, G. Steinbach, et al. An expanded genome-wide association study of type 2 diabetes in europeans. *Diabetes*, 66(11):2888–2902, 2017.

[2] Consortium Schizophrenia Working Group of the Psychiatric Genomics. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–7, 2014.

[3] V. Turcot, Y. Lu, H. M. Highland, C. Schurmann, A. E. Justice, R. S. Fine, J. P. Bradfield, T. Esko, A. Giri, M. Graff, X. Guo, A. E. Hendricks, T. Karaderi, A. Lempradl, A. E. Locke, A. Mahajan, E. Marouli, S. Sivapalaratnam, K. L. Young, T. Alfred, M. F. Feitosa, N. G. D. Masca, A. K. Manning, C. Medina-Gomez, P. Mudgal, M. C. Y. Ng, A. P. Reiner, S. Vedantam, S. M. Willems, T. W. Winkler, G. Abecasis, K. K. Aben, D. S. Alam, S. E. Alharthi, M. Allison, P. Amouyel, F. W. Asselbergs, P. L. Auer, B. Balkau, L. E. Bang, I. Barroso, L. Bastarache, M. Benn, S. Bergmann, L. F. Bielak, M. Bluher, M. Boehnke, H. Boeing, E. Boerwinkle, C. A. Boger, J. Bork-Jensen, M. L. Bots, E. P. Bottinger, D. W. Bowden, I. Brandslund, G. Breen, M. H. Brilliant, L. Broer, M. Brumat, A. A. Burt, A. S. Butterworth, P. T. Campbell, S. Cappellani, D. J. Carey, E. Catamo, M. J. Caulfield, J. C. Chambers, D. I. Chasman, Y. I. Chen, R. Chowdhury, C. Christensen, A. Y. Chu, M. Cocca, F. S. Collins, J. P. Cook, J. Corley, J. Corominas Galbany, A. J. Cox, D. S. Crosslin, G. Cuellar-Partida, A. D'Eustacchio, J. Danesh, G. Davies, P. I. W. Bakker, M. C. H. Groot, R. Mutsert, I. J. Deary, G. Dedoussis, E. W. Demerath, M. Heijer, A. I. Hollander, H. M. Ruijter, J. G. Dennis, J. C. Denny, E. Di Angelantonio, F. Drenos, M. Du, M. P. Dube, A. M. Dunning, D. F. Easton, et al. Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. *Nat Genet*, 50(1):26–41, 2018.

[4] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7):499–511, 2010.

[5] L. Yengo, J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood, M. N. Weedon, T. M. Frayling, J. Hirschhorn, J. Yang, P. M. Visscher, and GIANT Consortium. Meta-analysis of genome-wide association studies for height and body mass index in 700,000 individuals of european ancestry. *bioRxiv*, 2018.

[6] A. Sekar, A. R. Bialas, H. de Rivera, A. Davis, T. R. Hammond, N. Kamitaki, K. Tooley, J. Presumey, M. Baum, V. Van Doren, G. Genovese, S. A. Rose, R. E. Handsaker, Consortium Schizophrenia Working Group of the Psychiatric Genomics, M. J. Daly, M. C. Carroll, B. Stevens, and S. A. McCarroll. Schizophrenia risk from complex variation of complement component 4. *Nature*, 530(7589):177–83, 2016.

[7] L. M. Boettger, R. M. Salem, R. E. Handsaker, G. M. Peloso, S. Kathiresan, J. N. Hirschhorn, and S. A. McCarroll. Recurring exon deletions in the hp (haptoglobin) gene contribute to lower blood cholesterol levels. *Nat Genet*, 48(4):359–66, 2016.

[8] E. M. Leffler, G. Band, G. B. J. Busby, K. Kivinen, Q. S. Le, G. M. Clarke, K. A. Bojang, D. J. Conway, M. Jallow, F. Sisay-Joof, E. C. Bougouma, V. D. Mangano, D. Modiano, S. B. Sirima, E. Achidi, T. O. Apinjoh, K. Marsh, C. M. Ndila, N. Peshu, T. N. Williams, C. Drakeley, A. Manjurano, H. Reyburn, E. Riley, D. Kachala, M. Molyneux, V. Nyirongo, T. Taylor, N. Thornton, L. Tilley, S. Grimsley, E. Drury, J. Stalker, V. Cornelius, C. Hubbart, A. E. Jeffreys, K. Rowlands, K. A. Rockett, C. C. A. Spencer, D. P. Kwiatkowski, and Network Malaria Genomic Epidemiology. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science*, 356(6343), 2017.

[9] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[10] A. J. Hannan. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. *Trends Genet*, 26(2):59–65, 2010.

[11] A. J. Hannan. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet*, 19(5):286–298, 2018.

[12] M. O. Press, K. D. Carlson, and C. Queitsch. The overdue promise of short tandem repeat variation for heritability. *Trends Genet*, 30(11):504–12, 2014.

[13] J. X. Sun, A. Helgason, G. Masson, S. S. Ebenesersdottir, H. Li, S. Mallick, S. Gnerre, N. Patterson, A. Kong, D. Reich, and K. Stefansson. A direct characterization of human mutation based on microsatellites. *Nat Genet*, 44(10):1161–5, 2012.

[14] T. Willems, M. Gymrek, G. Highnam, Consortium Genomes Project, D. Mittelman, and Y. Erlich. The landscape of human str variation. *Genome Res*, 24(11):1894–904, 2014.

[15] R. Acuna-Hidalgo, J. A. Veltman, and A. Hoischen. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol*, 17(1):241, 2016.

[16] T. Willems, M. Gymrek, G. D. Poznik, C. Tyler-Smith, Y. Group Genomes Project Chromosome, and Y. Erlich. Population-scale sequencing data enable precise estimates of y-str mutation rates. *Am J Hum Genet*, 98(5):919–933, 2016.

[17] M. Gymrek, T. Willems, A. Guilmatre, H. Zeng, B. Markus, S. Georgiev, M. J. Daly, A. L. Price, J. K. Pritchard, A. J. Sharp, and Y. Erlich. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet*, 48(1):22–9, 2016.

[18] J. Quilez, A. Guilmatre, P. Garg, G. Highnam, M. Gymrek, Y. Erlich, R. S. Joshi, D. Mittelman, and A. J. Sharp. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and dna methylation in humans. *Nucleic Acids Res*, 44(8):3750–62, 2016.

[19] J. R. Tollervey, T. Curk, B. Rogelj, M. Briese, M. Cereda, M. Kayikci, J. Konig, T. Hortobagyi, A. L. Nishimura, V. Zupunski, R. Patani, S. Chandran, G. Rot, B. Zupan, C. E. Shaw, and J. Ule. Characterizing the rna targets and position-dependent splicing regulation by tdp-43. *Nat Neurosci*, 14(4):452–8, 2011.

[20] J. Hui, L. H. Hung, M. Heiner, S. Schreiner, N. Neumuller, G. Reither, S. A. Haas, and A. Bindereif. Intronic ca-repeat and ca-rich elements: a new class of regulators of mammalian alternative splicing. *EMBO J*, 24(11):1988–98, 2005.

[21] T. W. Hefferon, J. D. Groman, C. E. Yurk, and G. R. Cutting. 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*, 101(10):3504–9, 2004.

[22] S. M. Mirkin. Expandable dna repeats and human disease. *Nature*, 447(7147):932–40, 2007.

[23] J. S. Sutcliffe, D. L. Nelson, F. Zhang, M. Pieretti, C. T. Caskey, D. Saxe, and S. T. Warren. Dna methylation represses fmr-1 transcription in fragile x syndrome. *Hum Mol Genet*, 1(6):397–400, 1992.

[24] M. van Blitterswijk, M. DeJesus-Hernandez, and R. Rademakers. How do c9orf72 repeat expansions cause amyotrophic lateral sclerosis and frontotemporal dementia: can we learn from other noncoding repeat expansion disorders? *Curr Opin Neurol*, 25(6):689–700, 2012.

[25] T. G. Grunewald, V. Bernard, P. Gilardi-Hebenstreit, V. Raynal, D. Surdez, M. M. Aynaud, O. Mirabeau, F. Cidre-Aranaz, F. Tirode, S. Zaidi, G. Perot, A. H. Jonker, C. Lucchesi, M. C. Le Deley, O. Oberlin, P. Marec-Berard, A. S. Veron, S. Reynaud, E. Lapouble, V. Boeva, T. Rio Frio, J. Alonso, S. Bhatia, G. Pierron, G. Cancel-Tassin, O. Cussenot, D. G. Cox, L. M. Morton, M. J. Machiela, S. J. Chanock, P. Charnay, and O. Delattre. Chimeric ewsr1-fli1 regulates the ewing sarcoma susceptibility gene egr2 via a ggaa microsatellite. *Nat Genet*, 47(9):1073–8, 2015.

[26] J. L. Mountain, A. Knight, M. Jobin, C. Gignoux, A. Miller, A. A. Lin, and P. A. Underhill. Snpstrs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome Res*, 12(11):1766–72, 2002.

[27] S. A. Tishkoff, E. Dietzsch, W. Speed, A. J. Pakstis, J. R. Kidd, K. Cheung, B. Bonne-Tamir, A. S. Santachiara-Benerecetti, P. Moral, and M. Krings. Global patterns of linkage disequilibrium at the cd4 locus and modern human origins. *Science*, 271(5254):1380–7, 1996.

[28] B. A. Payseur, M. Place, and J. L. Weber. Linkage disequilibrium between strps and snps across the human genome. *Am J Hum Genet*, 82(5):1039–50, 2008.

[29] M. Gymrek, T. Willems, D. Reich, and Y. Erlich. Interpreting short tandem repeat variations in humans using mutational constraint. *Nat Genet*, 49(10):1495–1501, 2017.

[30] Y. Lai, D. Shinde, N. Arnheim, and F. Sun. The mutation process of microsatellites during the polymerase chain reaction. *J Comput Biol*, 10(2):143–55, 2003.

[31] Y. Lai and F. Sun. Microsatellite mutations during the polymerase chain reaction: mean field approximations and their applications. *J Theor Biol*, 224(1):127–37, 2003.

[32] T. Willems, D. Zielinski, J. Yuan, A. Gordon, M. Gymrek, and Y. Erlich. Genome-wide profiling of heritable and de novo str variations. *Nat Methods*, 14(6):590–592, 2017.

[33] L. G. Gouw, M. A. Castaneda, C. K. McKenna, K. B. Digre, S. M. Pulst, S. Perlman, M. S. Lee, C. Gomez, K. Fischbeck, D. Gagnon, E. Storey, T. Bird, F. R. Jeri, and L. J. Ptacek. Analysis of the dynamic mutation in the sca7 gene shows marked parental effects on cag repeat transmission. *Hum Mol Genet*, 7(3):525–32, 1998.

[34] M. D. Figley, A. Thomas, and A. D. Gitler. Evaluating noncoding nucleotide repeat expansions in amyotrophic lateral sclerosis. *Neurobiol Aging*, 35(4):936 e1–4, 2014.

[35] K. K. Ambrose, T. Ishak, L. H. Lian, K. J. Goh, K. T. Wong, A. Ahmad-Annuar, and M. K. Thong. Analysis of ctg repeat length variation in the dmpk gene in the general

population and the molecular diagnosis of myotonic dystrophy type 1 in malaysia. *BMJ Open*, 7(3):e010711, 2017.

[36] E. Majounie, M. Wardle, M. Muzaimi, W. C. Cross, N. P. Robertson, N. M. Williams, and H. R. Morris. Case control analysis of repeat expansion size in ataxia. *Neurosci Lett*, 429(1):28–32, 2007.

[37] H. Tang, E. F. Kirkness, C. Lippert, W. H. Biggs, M. Fabani, E. Guzman, S. Ramakrishnan, V. Lavrenko, B. Kakaradov, C. Hou, B. Hicks, D. Heckerman, F. J. Och, C. T. Caskey, J. C. Venter, and A. Telenti. Profiling of short-tandem-repeat disease alleles in 12,632 human whole genomes. *Am J Hum Genet*, 101(5):700–715, 2017.

[38] Consortium Genomes Project, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[39] M. Gymrek, D. Golan, S. Rosset, and Y. Erlich. lobstr: A short tandem repeat profiler for personal genomes. *Genome Res*, 22(6):1154–62, 2012.

[40] S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, 81(5):1084–97, 2007.

[41] M. D. Edge, B. F. B. Algee-Hewitt, T. J. Pemberton, J. Z. Li, and N. A. Rosenberg. Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. *Proc Natl Acad Sci U S A*, 114(22):5671–5676, 2017.

[42] Consortium International HapMap. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, 2005.

[43] B. A. Payseur and P. Jing. A genomewide comparison of population structure at strps and nearby snps in humans. *Mol Biol Evol*, 26(6):1369–77, 2009.

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

[45] A. Contente, A. Dittmer, M. C. Koch, J. Roth, and M. Dobbelstein. A polymorphic microsatellite that mediates induction of pig3 by p53. *Nat Genet*, 30(3):315–20, 2002.

[46] G. TEx Consortium. Human genomics. the genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–60, 2015.

[47] F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, and E. Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 2014.

[48] C. Borel, E. Migliavacca, A. Letourneau, M. Gagnebin, F. Bena, M. R. Sailani, E. T. Dermitzakis, A. J. Sharp, and S. E. Antonarakis. Tandem repeat sequence variation as causative cis-eqtls for protein-coding gene expression variation: the case of cstb. *Hum Mutat*, 33(8):1302–9, 2012.

[49] M. D. Lalioti, H. S. Scott, C. Buresi, C. Rossier, A. Bottani, M. A. Morris, A. Malafosse, and S. E. Antonarakis. Dodecamer repeat expansion in cystatin b gene in progressive myoclonus epilepsy. *Nature*, 386(6627):847–51, 1997.

[50] L. M. Brenman. Spinocerebellar ataxia type 6 (sca 6) phenotype in a patient with an intermediate mutation range cacna 1 a allele.

[51] A. D. Ha, C. A. Beck, and J. Jankovic. Intermediate cag repeats in huntington's disease: Analysis of cohort. *Tremor Other Hyperkinet Mov (N Y)*, 2, 2012.

[52] A. C. Wheeler, Jr. Bailey, D. B., E. Berry-Kravis, J. Greenberg, M. Losh, M. Mailick, M. Mila, J. M. Olichney, L. Rodriguez-Revenga, S. Sherman, L. Smith, S. Summers, J. C. Yang, and R. Hagerman. Associated features in females with an fmr1 premutation. *J Neurodev Disord*, 6(1):30, 2014.

[53] D. Y. Lee and C. T. McMurray. Trinucleotide expansion in disease: why is there a length threshold? *Curr Opin Genet Dev*, 26:131–40, 2014.

[54] R. Koide, T. Ikeuchi, O. Onodera, H. Tanaka, S. Igarashi, K. Endo, H. Takahashi, R. Kondo, A. Ishikawa, T. Hayashi, and et al. Unstable expansion of cag repeat in hereditary dentatorubral-pallidoluysian atrophy (drpla). *Nat Genet*, 6(1):9–13, 1994.

[55] L. Veneziano, E. Mantuano, C. Catalli, C. Gellera, A. Durr, S. Romano, M. Spadaro, M. Frontali, and A. Novelletto. A shared haplotype for dentatorubropallidoluysian atrophy (drpla) in italian families testifies of the recent introduction of the mutation. *J Hum Genet*, 59(3):153–7, 2014.

[56] J. M. Laffita-Mesa, J. M. Rodriguez Pupo, R. Moreno Sera, Y. Vazquez Mojena, V. Kouri, L. Laguna-Salvia, M. Martinez-Godales, J. A. Valdevila Figueira, P. O. Bauer, R. Rodriguez-Labrada, Y. Gonzalez Zaldivar, M. Paucar, P. Svenningsson, and L. Velazquez Perez. De novo mutations in ataxin-2 gene and als risk. *PLoS One*, 8(8):e70560, 2013.

[57] I. Paradisi, V. Ikonomu, and S. Arias. Huntington disease-like 2 (hdl2) in venezuela: frequency and ethnic origin. *J Hum Genet*, 58(1):3–6, 2013.

[58] S. R. Gan, W. Ni, Y. Dong, N. Wang, and Z. Y. Wu. Population genetics and new insight into range of cag repeats of spinocerebellar ataxia type 3 in the han chinese population. *PLoS One*, 10(8):e0134405, 2015.

[59] G. Gibson. Rare and common variants: twenty arguments. *Nat Rev Genet*, 13(2):135–45, 2012.

[60] C. Fuchsberger, J. Flannick, T. M. Teslovich, A. Mahajan, V. Agarwala, K. J. Gaulton, C. Ma, P. Fontanillas, L. Moutsianas, D. J. McCarthy, M. A. Rivas, J. R. B. Perry, X. Sim, T. W. Blackwell, N. R. Robertson, N. W. Rayner, P. Cingolani, A. E. Locke, J. F. Tajes, H. M. Highland, J. Dupuis, P. S. Chines, C. M. Lindgren, C. Hartl, A. U. Jackson, H. Chen, J. R. Huyghe, M. van de Bunt, R. D. Pearson, A. Kumar, M. Muller-Nurasyid, N. Grarup, H. M. Stringham, E. R. Gamazon, J. Lee, Y. Chen, R. A. Scott, J. E. Below, P. Chen, J. Huang, M. J. Go, M. L. Stitzel, D. Pasko, S. C. J. Parker, T. V. Varga, T. Green, N. L. Beer, A. G. Day-Williams, T. Ferreira, T. Fingerlin, M. Horikoshi, C. Hu, I. Huh, M. K. Ikram, B. J. Kim, Y. Kim, Y. J. Kim, M. S. Kwon, J. Lee, S. Lee, K. H. Lin, T. J. Maxwell, Y. Nagai, X. Wang, R. P. Welch, J. Yoon, W. Zhang, N. Barzilai, B. F. Voight, B. G. Han, C. P. Jenkinson, T. Kuulasmaa, J. Kuusisto, A. Manning, M. C. Y. Ng, N. D. Palmer, B. Balkau, A. Stancakova, H. E. Abboud, H. Boeing, V. Giedraitis, D. Prabhakaran, O. Gottesman, J. Scott, J. Carey, P. Kwan, G. Grant, J. D. Smith, B. M. Neale, S. Purcell, A. S. Butterworth, J. M. M. Howson, H. M. Lee, Y. Lu, S. H. Kwak, W. Zhao, J. Danesh, V. K. L. Lam, K. S. Park, D. Saleheen, et al. The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–47, 2016.

[61] Egor Dolzhenko, Joke JFA van Vugt, Richard J Shaw, Mitchell A Bekritsky, Marka van Blitterswijk, Giuseppe Narzisi, Subramanian S Ajay, Vani Rajan, Bryan R Lajoie, Nathan H Johnson, et al. Detection of long repeat expansions from pcr-free whole-genome sequence data. *Genome research*, 27(11):1895–1903, 2017.

[62] M. Bakhtiari, S. Shleizer-Burko, M. Gymrek, V. Bansal, and V. Bafna. Targeted genotyping of variable number tandem repeats with advntr. *bioRxiv*, 2018.

[63] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nat Genet*, 43(5):491–8, 2011.

[64] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Subgroup Genome Project Data Processing. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, 2009.

[65] H. Li. Tabix: fast retrieval of sequence features from generic tab-delimited files. *Bioinformatics*, 27(5):718–9, 2011.

[66] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. The ucsc table browser data retrieval tool. *Nucleic Acids Res*, 32(Database issue):D493–6, 2004.

[67] A. R. Quinlan and I. M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–2, 2010.

[68] S. A. Fisher, C. M. Lewis, and L. H. Wise. Detecting population outliers and null alleles in linkage data: application to gaw12 asthma studies. *Genet Epidemiol*, 21 Suppl 1:S18–23, 2001.

[69] O. Delaneau, J. Marchini, and J. F. Zagury. A linear complexity phasing method for thousands of genomes. *Nat Methods*, 9(2):179–81, 2011.

[70] J. O'Connell, D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, M. Cocca, M. Traglia, J. Huang, J. E. Huffman, I. Rudan, R. McQuillan, R. M. Fraser, H. Campbell, O. Polasek, G. Asiki, K. Ekoru, C. Hayward, A. F. Wright, V. Vitart, P. Navarro, J. F. Zagury, J. F. Wilson, D. Toniolo, P. Gasparini, N. Soranzo, M. S. Sandhu, and J. Marchini. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet*, 10(4):e1004234, 2014.

[71] T. J. Pemberton, C. I. Sandefur, M. Jakobsson, and N. A. Rosenberg. Sequence determinants of human microsatellite variability. *BMC Genomics*, 10:612, 2009.

[72] G. TEx Consortium, Data Analysis Laboratory, Group Coordinating Center Analysis Working, Group Statistical Methods groups Analysis Working, GTEx groups Enhancing, N. I. H. Common Fund, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, Ndri Biospecimen Collection Source Site, Rpci Biospecimen Collection Source Site, Vari Biospecimen Core Resource, Bank Brain Bank Repository-University of Miami Brain Endowment, Management Leidos Biomedical-Project, Elsi Study, Integration Genome Browser Data, E. B. I. Visualization, Integration Genome Browser Data, University of California Santa Cruz Visualization-Ucsc Genomics Institute, analysts Lead, Data Analysis Laboratory, Center Coordinating, N. I. H. program management, collection Biospecimen, Pathology, Q. T. L. manuscript working group e, A. Battle, C. D. Brown, B. E. Engelhardt, and S. B. Montgomery. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

[73] O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc*, 7(3):500–7, 2012.

Supplementary Information

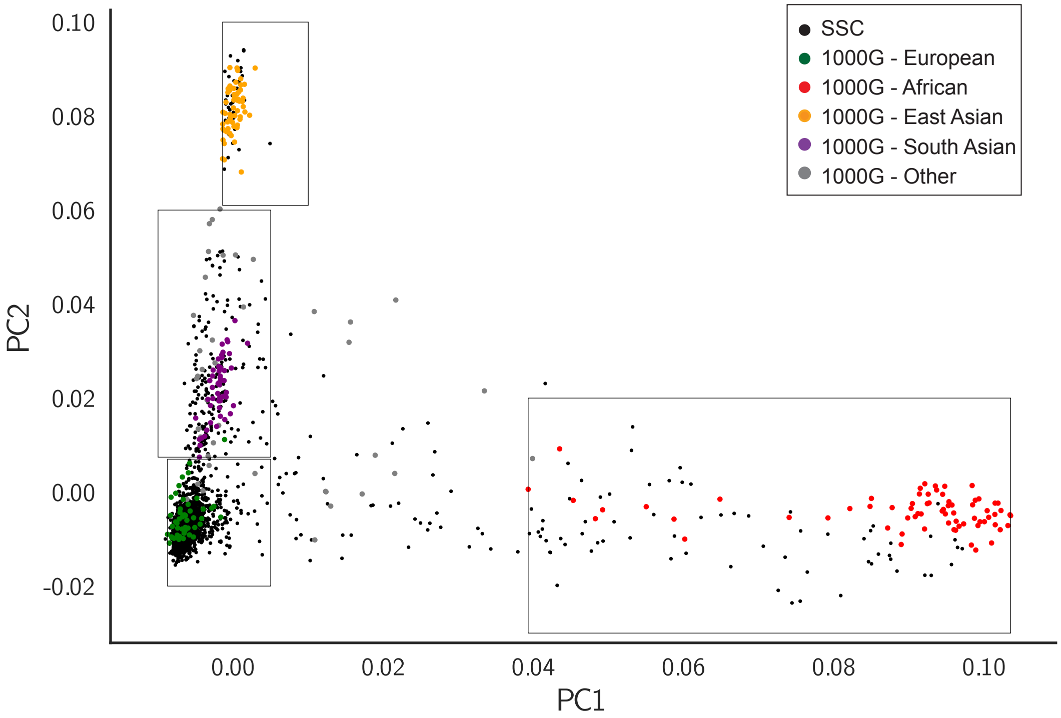


Figure S2.1: **Analysis of SSC populations.** Principal component analysis was performed using SNP genotypes from the SSC and 1000G cohorts. Boxes show inferred ancestry groups based on 1000 Genomes samples. Boxes for European, East Asian, South Asian, and African populations contain 1,585, 39, 172, and 69 SSC samples respectively. 51 SSC samples could not be confidently assigned to a population group.

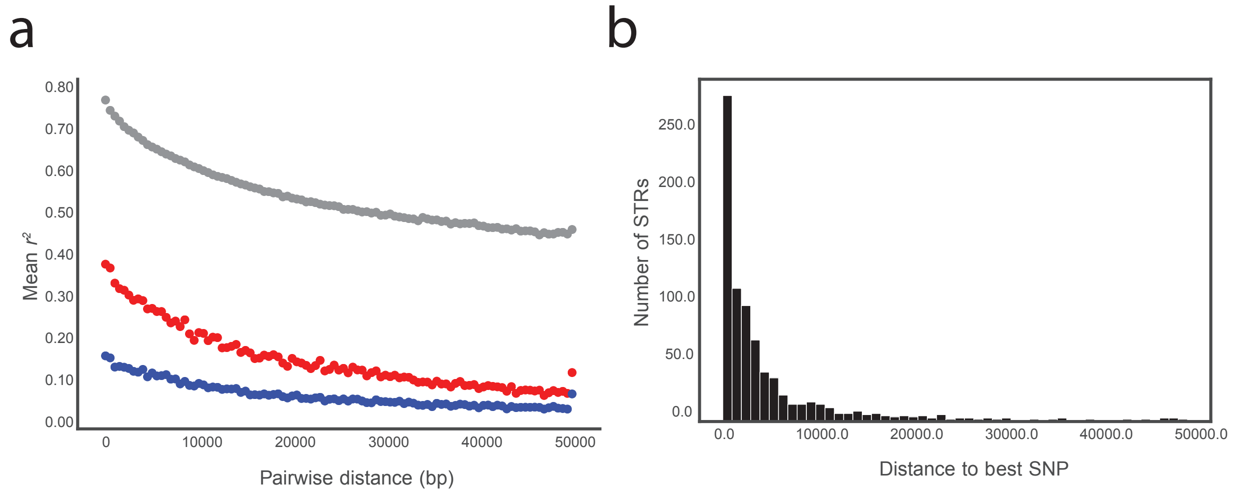


Figure S2.2: **a. SNP-SNP LD is stronger than STR-SNP LD.** Gray dots give the average pairwise SNP-SNP LD as a function of distance. Red dots give length r^2 computed as the squared Pearson correlation between STR length and SNP genotype (0, 1, 2). Blue dots give the allele r^2 , defined as the squared Pearson correlation between each SNP and each STR allele treated as a separate bi-allelic marker. **b. Distribution of distances from each STR to its best tag SNP.** The best tag SNP is defined as the SNP within 50kb with the highest length r^2 . The x-axis gives distance in bp.

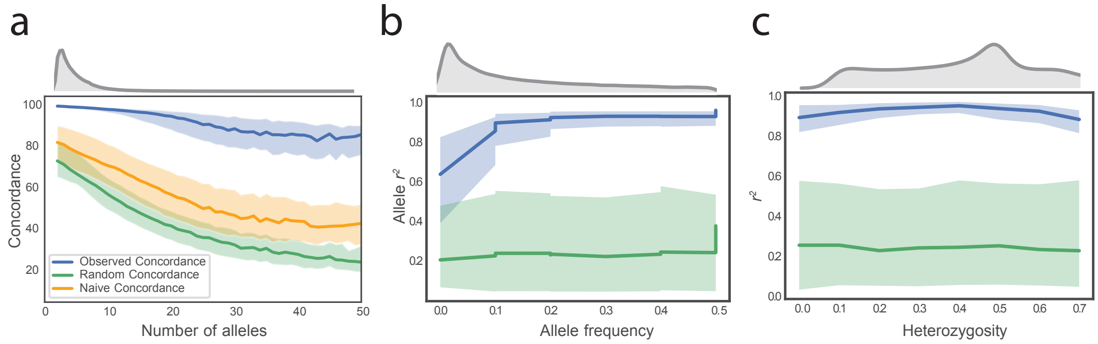


Figure S2.3: **Imputation performance is strongest at the least polymorphic STRs.** Plots show per-locus concordance vs. the number of alleles **(a)**, allelic r^2 vs. allele frequency **(b)**, and length r^2 vs. heterozygosity **(c)**. Upper gray plots give the relative frequency of points along the x-axis. Blue denotes observed per-locus values, green denotes values expected under a random model and orange denotes values expected under a naive model as defined in the **Online Methods.** Solid lines give median values for each bin and filled areas span the 25th to 75th percentile of values in each bin. X-axis values for **a. b.**, and **c.** were binned by 1, 0.05, and 0.1, respectively.

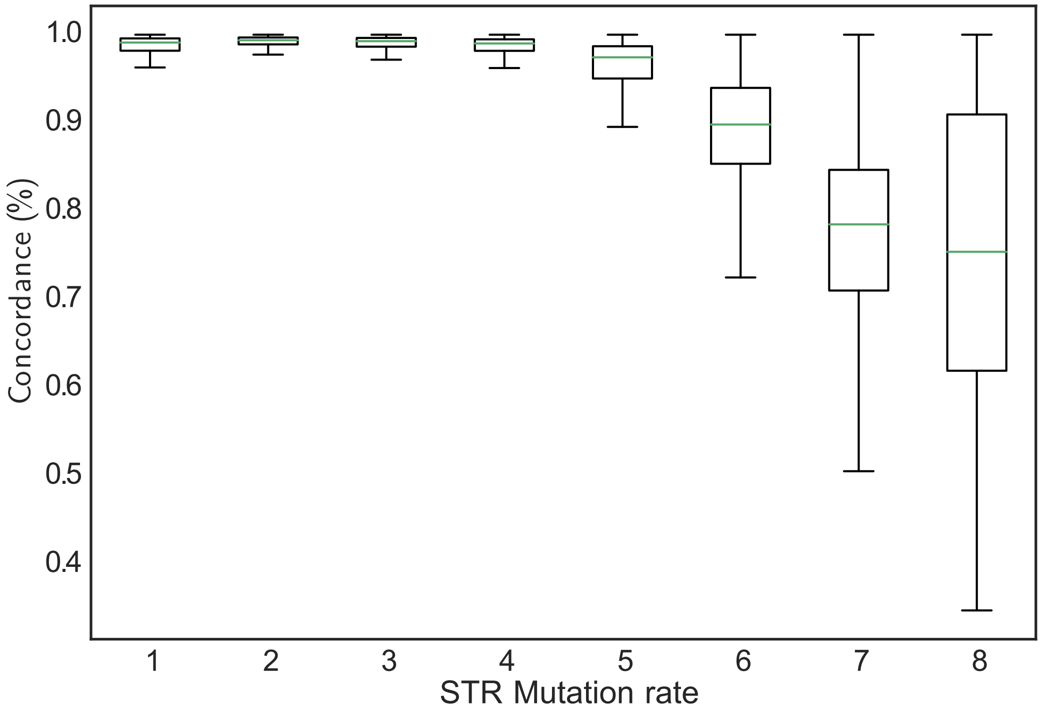


Figure S2.4: **Imputation concordance vs. mutation rate.** The x-axis gives the estimated mutation rate of each locus and y-axis gives concordance between imputed vs. HipSTR genotypes at each locus based on the leave-one-out analysis in SSC samples. Mutation rates were inferred by correlating local sequence heterozygosity with observed population-wide STR variation using the method described in Gymrek, et al . Green lines give median values. Boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1-1.5*IQR (bottom) and Q3+1.5*IQR, where IQR gives the interquartile range (Q3-Q1).

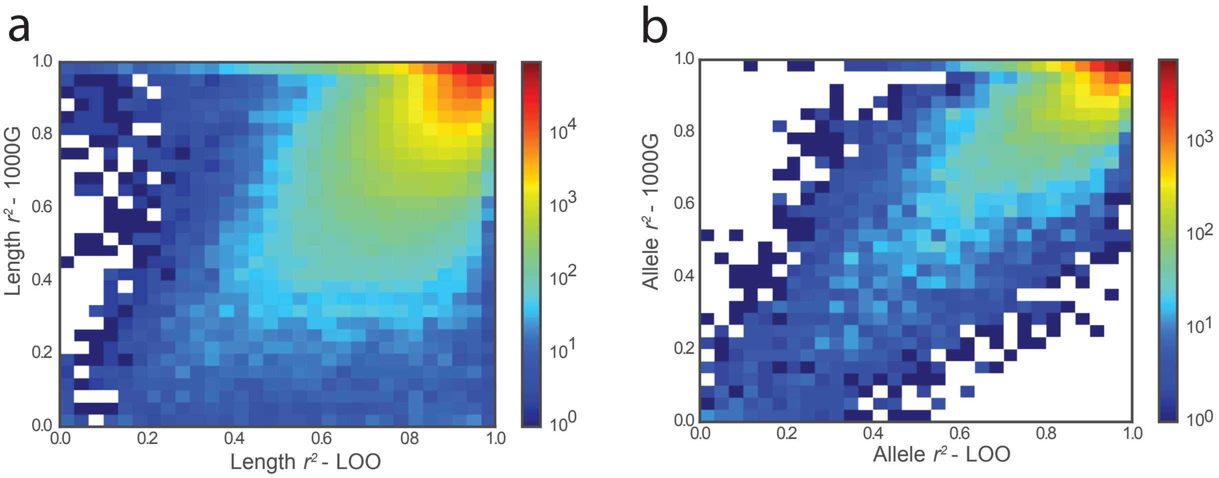


Figure S2.5: **Comparison of per-STR imputation metrics in the SSC dataset (leave-one-out analysis) vs.in the 1000 Genomes European samples. a.** and **b.** compare per-locus length r^2 and allelic r^2 , respectively. Color scale gives the $\log_{10}$ number of STRs represented in each bin. 1000 Genomes values are based on comparing HipSTR genotypes obtained from deep WGS for 49 European samples vs. STR genotypes imputed into 1000 Genomes Phase 3 SNP data obtained from low coverage WGS.

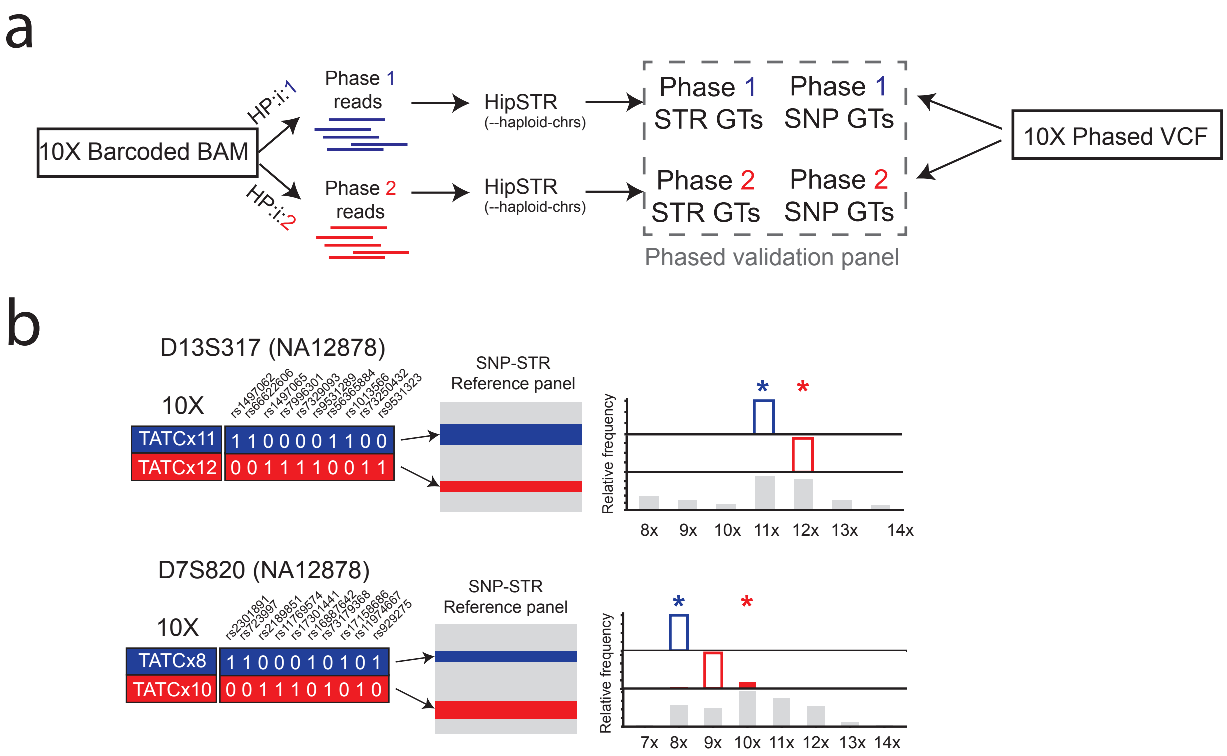


Figure S2.6: **Evaluating imputation and phasing accuracy using 10X Genomics. a. Schematic of pipeline to create a phased SNP-STR validation set in NA12878.** Barcoded BAMs were separated by phase and HipSTR was called in haploid mode separately on each set of reads. HipSTR genotypes from each read set were concatenated to form phased diploid genotypes. Phased STR and SNP genotypes were combined into a single phased validation panel. **b. Imputation vs. 10X results at example STRs.** Representative SNP-STR haplotypes are shown for NA12878 at two CODIS STRs. Blue denotes "phase 1" and red denotes "phase 2" as annotated in the 10X data. Values for each SNP (denoted by rsids) are 0 for the reference allele and 1 for the alternate allele. "10X" on the left denotes phased STR genotypes obtained using the pipeline in **a.** In each example all SNPs shown were identically genotyped in the 1000 Genomes Project panel and by 10X. Histograms on the right indicate STR allele frequencies in the SSC reference panel for the phase 1 SNP haplotype (blue), phase 2 SNP haplotype (red), and across the entire panel (gray). Filled bars give the imputed STR allele for each allele and stars give the expected value based on 10X genotypes. Both alleles at the top locus (D13S317) were imputed correctly. The second allele at the bottom locus (D7S820) was imputed incorrectly, likely because most haplotypes matching NA12878 contain 9, rather than 10, copies of TATC.

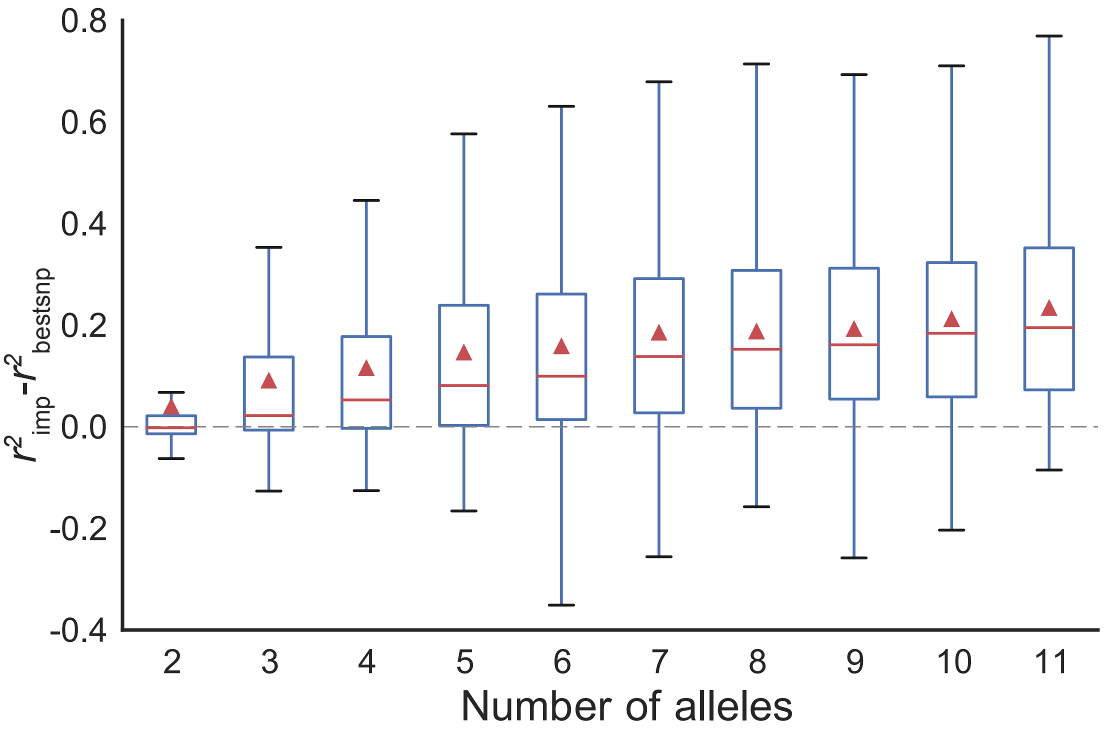


Figure S2.7: **Gain in length r^2 for imputed STR genotypes compared to the best tag SNP vs. number of STR alleles.** Data is shown for chr21 only. Red lines give medians and red triangles give mean values. Boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1-1.5*IQR (bottom) and Q3+1.5*IQR, where IQR gives the interquartile range (Q3-Q1). The best tag SNP is defined as the SNP within 50kb of the STR with the highest length r^2.

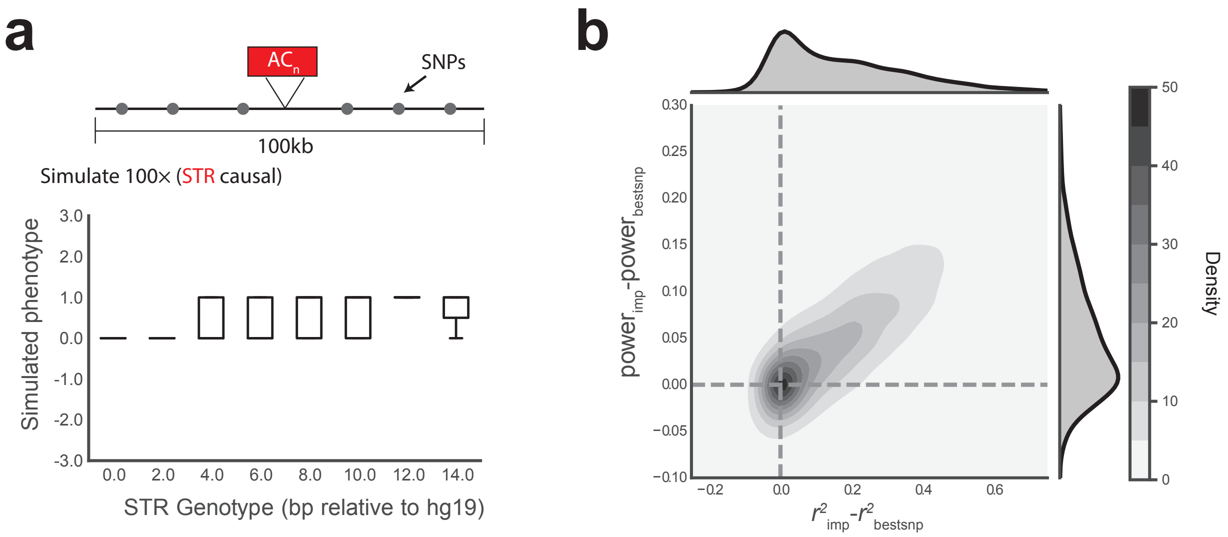


Figure S2.8: **STR imputation improves power to detect STR associations - case control phenotype. a. Example simulated case control phenotype.** Simulation is based on observed SSC STR genotypes. Boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1-1.5*IQR (bottom) and Q3+1.5*IQR, where IQR gives the interquartile range (Q3-Q1). For case/control simulations all phenotype values are either 0 or 1. **b. The gain in power using imputed genotypes is linearly related to the gain in r^2 compared to the best tag SNP.** Gray contours give the bivariate kernel density estimate. Top and right gray area gives the distribution of points along the x- and y-axes, respectively. Power was calculated based on the number of simulations out of 100 with nominal *p*-value ¡ 0.05.

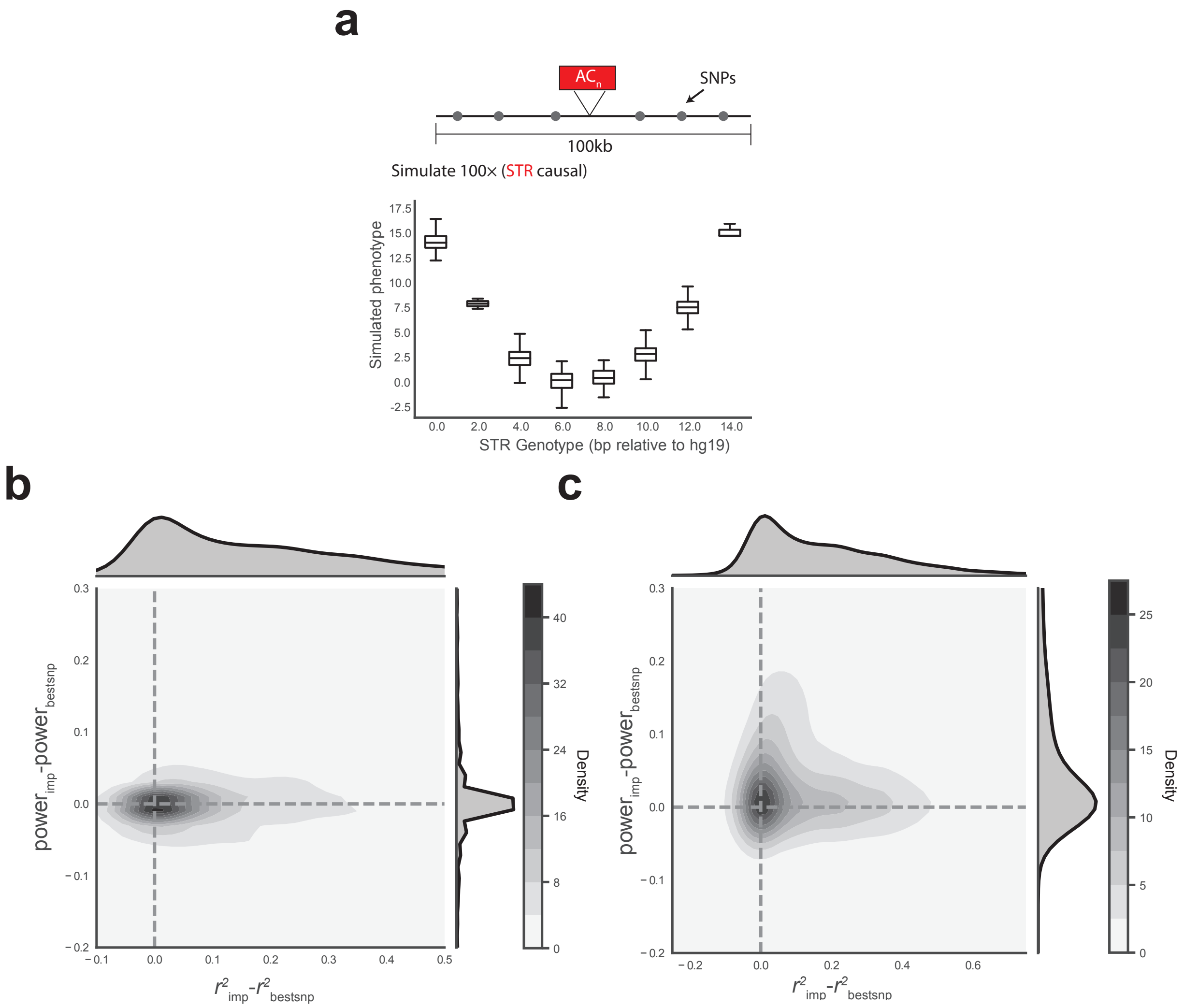


Figure S2.9: **STR imputation improves power to detect STR associations - non-additive phenotype model. a. Example simulated non-additive phenotype.** Simulation is based on observed SSC STR genotypes and uses a quadratic model as described in **Online Methods.** Black horizontal lines in the center of each box give median values. Boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1-1.5*IQR (bottom) and Q3+1.5*IQR, where IQR gives the interquartile range (Q3-Q1). **b. Gain in power using imputed genotypes compared to the best tag SNP.** STR association tests were conducted by regressing the imputed STR repeat dosage vs. phenotype. **c. Gain in power using per-allele STR association tests compared to the best tag SNP.** A separate association test was performed for each STR allele treating the allele as a bi-allelic marker. For the STR, power was determined using the most strongly associated allele. For **b.** and **c.**, gray contours give the bivariate kernel density estimate. Top and right gray area gives the distribution of points along the x- and y-axes, respectively. Power was calculated based on the number of simulations out of 100 with nominal *p*-value ¡ 0.05.

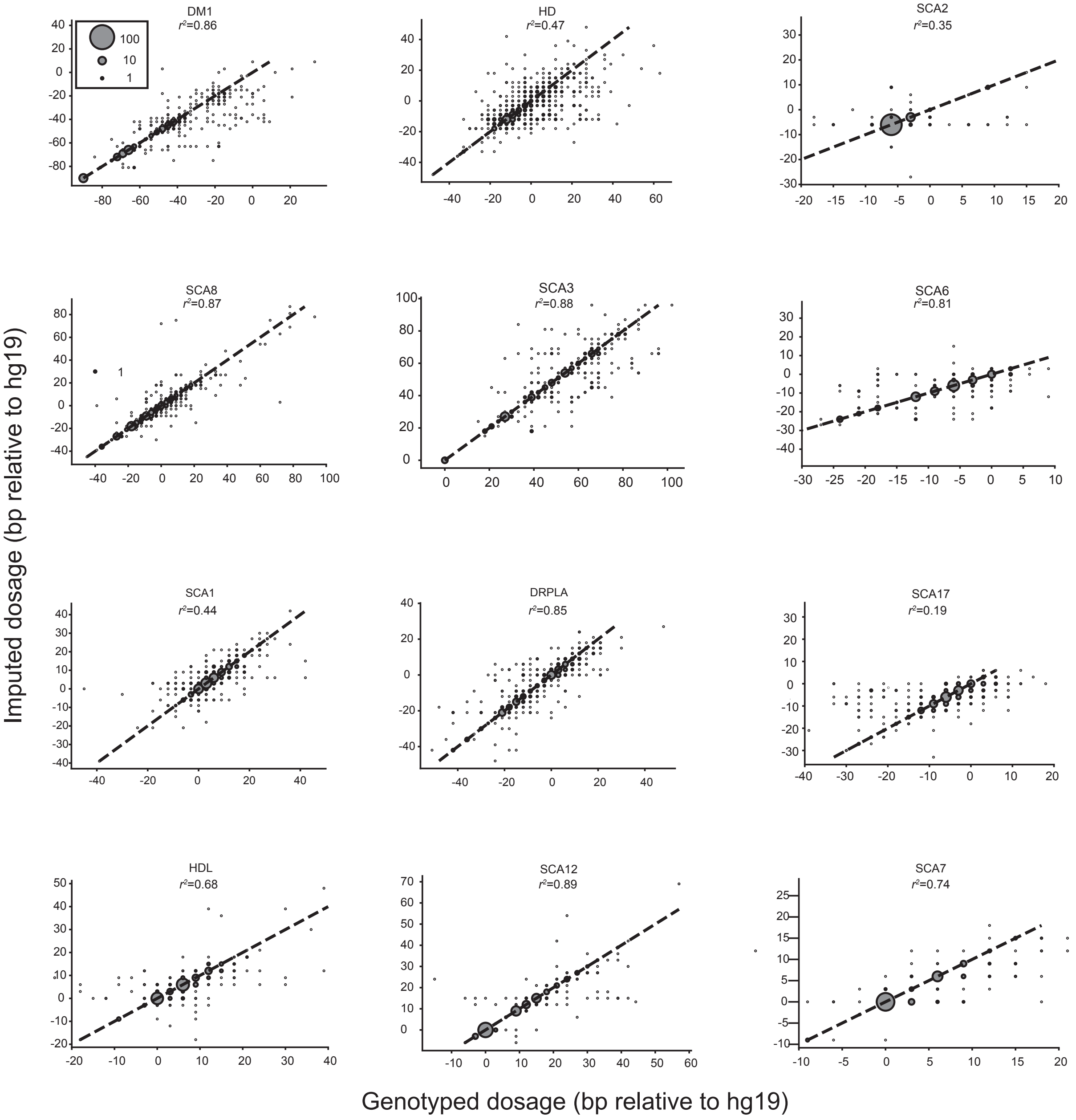


Figure S2.10: **STR imputation performance at known pathogenic STRs.** Panels show the genotyped vs. imputed dosage at each locus in the SSC cohort. Dashed lines give the diagonal. Bubble size scales with the number of points represented by each bubble as in **Figure 4.**

Table S2.1: **Set of known pathogenic STRs genotyped using Tredparse.** Concordance was estimated for the STRs genotyped using HipSTR. # Alleles gives the number of alleles occurring at least once in the Tredparse calls for SSC.

Coordinate (hg19)	Disorder	Locus ID	Motif	Call rate	# Alleles	Tredparse vs. HipSTR concordance
2:176957787	Syndactyly	SD5	GCN	0.01	6	
3:128891420	Myotonic dystrophy 2	DM2	CAGG	0	-	
3:138664863	Blepharophimosis, epicanthus inversus, and ptosis	BPES	NGC	0	-	
3:63898362	Spinocerebellar ataxia 7	SCA7	CAG	0.43	14	99.60%
4:3076604	Huntington disease	HD	CAG	0.99	28	
4:41747989	Central hypoventilation syndrome	CCHS	NGC	0	-	
5:146258292	Spinocerebellar ataxia 12	SCA12	CAG	0.65	21	99.90%
6:16327867	Spinocerebellar ataxia 1	SCA1	CAG	0.89	27	

6:170870996	Spinocerebellar ataxia 17	SCA17	CAG	0.85	34	
6:45390488	Cleidocranial dysplasia	CCD	GCN	0.13	9	
7:27239544	Hand-foot-uterus syndrome	HFG	GCN	0.01	5	
9:27573527	Amyotrophic lateral sclerosis	ALS	GGC CCC	0	-	
9:71652203	Friedreich ataxia	FRDA	GAA	0	-	
12:112036755	Spinocerebellar ataxia 2	SCA2	CAG	0.99	20	99.40%
12:7045892	Dentatorubral-pallidoluysian atrophy	DR-PLA	CAG	0.87	21	99.80%
13:100637703	Holoprosencephaly-5	HPE5	GCN	0.01	7	
13:70713516	Spinocerebellar ataxia 8	SCA8	CTG / CAG	0.95	36	
14:23790682	Oculopharyngeal muscular dystrophy	OPMD	GCN	0	-	
14:92537355	Spinocerebellar ataxia 3	SCA3	CAG	0.94	26	98.40%

16:87637894	Huntington disease-like 2	HDL	CTG	0.74	20	99.70%
19:13318673	Spinocerebellar ataxia 6	SCA6	CAG	0.9	12	
19:46273463	Myotonic dystrophy 1	DM1	CTG	1	33	99.30%
20:2633380	Spinocerebellar ataxia 36	SCA36	GGCCTG	0	-	
21:45196325	Unverricht-Lundborg Disease	ULD	CGCGGGGCGGGG	0	-	
22:46191235	Spinocerebellar ataxia 10	SCA10	ATTCT	0	-	

Table S2.2: **Imputation performance at CODIS markers.** Values were computed using leave-one-out analysis in the SSC cohort as described in the main text.

Position	ID	Length r^2	Concordance	Edge, et al. Concordance	# Alleles	Motif Length
5:149455884	CSF1PO	0.39	63%	60%	10	4
13:82722160	D13S317	0.75	69%	61%	10	4
18:60948895	D18S51	0.6	51%	32%	18	4
19:30417140	D19S433	0.61	70%	NA	15	4
3:45582231	D3S1358	0.66	67%	59%	8	4
5:123111245	D5S818	0.53	70%	60%	9	4
7:83789542	D7S820	0.71	70%	63%	8	4
8:125907107	D8S1179	0.78	69%	59%	10	4
4:155508888	FGA	0.6	48%	41%	17	4
15:97374244	PentaE	0.93	77%	NA	11	5

11:2192318	TH01	0.93	94%	83%	7	4
2:1493425	TPOX	0.87	90%	85%	7	4

Table S2.3: **Comparison of imputation performance in 1000 Genomes samples across genotyping platforms.** Mean concordance and length r^2 for the different datasets were found by comparing the imputed genotypes against the real genotypes called on high-coverage WGS samples using HipSTR.

	Platform	Number of samples	Mean conc.	Mean length r^2
1000 Genomes - EUR	WGS	49	97.00%	0.91
1000 Genomes - EUR	Affy 6.0	8	96.50%	0.86
1000 Genomes - EUR	Omni 2.5	50	96.70%	0.9
1000 Genomes - EAS	WGS	45	93.80%	0.79
1000 Genomes - EAS	Affy 6.0	18	92.40%	0.7
1000 Genomes - EAS	Omni 2.5	48	93.40%	0.77
1000 Genomes - AFR	WGS	46	90.60%	0.71
1000 Genomes - AFR	Affy 6.0	50	87.70%	0.6
1000 Genomes - AFR	Omni 2.5	0	-	-

Table S2.4: **Putative causal eSTRs.** CAVIAR score gives the posterior probability of causality of the STR. Best tag SNP gives the SNP within 50kb of the STR with highest length r^2.

Gene	STR (hg19)	CAVIAR score	HipSTR eSTR *p*-value	Imputed eSTR *p*-value	Best tag SNP	SNP-STR length r^2	Best tag SNP *p*-value
DSCR3	21:38733174	0.35	$2.05\text{x}10^{-4}$	$5.0\text{x}10^{-5}$	rs9976222	0.46	$3.5\text{x}10^{-3}$
CSTB	21:45196326	0.15	$1.36\text{x}10^{-12}$	$5.9\text{x}10^{-13}$	rs35285321	0.58	$3.2\text{x}10^{-9}$
C21orf62	21:34133199	0.094	$2.87\text{x}10^{-2}$	$2.9\text{x}10^{-1}$	rs9967977	0.77	0.5

Table S2.5: **Comparison of imputation performance using leave-one-out analysis on known pathogenic STRs called using both HipSTR and Tredparse.** Tredparse metrics were computed as described in the main text and Online Methods. For comparison, HipSTR metrics were re-computed by imputing each STR separately considering all SNPs within a 50kb region surrounding the STR. This was found to give slightly better imputation results compared to imputing all genome-wide STRs simultaneously as is done in the main text. [a]SCA=spinocerebellar ataxia; DRPLA=Dentatorubral-pallidoluysian Atrophy; DM1=Myotonic Dystrophy Type 1; HDL=Huntington's Disease-Like 2.

Locus	Disease[a]	HipSTR length r^2	HipSTR concordance	HipSTR # alleles	Tredparse length r^2	Tredparse concordance	Tredparse # alleles
3:63898362	SCA7	0.79	92.60%	10	0.8	91.90%	14
5:146258292	SCA12	0.9	94.90%	14	0.9	94.90%	21
12:112036755	SCA2	0.37	95.70%	13	0.48	96.20%	20
12:7045892	DRPLA	0.85	81.60%	19	0.85	81.20%	21
14:92537355	SCA3	0.86	87.10%	20	0.88	86.40%	26
16:87637894	HDL	0.65	88.50%	15	0.7	88.30%	20
19:46273463	DM1	0.88	85.40%	25	0.86	86.90%	33

Acknowledgements

Research reported in this publication was supported in part by the Office Of The Director, National Institutes of Health under Award Number DP5OD024577 and by a SFARI Explorer Award Number 515568. M.G. was supported in part by NIH/NIMH grant R01 MH113715. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) comet resource at the San Diego Supercomputing Center through allocations ddp268 and csd568. XSEDE is supported by National Science Foundation grant number ACI-1548562. We thank Alon Goren for helpful comments on the manuscript. We additionally thank Vineet Bafna and Vikas Bansal for helpful discussions and providing access to compute resources. We are grateful to all of the families that participated in the Simons Simplex Collection as well as the principal investigators.

Chapter 2, in full, contains 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.” Nature Communications (2018). I was a primary investigator and author of this paper.

Chapter 3

Genome-wide analysis of the contributions of short tandem repeat variants to schizophrenia risk

3.1 Introduction

Schizophrenia is a highly heritable disorder affecting around 1% of the population[1]. Genome-wide association studies (GWAS) based on common single nucleotide polymorphisms (SNPs) have now identified more than 200 independent genomic loci associated with schizophrenia risk[2, 3]. However, only a handful of loci identified so far point to a single plausible common SNP[3]. Indeed, dissection of the strongest association signal located in the major histocompatibility complex, revealed a poorly tagged multiallelic copy number variant in the gene C4 to be the causal variant[4]. Subsequent efforts in schizophrenia and other traits have often revealed complex variant types only partially tagged by SNPs to be driving many published association signals[5, 6, 7, 8].

Short tandem repeats (STRs), consisting of repeated motifs of 1-6bp, represent some of the most polymorphic regions of the human genome. Multiple lines of evidence point to a role for STRs in psychiatric disorders. STRs exhibit rapid mutation rates which often result in only moderate linkage disequilibrium (LD) with nearby SNPs. Further, variation in STR copy number has been shown to play a significant role in regulating gene expression[9, 10] and splicing[11]. Intriguingly, >30 Mendelian disorders, such as Huntington's Disease and Fragile X Syndrome, are caused by STR expansions[12]. Nearly all repeat disorders involve neurological phenotypes and many have psychiatric components[13, 14]. Finally, we and others have recently demonstrated the contribution of de novo mutations at STRs to autism spectrum disorders[15, 16].

Existing technologies have not allowed for systematic STR association studies. The majority of GWAS datasets have been generated with commodity genotype arrays, and cannot be used to directly analyze STR variants. STRs can be directly profiled from next-generation sequencing (NGS)[17, 18]. However, sample sizes of available NGS datasets for schizophrenia are still insufficient to detect common variants with modest effect size as expected based on the genetic architecture of schizophrenia. To overcome this challenge, we recently generated a reference STR+SNP haplotype panel that enables imputation of STR genotypes into SNP genotypes available for most GWAS cohorts[19]. Our imputation pipeline achieves an average imputed genotype concordance of 97% on European samples and can be used to impute nearly 500,000 STRs genome-wide.

Here, we leveraged our reference haplotype panel to impute STRs into GWAS data for more than 50,000 samples from the Psychiatric Genomics Consortium (PGC) to perform a genome-wide analysis of associations between STR lengths and schizophrenia. We performed statistical fine-mapping of previously published GWAS loci and identified five independent loci predicted to be driven by an underlying causal STR. We demonstrate that the second-strongest GWAS signal (after MHC) for schizophrenia may be driven by a penta-allelic tetranucleotide repeat associated with expression of the host gene for mir137, known to play a key role in multiple psychiatric disorders[20], and the nearby gene *DPYD*. Additionally, we show that a signal overlapping *AS3MT* previously suggested to be driven by multiple independent signals[2] can be explained by a single tri-allelic dinucleotide repeat. Taken together, these results reveal new potential biological mechanisms at these loci and highlight the need to consider additional variant types in future efforts to fine-map associations identified by GWAS.

3.2 Results

3.2.1 Performing a genome-wide STR association study for schizophrenia

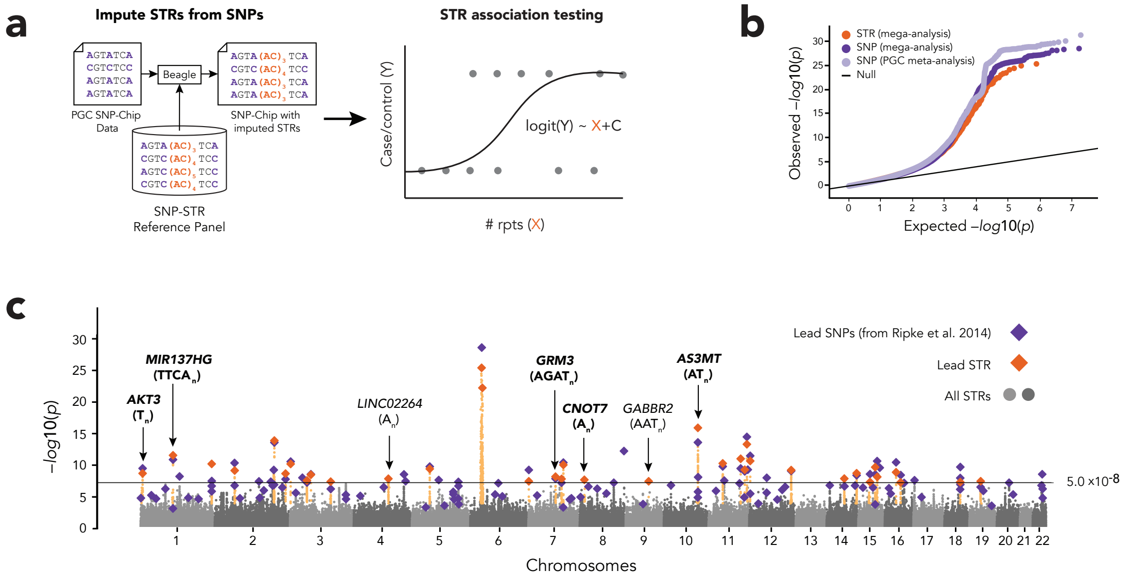


Figure 3.1: **STR-based genome-wide association testing in schizophrenia. a. Study overview:** we imputed genome-wide STRs into available SNP genotypes from the PGC cohort, and performed a logistic regression to test each STR for association with schizophrenia. **b. Quantile-quantile plot of STR and SNP associations.** The QQ plot shows the distribution of association *p*-values compared to the null expectation (light purple=published SNP summary statistics from the PGC SCZ2 cohort; dark purple=SNP mega-analysis, orange=STR mega-analysis). **c. Manhattan plot of STR associations:** The *x*-axis gives chromosome position and the *y*-axis gives the -$\log_{10}$ *p*-value for each tested STR. Purple diamonds show locations and *p*-values of previously published genome-wide significant loci based on SNPs. Orange diamonds show lead STRs at each LD-independent signal identified. The horizontal line shows the standard GWAS *p*-value threshold of 5×10^{-8}. Arrows denote significant loci only identified by STRs (non-bold) or loci for which fine-mapping indicated an STR as the most probable causal variant (bold).

We used Beagle[21] in combination with our published SNP+STR reference haplotype panel[19] to impute STR genotypes into genome-wide SNP data for 25,578 cases and 31,957 controls across 36 cohorts genotyped as part of the Psychiatric Genomics Consortium (PGC) **(Methods; Supplementary Table S3.1; Fig. 3.1a)**. Our reference panel focuses on STRs that can be reliably genotyped using short reads by HipSTR[17] and thus excluded most STRs known to be involved in repeat expansion disorders such as Huntington's Disease and hereditary ataxias[12]. We filtered STRs with low imputation quality in PGC and with minor allele frequency (MAF) $<5\%$ **(Methods)**. After filtering, 383,813 autosomal STRs remained for analysis.

We next used imputed STR genotypes to test each STR for association with schizophrenia in each cohort. As in previous studies of STR associations[9], we assumed an additive relationship between STR copy number and phenotype. For each diploid genotype, we computed a dosage score to account for uncertainty in the average copy number **(Methods)**. For each STR, we performed a logistic regression between the dosages in each person and case/control labels using a mega-analysis framework, controlling for population structure (top 10 SNP PCs) and cohort as covariates **(Methods, Fig. 3.1a, Supplementary Dataset 1)**. For comparison, we applied the same mega-analysis pipeline using SNPs **(Methods, Supplementary Dataset 2)**. We found that our results match closely to previous results on an overlapping dataset based on meta-analysis **(Supplementary Fig. S3.1)**.

Consistent with the known highly polygenic architecture of SCZ[2], STR association *p*-values showed strong departures from the null expectation (**Fig. 3.1b**; λ_{GC}=1.50). Our analysis identified 36 significant linkage-disequilibrium-independent STRs at $p < 5 \times 10^{-8}$ **(Supplementary Table S3.2)**. The majority of significant STR associations overlap with loci previously identified by SNP-based GWAS **(Fig. 3.1c; Methods)**. Four signals identified by STRs were not within 1 Mb of the lead SNPs for any of the previously reported significant loci based on meta-analysis of SNP genotypes in an overlapping dataset **(Supplementary Table S3.2)**. One of these falls in the complex HLA region on chromosome 6 and was excluded from further analysis. Significant SNP signals for the three remaining cases **(Supplementary Fig. S3.2)** have been identified by subsequent GWAS for SCZ[3, 22] indicating they are not truly novel loci.

To evaluate the robustness of our STR association signals, we repeated the analysis above using a meta-analysis framework. We analyzed each STR for association with SCZ in each cohort separately and combined results using an inverse-variance weighted fixed effects model[23]. Of the 36 LD-independent signals identified above, 30 passed genome-wide significance ($p < 5 \times 10^{-8}$) and all had $p < 4.33 \times 10^{-7}$ in the meta-analysis. Overall, our results suggest that STR association signals are robust and largely overlap with genomic regions previously identified using standard SNP-based GWAS.

3.2.2 Fine-mapping prioritize STRs at multiple SCZ-associated loci

We fine-mapped SCZ-associated loci to determine whether each signal could plausibly be explained by an underlying causal STR. To identify robustly fine-mapped STRs, we applied two orthogonal methods **(Methods)**. The first, FINEMAP[24], operates on association summary statistics and pairwise variant LD. The second, a Bayesian method based on a flat prior with steepest descent approximation[25], performs fine-mapping using individual-level genotype information for each variant. We refer to the latter method below as "fmgt". To further ensure the robustness of our fine-mapping results, we repeated each method using either best guess genotypes or genotype dosage values.

For each locus either previously implicated[2] or identified as meeting genome-wide significance based on our STR analysis, we applied fine-mapping to all STRs and SNPs within a 1 Megabase window centered at the index variant with $p < 10^{-6}$ **(Supplementary Table S3.3)**. At five loci, either all four fine-mapping settings identified the same STR as the variant with the highest posterior probability or at least two settings identified the same STR at posterior probability >50% **(Table ??, Supplementary Fig. S3.3)**. These include four intronic STRs (in *MIR137HG*, *GRM3*, *AS3MT*, and *AKT3*) and an STR in the 3'UTR of *CNOT7*.

Table 3.1: **Schizophrenia risk loci fine-mapped to STRs.** The table shows loci for which the same STR was the top variant in all four fine-mapping settings (*MIR137HG, GRM3, and CNOT7*) or for which at least two settings indicated the same STR at >50% posterior probability (*AS3MT and AKT3*). The lead SNP is that reported in Supplementary Table 2 of Ripke, et al. 2014. SNP and STR *p*-values are based on the mega-analysis performed here. Full fine-mapping results for each locus are provided in **Supplementary Table 3.**

Locus	Lead SNP (best fine-mapping posterior)	Lead SNP *p*-value	STR (best fine-mapping posterior*)	Annotation	Repeat unit	STR *p*-value
chr1:98001984-99001984	rs1702294 (1.54%)	1.19×10^{-11}	chr1:98506615 (69%)	*MIR137HG* (intron)	CATT	2.49×10^{-12}
chr1:243055105-244055105	rs77149735	4.15×10^{-8}	chr1:243671958 (77%)	*AKT3* (intron)	T	5.60×10^{-11}
chr7:85927626-86927626	rs12704290 (9.17%)	1.21×10^{-8}	chr7:86454144 (31%)	*GRM3* (intron)	AGAT	5.88×10^{-9}
chr8:16584523-17584523	NA	NA	chr8:17084523 (39%)	*CNOT7* (3'UTR)	A	1.82×10^{-8}
chr10:104457618-105457618	rs7907645	1.14×10^{-5}	chr10-104639652 (77%)	*AS3MT* (intron)	AT	1.13×10^{-16}
	rs55833108	1.42×10^{-6}				
	chr10-104957618-I (0.069%)	6.94×10^{-9}				
	rs11191419 (1.52%)	2.25×10^{-14}				

3.2.3 A candidate causal STR at the MIR137HG locus

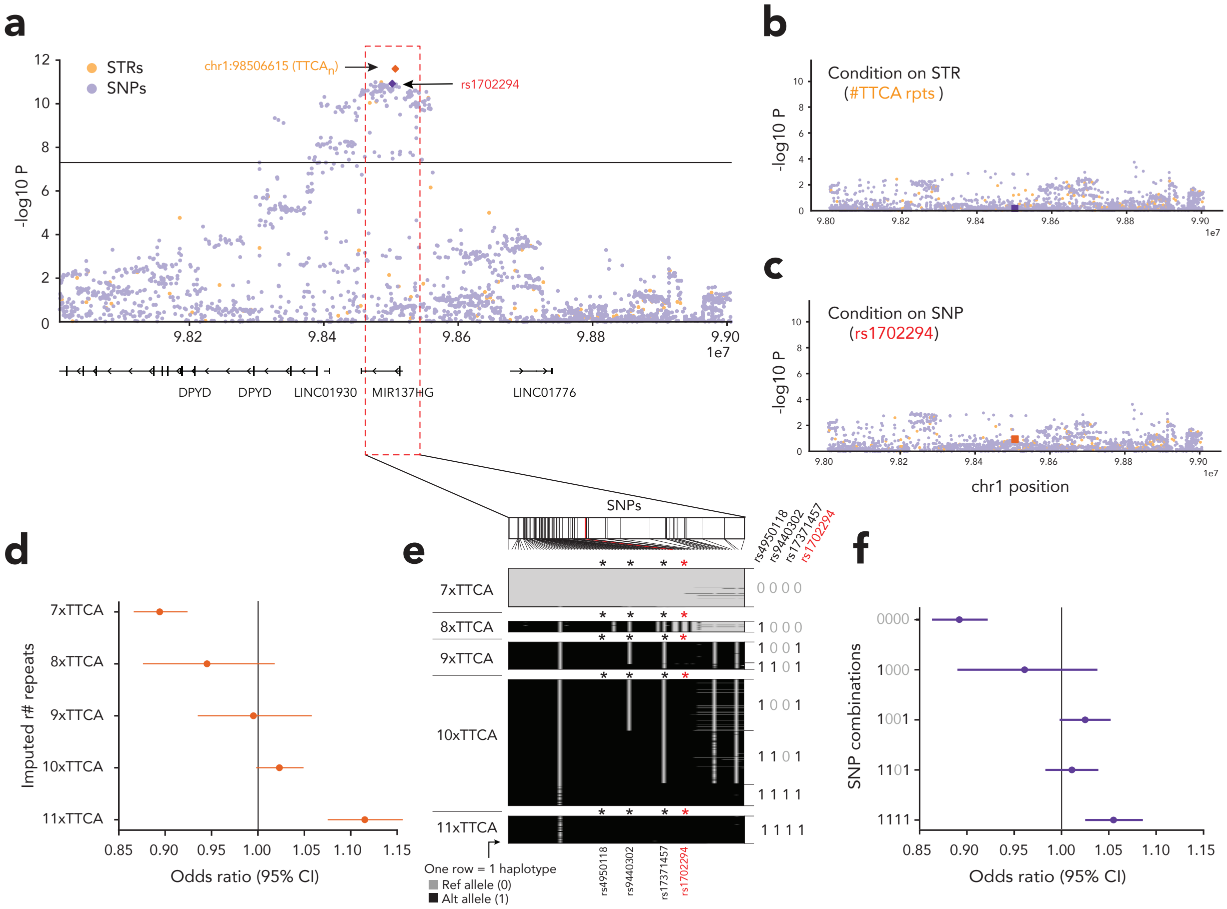


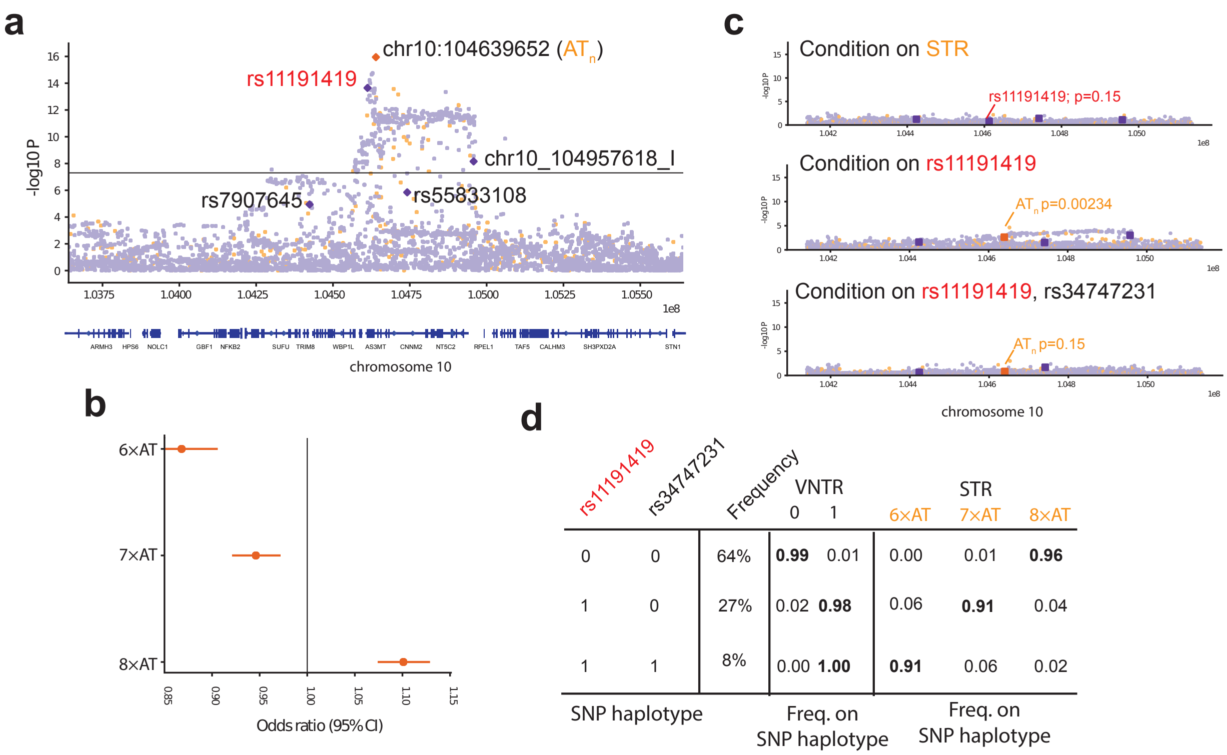
Figure 3.2: **Fine-mapping the MIR137HG locus. a. SNP and STR association signals at the MIR137HG locus.** The x-axis shows the position on chromosome 1 (hg19) and the y-axis shows -log$_{10}$ association p-values for SNPs (light purple) and STRs (light orange) based on mega-analysis. The lead SNP (rs1702294) and lead STR (chr1:98506615) are annotated. **b-c. Conditional regression analysis of the MIR137HG locus.** Plots show -log$_{10}$ association p-values after conditioning on the lead STR **(b)** and lead SNP **(c)** in the region. **d. Per-allele association tests.** The x-axis shows odds ratios and 95% confidence intervals from testing each allele length (7-11 TTCA repeats) for association with schizophrenia. **e. SNP-STR haplotypes.** Each column represents a SNP in the region highlighted by the dashed red box. Each row represents a single haplotype from a European individual from the 1000 Genomes Project. Gray or black denotes that a haplotype harbors the reference or alternate allele, respectively, at that SNP. Haplotypes are grouped by the number of copies of the tetranucleotide repeat. The index SNP is highlighted in red. **f. Haplotype association tests.** Four representative SNPs, including the index SNP, were used to perform haplotype association tests (0=ref allele, 1=alt allele). The x-axis shows the odds ratio and 95% confidence interval from testing each haplotype for association with schizophrenia. 4-SNP haplotypes and their relationship to each STR allele are annotated in **e.**

The role of the micro-RNA mir137 in psychiatric disorders has been extensively studied [20], and its corresponding locus on chromosome 1 **(Fig. 3.2a)** is one of the genomic regions most strongly associated with SCZ. The top variant identified in this region is a multi-allelic tetranucleotide (TTCA) repeat in the *MIR137* host gene (*MIR137HG*). Performing conditional regression conditioned on either the lead STR or the lead SNP (rs1702294) leaves no additional significant signals remaining **(Fig. 3.2b-c)** suggesting there is a single independent signal at this locus and reflecting the significant LD between the two variants (r^2=0.81). However, all four fine-mapping settings above indicated the STR as the top variant (67% and 10% posterior for FINEMAP and fmgt, respectively) for this locus **(Supplementary Fig. S3.3)**.

A previous study suggested a polymorphic VNTR upstream of *MIR137HG* may act as a causal variant at this locus. To test this hypothesis, we genotyped this VNTR in 503 European samples from the 1000 Genomes Project[26] using adVNTR[27], phased VNTR genotypes onto local SNP haplotypes, and imputed the VNTR into the PGC cohort **(Methods)**. VNTR length showed only modest LD with rs1702294 in the 1000 Genomes cohort (r^2=0.049). Further, imputed VNTR length did not show a significant association with SCZ risk ($p = 2.48 \times 10^{-2}$) suggesting it is unlikely to be the main causal variant driving this signal.

To further examine this association, we repeated association tests treating each STR allele (7-11 copies of TTCA) as a bi-allelic marker. These allele-specific association tests demonstrated a monotonically increasing risk for SCZ with each additional copy of the TTCA repeat **(Fig. 3.2d)**. We next examined the relationship between STR alleles and local SNP haplotypes **(Fig. 3.2e)** in European samples from the 1000 Genomes Project **(Methods)**. We found that rs1702294 tags short (7-8 copies) vs. long (9+ copies) of the repeat. On the other hand, multi-SNP haplotypes better tag individual STR alleles. Performing association tests with these haplotypes recapitulates the increasing trend between repeat copy number and SCZ risk **(Fig. 3.2e)**. Although we cannot rule out other variants not considered in our analysis, these results suggest the tetranucleotide STR as the most probable causal variant for this locus.

3.2.4 A candidate causal STR at the AS3MT locus



rs11191419	rs34747231	Frequency	VNTR 0	VNTR 1	STR 6×AT	STR 7×AT	STR 8×AT
0	0	64%	**0.99**	0.01	0.00	0.01	**0.96**
1	0	27%	0.02	**0.98**	0.06	**0.91**	0.04
1	1	8%	0.00	**1.00**	**0.91**	0.06	0.02

Figure 3.3: **Fine-mapping the AS3MT locus. a. SNP and STR association signals at the AS3MT locus.** The *x*-axis shows the position on chromosome 10 (hg19) and the *y*-axis shows -log$_{10}$ association *p*-values for SNPs (light purple) and STRs (light orange) based on mega-analysis. The lead SNP from PGC (rs11191419) and lead STR (chr10:104639652) are annotated. **b. Per-allele association tests.** The *x*-axis shows odds ratios and 95% confidence intervals from testing each allele length (6-8 AT repeats) for association with schizophrenia. **c. Conditional regression analysis of the AS3MT locus.** Plots show -log$_{10}$ association *p*-values after conditioning on the lead STR (top), the lead SNP (middle), or both the lead SNP and the remaining strongest SNP (bottom). **d. Haplotype analysis.** The table shows the three common two-SNP haplotypes for rs11191419 and rs34747231. "Frequency" gives the frequency of each SNP haplotype in Europeans. "VNTR" gives the frequency of the reference (0) and alternate (1) allele on each SNP haplotype. "STR" gives the frequency of each STR allele (6-8 repeats) on each SNP haplotype. Frequencies are based on European samples from the 1000 Genomes Project dataset.

Fine-mapping with fmgt indicated a tri-allelic dinucleotide ("AT") STR in an intron of AS3MT as the top signal with posterior probability >50% at the locus spanning chr10:104457618-105457618 **(Fig. 3.3a, Supplementary Fig. S3.3)**, which is the third most significant signal for schizophrenia. Testing each allele separately (6, 7, and 8 copies of AT) shows a monotonically increasing trend between repeat length and schizophrenia risk **(Fig. 3.3b)**. We further tested whether the STR best explains the signal in this region using conditional regression **(Fig. 3.3c)**. Conditioning on the STR leaves no remaining significant loci. On the other hand, after conditioning on the lead SNP reported by PGC, rs11191419, the STR remains nominally significant ($p = 0.0023$). Further conditioning on the top remaining SNP (rs34747231) leaves no remaining signal. Thus, whereas at least two independent SNPs are required, copy number variation at the STR alone is sufficient to explain the signal in this region, consistent with the fine-mapping results.

We investigated the SNP-STR haplotype structure at this locus using the two SNPs from our conditional regression above **(Fig. 3.3d)**. We found that the lead SNP from PGC (rs11191419) tags 8×AT vs. other alleles but cannot distinguish between 6-7×AT. On the other hand, rs34747231 tags 6×AT vs. 7-8×AT. While neither SNP alone is in strong LD with all three STR alleles, the three common two-SNP haplotypes correspond tightly with the three separate STR alleles. These results are consistent with our conditional regression analysis, in which both SNPs, but only a single STR, are needed to explain the signal.

Previous work suggested a bi-allelic 36-mer VNTR in the promoter of AS3MT as the likely causal variant for this locus[28]. To test whether this VNTR might best explain the GWAS signal, we genotyped it in European samples from the 1000 Genomes Project, imputed VNTR genotypes into PGC, and tested for association with schizophrenia. We found that the VNTR is in high LD (r^2=0.94) with rs11191419 **(Fig. 3.3d)**, which we showed above is not sufficient to explain the signal in this region. As expected, the VNTR is strongly associated with schizophrenia (mega-analysis $p = 2.17 \times 10^{-15}$). However, the association is not significant after conditioning on the STR (p=0.15). On the other hand, the STR association remains significant after conditioning on the VNTR (p=0.0047), suggesting the STR is a stronger candidate causal variant.

3.3 Discussion

This study uses our previously published SNP+STR reference haplotype panel to impute STR genotypes into genome-wide SNP data for 25,578 cases and 31,957 controls across 36 cohorts genotyped as part of the Psychiatric Genomics Consortium (PGC). We used imputed STR genotypes to test each STR for association with schizophrenia in each cohort. Consistent with the known highly polygenic architecture of SCZ, STR association *p*-values showed strong departures from the null expectation. Our analysis identified 36 significant linkage-disequilibrium-independent STRs and the majority of significant STR associations overlap with loci previously identified by SNP-based GWAS.

We fine-mapped SCZ-associated loci to determine whether each signal could plausibly be explained by an underlying causal STR using two orthogonal methods in four settings. At five loci, either all four fine-mapping settings identified the same STR as the variant with the highest posterior probability, or at least two settings identified the same STR at posterior probability $>50\%$. These include four intronic STRs (in *MIR137HG, GRM3, AS3MT, and AKT3*) and an STR in the 3'UTR of *CNOT7*.

The top variant identified in the micro-RNA mir137 region is a multi-allelic tetranucleotide (TTCA) repeat in the MIR137 host gene (*MIR137HG*). All four fine-mapping settings indicated the STR as the top variant. We show the allele-specific association tests demonstrate a monotonically increasing risk for SCZ with each additional copy of the TTCA repeat.

Further, fine-mapping of *AS3MT* locus indicated a tri-allelic dinucleotide ("AT") STR in an intron as the top signal. Testing each allele separately (6, 7, and 8 copies of AT) shows a monotonically increasing trend between repeat length and schizophrenia risk. We investigated the SNP-STR haplotype structure at this locus using the two SNPs from our conditional regression and found that the lead SNP from PGC (rs11191419) tags 8×AT vs. other alleles but cannot distinguish between 6-7×AT. On the other hand, rs34747231 tags 6×AT vs. 7-8×AT.

We identify the unavailability of pathogenic repeats expansion data and low accuracy STR imputation for highly polymorphic repeats as a limitation of this study. This aspect can be improved in the future with the availability of high coverage WGS data, allowing for direct genotyping of STRs. We also note that statistical fine-mapping is a very new field, with no clear protocols and best practices. While the availability of epigenomic data for SNP variants greatly improves their fine-mapping accuracy, the unavailability of similar data for STRs makes interpretation of fine-mapping results difficult. Overall, our STR imputation, GWAS, and fine-mapping framework allow for an entirely new class of variation to be interrogated by reanalyzing hundreds of thousands of existing datasets, with the potential to lead to novel genetic discoveries across a broad range of phenotypes.

3.4 Methods

3.4.1 Dataset and preprocessing

We used data from 36 European-ancestry cohorts with individual-level data available through PGC. The cohorts used are listed in **Supplementary Table S3.1**. As described previously [2], all subjects provided written and informed consent. This study was conducted in concordance with an analysis proposal approved by the PGC Schizophrenia Working Group. All analyses of individual-level genotype data were conducted on PGC's approved server in the Netherlands.

Published meta-analysis summary statistics were obtained from (`https://www.med.unc.edu/pgc/download-results/`).

3.4.2 STR Imputation

STR genotypes were imputed into SNP VCFs containing genotype data based on stringent quality filtering (".bgs" files from RICOPILI[29]) for all cohorts with Beagle version 5.1 using our published SNP+STR reference haplotype panel[19] based on 957 unrelated samples. Imputed VCFs were merged into a joint VCF file and split by chromosome using bcftools version 1.9 (`http://samtools.github.io/bcftools/bcftools.html`). We removed STRs for which the posterior probability of the best guess genotype had an average value across all samples of less than 0.5. We additionally removed STRs with MAF<5% in imputed genotypes.

3.4.3 STR association testing

We developed a custom utility (plinkSTR; `https://github.com/gymreklab/plinkSTR`) for performing association tests between STR length and a phenotype of interest. The script is modeled after plink[30], which currently does not support association tests based on STR dosage. PlinkSTR takes as input STR genotypes in VCF format, a covariates file, and a Plink FAM file.

We consider each STR genotype GT_i as the sum of the difference of allele lengths from hg19 reference allele. To account for uncertainty in imputed genotypes, we computed STR dosages rather than using hard genotype calls. Dosages were computed as:

$$Pr(GT_1) * GT_1 + Pr(GT_2) * GT_2 + Pr(GT_3) * GT_3 + ...$$

Where $Pr(GT_i)$ is based on the posterior probability of each genotype reported by Beagle.

For case-control phenotypes, plinkSTR uses the Python statsmodel[31] library's Logit module to perform logistic regression. PlinkSTR outputs an odds ratio, standard error, and p-value for each STR tested. We used plinkSTR to perform STR-based GWAS in the PGC data separately for each cohort using the top 10 population PCs (obtained from the PGC SCZ working group) and cohort as covariates. At each locus, we filtered individuals with outlier genotypes with dosages more than (1.5*IQR above the third quartile or the maximum dosage with 100 samples, whichever if more) or (1.5*IQR below the first quartile or the minimum dosage with 100 samples, whichever is less).

For meta-analysis, we analyzed each cohort separately using PlinkSTR and used METAL [23] (release 2011-03-25) to perform a meta-analysis across all cohorts using default options. Genome-wide STR summary statistics for STR mega- and meta- analyses are available in **Supplementary Dataset 1**.

To obtain a list of LD-independent STR associations, we created a custom utility (plinkSTR-clump.py in the plinkSTR package), modeled after plink –clump utility, which can handle STR genotypes. plinkSTR_clump.py takes as input a VCF file with STR genotypes and a summary statistics file produced by plinkSTR. It computes pairwise LD based on the correlation between STR genotype lengths at a pair of STR loci. plinkSTR_clump.py was run using the following options: –clump-p1 0.000001 –clump-p2 0.000 –clump-r2 0 –clump-kb 3000.

3.4.4 SNP association testing

Published SNP summary statistics for the PGC SCZ2 cohort were computed using a "meta-analysis" framework in which association testing is first performed separately in each cohort and then combined. Further, it included some cohorts for which individual-level genotypes were not available. To enable a more direct comparison of our STR results to SNP-based GWAS, we recomputed SNP association statistics using an identical set of cohorts and covariates as used for our STR analysis.

SNP genotypes were obtained from genotype data based on best guess genotypes with moderate quality filtering (".bgn" files from RICOPILI[29]) for all cohorts. All cohorts were merged into a single plink BED file using plink –bmerge utility. We used plink v1.90[30] for performing association tests between SNP genotypes and the phenotype. plink takes as input SNP genotypes in a binary BED file format and a covariates file. To perform logistic regression, we used the plink –logit method that outputs an odds ratio, standard error, and *p*-value using the top 10 population PCs (obtained from PGC SCZ working group) and cohort as covariates.

3.4.5 Fine-mapping STRs

We used FINEMAP v1.4[24] to fine map association signals. We considered all previously reported 108 loci based on SNP meta-analysis in this cohort2 as well as the 3 additional LD-independent STR signals identified here. For each locus, we considered all variants within a 1Mb window centered at the lead variant with mega-analysis $p < 10^{-6}$. Loci for which no STR reached this threshold were removed from further analysis. We further excluded the MHC region, which was fine-mapped previously to a multi-allelic CNV[2, 4].

FINEMAP takes as input an LD file with pairwise LD (pearson correlation coefficient), and a Z file with summary statistics data for each variant. We use a custom python script (generate_finemap_files.py) to extract SNP and STR genotypes from VCF files, extract summary statistics data for each variant, and find LD between each pair of variants. We ran FINEMAP using default options which sets the maximum number of allowed causal SNPs to 5. We report the model-averaged posterior summaries for the top SNP and top STR at each locus in **Supplementary Table S3.3**.

We additionally applied a previously described fine-mapping method[25] that uses individual genotype data (https://github.com/hailianghuang/Fine-mapping), which refer to here as fmgt We modified the existing method (https://github.com/shubhamsaini/Fine-mapping) to allow one phenotype as input by removing any dependencies on the second phenotype. We ran fmgt on an imputed set of SNPs and STRs obtained from Beagle during STR imputation process described previously since fmgt does not let us work with missing genotype data which is a limitation of underlying R nnet [32] library. We additionally included identical covariates (cohort and population PCs) as were used in the original mega-analysis (note, FINEMAP, which is based on summary statistics, cannot handle covariates). For each locus, we report the posterior probability for the top SNP and top STR based on the best model identified by fmgt.

3.4.6 Conditional Regression

We used plinkSTR to perform conditional regression analysis. plinkSTR accepts a comma separated list of variant positions to condition on and the conditional variants are included as covariates in addition to the principal components and cohort information. The rest of the process is done like regular STR case-control association testing as described previously.

3.4.7 Haplotype Association Tests

We used 1000 Genomes Project phased SNP+STR data published previously [19] for 503 samples of European ancestry to determine haplotypes that harbor distinct STR alleles. We train a ElasticNet regression model using Python scikit-learn [33] library on SNP haplotypes spanning

500kb around the STR. We consider each SNP as an input variable and the STR allele as the output variable. We choose haplotypes consisting of 20 SNPs with the greatest absolute effect sizes and the lead SNPs from SCZ2 [2] study, and group them for each unique STR allele.

Next, we select a minimal subset of SNPs from these haplotypes that uniquely identify STR alleles. We use these minimal subset to run a haplotype based association test using plinkSTR by regressing these haplotypes against the phenotype. For the haplotype based association test, we consider the number of copies of each haplotype as the genotype for each sample.

3.4.8 Analysis of target STRs and VNTRs in the 1000 Genomes Project

We used 503 high-coverage 1000 Genomes data samples of European ancestry generated by New York Genome Center. Using adVNTR[27] version 1.4.0, we genotyped the VNTRs in the *MIR137HG* (chr1:98046173-98046233, hg38) and *AS3MT* (chr10:102869497-102869605, hg38) region.

We next merged these VNTRs with our published SNP+STR[19] phased haplotype reference panel using bcftools version 1.9 (`http://samtools.github.io/bcftools/bcftools.html`), and phased the VNTRs onto the SNP+STR haplotypes using Beagle version 5.1.

Data Availability

STR and SNP summary statistics are available in Supplementary Datasets 1-2. Upon acceptance of this study for publication, individual-level STR genotypes will be made available through PGC.

Code Availability

The plinkSTR tool is available on Github: `https://github.com/gymreklab/plinkSTR`.

Bibliography

[1] World Health Organization et al. *The global burden of disease: 2004 update*. World Health Organization, 2008.

[2] Stephan Ripke, Benjamin M Neale, Aiden Corvin, James TR Walters, Kai-How Farh, Peter A Holmans, Phil Lee, Brendan Bulik-Sullivan, David A Collier, Hailiang Huang, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421, 2014.

[3] Max Lam, Chia-Yen Chen, Zhiqiang Li, Alicia R Martin, Julien Bryois, Xixian Ma, Helena Gaspar, Masashi Ikeda, Beben Benyamin, Brielin C Brown, et al. Comparative genetic architectures of schizophrenia in east asian and european populations. *Nature genetics*, 51(12):1670–1678, 2019.

[4] Aswin Sekar, Allison R Bialas, Heather De Rivera, Avery Davis, Timothy R Hammond, Nolan Kamitaki, Katherine Tooley, Jessy Presumey, Matthew Baum, Vanessa Van Doren, et al. Schizophrenia risk from complex variation of complement component 4. *Nature*, 530(7589):177–183, 2016.

[5] Ellen M Leffler, Gavin Band, George BJ Busby, Katja Kivinen, Quang Si Le, Geraldine M Clarke, Kalifa A Bojang, David J Conway, Muminatou Jallow, Fatoumatta Sisay-Joof, et al. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science*, 356(6343), 2017.

[6] Linda M Boettger, Rany M Salem, Robert E Handsaker, Gina M Peloso, Sekar Kathiresan, Joel N Hirschhorn, and Steven A McCarroll. Recurring exon deletions in the hp (haptoglobin) gene contribute to lower blood cholesterol levels. *Nature genetics*, 48(4):359–366, 2016.

[7] Janet HT Song, Craig B Lowe, and David M Kingsley. Characterization of a human-specific tandem repeat associated with bipolar disorder and schizophrenia. *The American Journal of Human Genetics*, 103(3):421–430, 2018.

[8] Ronen E Mukamel, Robert E Handsaker, Maxwell A Sherman, Alison R Barton, Yiming Zheng, Steven A McCarroll, and Po-Ru Loh. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *bioRxiv*, 2021.

[9] Stephanie Feupe Fotsing, Jonathan Margoliash, Catherine Wang, Shubham Saini, Richard Yanicky, Sharona Shleizer-Burko, Alon Goren, and Melissa Gymrek. The impact of short tandem repeat variation on gene expression. *Nature genetics*, 51(11):1652–1659, 2019.

[10] Javier Quilez, Audrey Guilmatre, Paras Garg, Gareth Highnam, Melissa Gymrek, Yaniv Erlich, Ricky S Joshi, David Mittelman, and Andrew J Sharp. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and dna methylation in humans. *Nucleic acids research*, 44(8):3750–3762, 2016.

[11] Timothy W Hefferon, Joshua D Groman, Catherine E Yurk, and Garry R Cutting. A variable dinucleotide repeat in the cftr gene contributes to phenotype diversity by forming rna secondary structures that alter splicing. *Proceedings of the National Academy of Sciences*, 101(10):3504–3509, 2004.

[12] Anthony J Hannan. Tandem repeats mediating genetic plasticity in health and disease. *Nature Reviews Genetics*, 19(5):286, 2018.

[13] Camille L Julien, Jennifer C Thompson, Sue Wild, Pamela Yardumian, Julie S Snowden, Gwen Turner, and David Craufurd. Psychiatric disorders in preclinical huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 78(9):939–943, 2007.

[14] Naoto Adachi, Kunimasa Arima, Takashi Asada, Masaaki Kato, Narihiro Minami, Yu-ichi Goto, Teiichi Onuma, Takeshi Ikeuchi, Shoji Tsuji, Masahiro Hayashi, et al. Dentatorubral-pallidoluysian atrophy (drpla) presenting with psychosis. *The Journal of neuropsychiatry and clinical neurosciences*, 13(2):258–260, 2001.

[15] Brett Trost, Worrawat Engchuan, Charlotte M Nguyen, Bhooma Thiruvahindrapuram, Egor Dolzhenko, Ian Backstrom, Mila Mirceta, Bahareh A Mojarad, Yue Yin, Alona Dov, et al. Genome-wide detection of tandem dna repeats that are expanded in autism. *Nature*, 586(7827):80–86, 2020.

[16] Ileena Mitra, Bonnie Huang, Nima Mousavi, Nichole Ma, Michael Lamkin, Richard Yanicky, Sharona Shleizer-Burko, Kirk E Lohmueller, and Melissa Gymrek. Patterns of de novo tandem repeat mutations and their role in autism. *Nature*, 589(7841):246–250, 2021.

[17] Thomas Willems, Dina Zielinski, Jie Yuan, Assaf Gordon, Melissa Gymrek, and Yaniv Erlich. Genome-wide profiling of heritable and de novo str variations. *Nature methods*, 14(6):590–592, 2017.

[18] Egor Dolzhenko, Mark F Bennett, Phillip A Richmond, Brett Trost, Sai Chen, Joke JFA van Vugt, Charlotte Nguyen, Giuseppe Narzisi, Vladimir G Gainullin, Andrew M Gross, et al. Expansionhunter denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome biology*, 21(1):1–14, 2020.

[19] Shubham Saini, Ileena Mitra, Nima Mousavi, Stephanie Feupe Fotsing, and Melissa Gymrek. A reference haplotype panel for genome-wide imputation of short tandem repeats. *Nature communications*, 9(1):1–11, 2018.

[20] E Mahmoudi and MJ Cairns. Mir-137: an important player in neural development and neoplastic transformation. *Molecular psychiatry*, 22(1):44–55, 2017.

[21] Brian L Browning, Ying Zhou, and Sharon R Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.

[22] Zhiqiang Li, Jianhua Chen, Hao Yu, Lin He, Yifeng Xu, Dai Zhang, Qizhong Yi, Changgui Li, Xingwang Li, Jiawei Shen, et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nature genetics*, 49(11):1576–1583, 2017.

[23] Cristen J Willer, Yun Li, and Gonçalo R Abecasis. Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 2010.

[24] Christian Benner, Chris CA Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, 2016.

[25] Hailiang Huang, Ming Fang, Luke Jostins, Maša Umićević Mirkov, Gabrielle Boucher, Carl A Anderson, Vibeke Andersen, Isabelle Cleynen, Adrian Cortes, François Crins, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*, 547(7662):173–178, 2017.

[26] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[27] Mehrdad Bakhtiari, Sharona Shleizer-Burko, Melissa Gymrek, Vikas Bansal, and Vineet Bafna. Targeted genotyping of variable number tandem repeats with advntr. *Genome research*, 28(11):1709–1719, 2018.

[28] Ming Li, Andrew E Jaffe, Richard E Straub, Ran Tao, Joo Heon Shin, Yanhong Wang, Qiang Chen, Chao Li, Yankai Jia, Kazutaka Ohi, et al. A human-specific as3mt isoform and borcs7 are molecular risk factors in the 10q24. 32 schizophrenia-associated locus. *Nature medicine*, 22(6):649, 2016.

[29] Max Lam, Swapnil Awasthi, Hunna J Watson, Jackie Goldstein, Georgia Panagiotaropoulou, Vassily Trubetskoy, Robert Karlsson, Oleksander Frei, Chun-Chieh Fan, Ward De Witte, et al. Ricopili: rapid imputation for consortias pipeline. *Bioinformatics*, 36(3):930–933, 2020.

[30] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[31] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. Austin, TX, 2010.

[32] Brian Ripley, William Venables, and Maintainer Brian Ripley. Package ‘nnet’. *R package version*, 7:3–12, 2016.

[33] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Supplementary Information

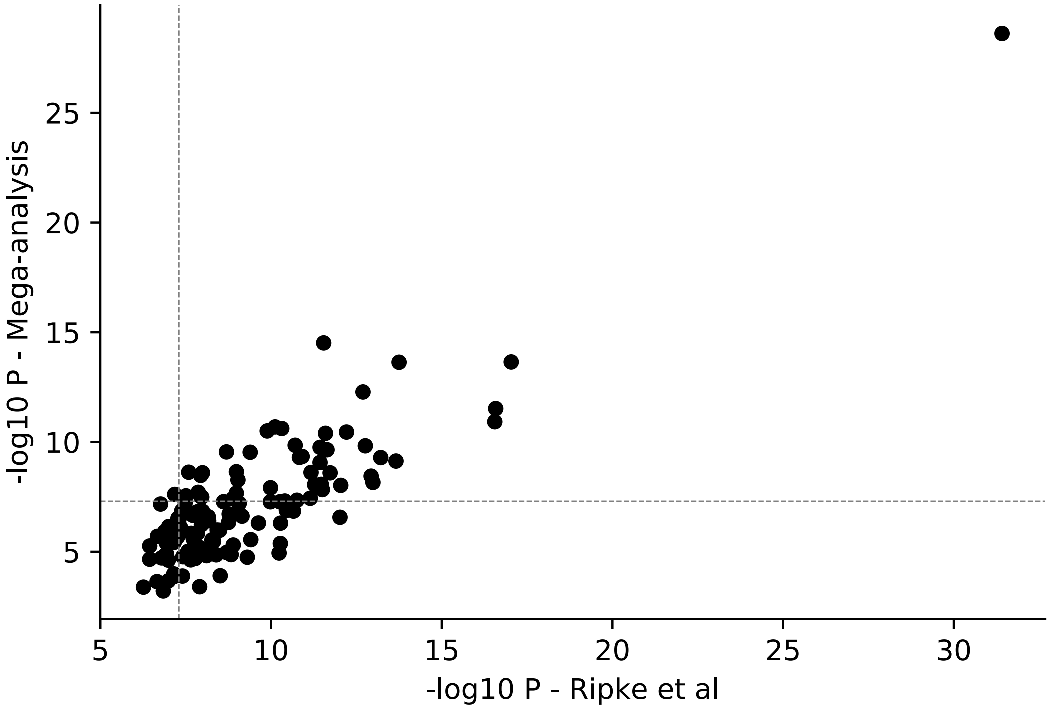


Figure S3.1: **Comparison of SNP p-values obtained by meta vs. mega analysis.** The *x*-axis gives published meta-analysis -$\log_{10}$ *p*-values from the PGC cohort (**Methods**). The *y*-axis gives -$\log_{10}$ *p*-values recomputed for this study using a mega-analysis framework. Only SNPs corresponding to lead SNPs published in Supplementary Table 2 of Ripke et al. are shown. Horizontal and vertical gray lines denote the genome-wide significance threshold of $p = 5 \times 10^{-8}$. Note, *p*-values from published meta-analysis results sometimes fail to reach genome-wide significance. This is because significant SNPs in Ripke et al. were determined based on analysis of both a discovery and replication cohort, but published *p*-values are only based on the discovery cohort. Further, our study contains a subset of samples from the original PGC dataset since some cohorts were not made available for analysis.

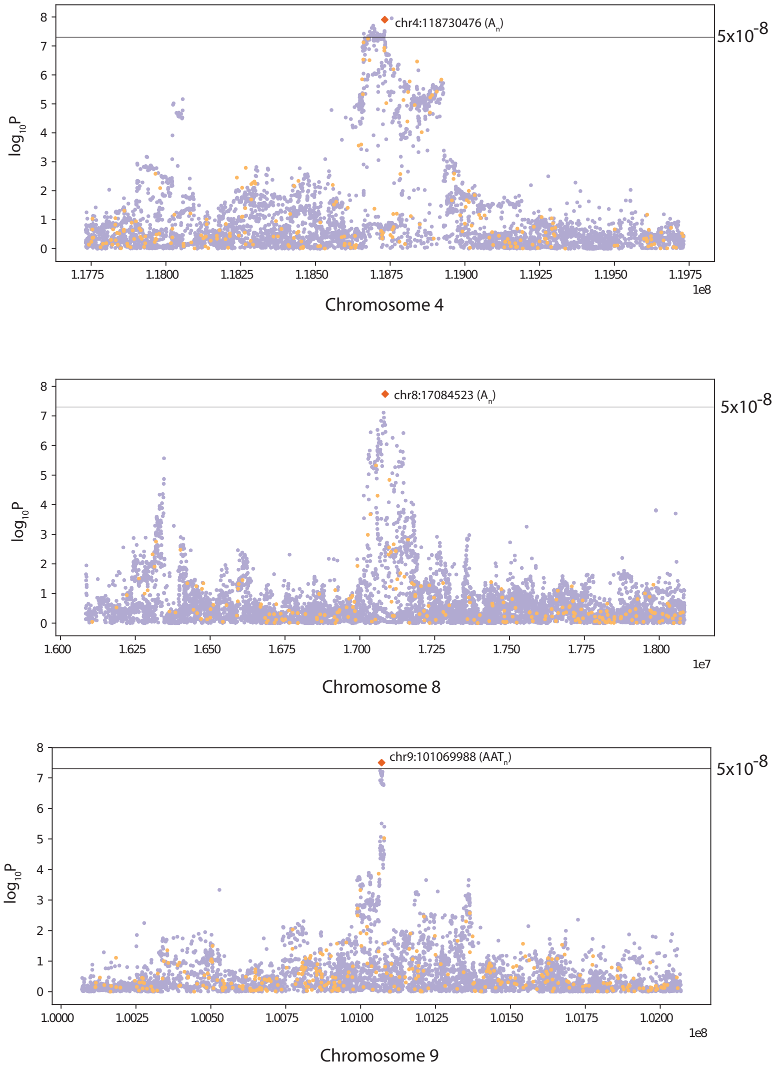


Figure S3.2: **Association signals at genome-wide significant STR signals not identified by SNPs.** Zoomed in Manhattan plots show association *p*-values for STRs (magenta) and SNPs (dark blue) in each region for STR signals not identified by SNP-based GWAS (**Supplementary table 2**). The red horizontal line indicates the genome-wide significance threshold of $p = 5 \times 10^{-8}$.

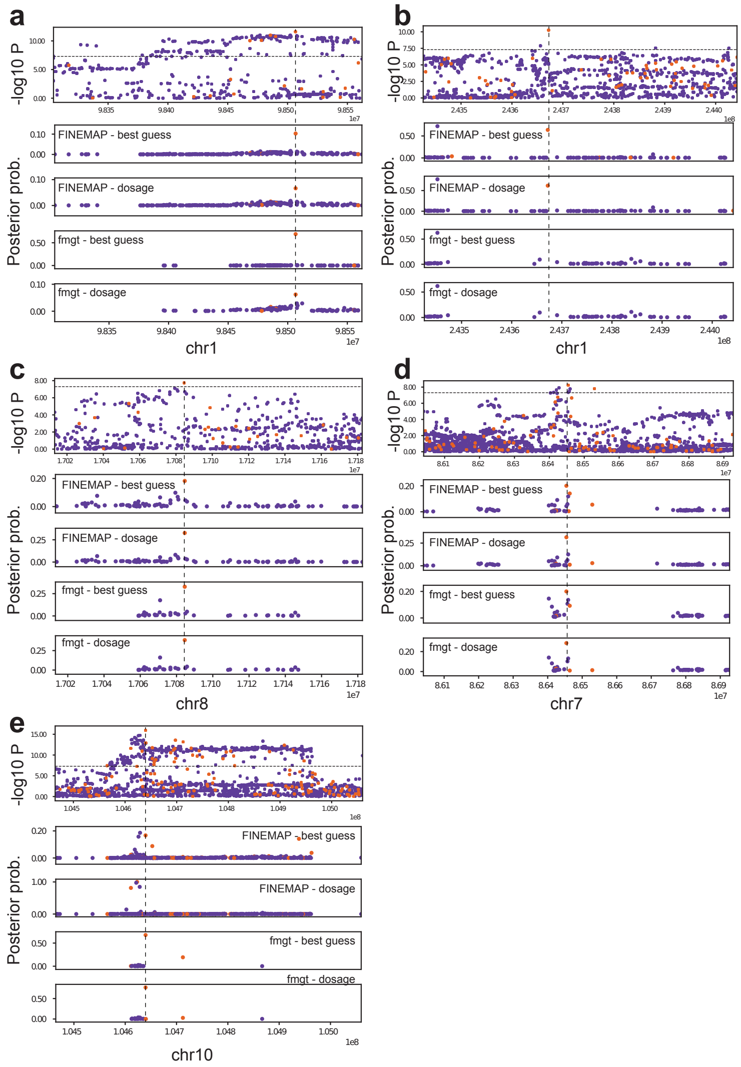


Figure S3.3: **Fine-mapping results of loci with a putative causal STR. a-e** highlight the five STRs shown in **Table 1**. For each panel, the top plot shows the zoomed in Manhattan plot. The *x*-axis gives chromosome position. The *y*-axis gives the -$\log_{10}$ association *p*-values. Purple=SNPs; orange=STRs. The bottom four plots show the posterior probability of causality computed baked on four different fine-mapping settings (**Methods**).

Table S3.1: **Summary of PGC cohorts included in this study.**

Cohort ID	Number of Controls	Number of Cases
scz_aber_eur	699	720
scz_ajsz_eur	1,595	896
scz_asrb_eur	310	509
scz_boco_eur	2,170	1847
scz_buls_eur	608	195
scz_cati_eur	392	409
scz_caws_eur	306	424
scz_cims_eur	69	71
scz_clm2_eur	4,297	3466
scz_clo3_eur	2,083	2150
scz_cou3_eur	693	540
scz_denm_eur	458	492
scz_dubl_eur	860	272
scz_edin_eur	284	368
scz_egcu_eur	1,177	239
scz_ersw_eur	332	322
scz_gras_eur	1,232	1086
scz_irwt_eur	1,022	1309
scz_lacw_eur	466	157
scz_lie2_eur	269	137
scz_lie5_eur	389	509
scz_msaf_eur	139	327
scz_munc_eur	351	437

scz_pewb_eur	1,892	641
scz_pews_eur	236	150
scz_port_eur	216	346
scz_s234_eur	2,341	2077
scz_swe1_eur	214	221
scz_swe5_eur	2,617	1801
scz_swe6_eur	1,219	1094
scz_top8_eur	403	377
scz_ucla_eur	637	705
scz_uclo_eur	494	521
scz_umeb_eur	584	375
scz_umes_eur	713	197
scz_zhh1_eur	190	191
Total	**31957**	**25578**

Table S3.2: **Genome-wide significant LD-independent STR signals**

Chromosome	Position (hg19)	Annotation	P-value (mega-analysis)	Odds ratio	Odds ratio (95% CI)	P-value (meta-analysis)	Index SNP
1	8443073	RERE (intron)	1.77×10^{-9}	1.083	1.056-1.112	1.39×10^{-9}	chr1-8424984-D
1	98506615	MIR137HG (intron)	2.49×10^{-12}	1.01	1.007-1.012	1.07×10^{-12}	rs1702294
1	243671958	AKT3 (intron)	5.60×10^{-11}	0.99	0.987-0.993	9.54×10^{-10}	rs77149735
2	58308993	VRK2 (intron)	6.26×10^{-10}	0.985	0.980-0.989	6.13×10^{-09}	rs11682175
2	200818832	TYW5 (intron)	1.15×10^{-14}	1.129	1.095-1.164	5.62×10^{-14}	chr2-200825237-I
2	233615442	GIGYF2 (intron)	1.89×10^{-9}	0.978	0.971-0.985	8.31×10^{-10}	rs6704768
3	2559900	CNTN4 (intron)	5.76×10^{-11}	1.027	1.019-1.035	9.63×10^{-11}	rs17194490
3	53015546	SFMBT1 (intron)	2.31×10^{-8}	1.023	1.015-1.031	4.22×10^{-08}	rs2535627

3	63849614	ATXN7 (5’UTR) / THOC7 (promoter)	2.85×10^{-9}	0.996	0.994-0.997	9.09×10^{-10}	rs832187
3	135987308	PCCB (intron)	3.61×10^{-8}	0.94	0.919-0.961	6.80×10^{-09}	rs7432375
4	**118731929**	**intergenic**	1.24×10^{-8}	**0.929**	**0.906-0.953**	4.48×10^{-08}	**NA**
5	60609636	intergenic	3.43×10^{-10}	0.953	0.939-0.967	2.40×10^{-09}	rs4391122
6	28176327	intergenic	3.87×10^{-26}	0.761	0.724-0.801	2.07×10^{-22}	rs115329265
6	**31719411**	**MSH5 (intron)**	5.59×10^{-23}	**0.849**	**0.822-0.877**	5.75×10^{-16}	**NA**
7	1989944	MAD1L1 (intron)	3.10×10^{-8}	0.981	0.974-0.987	4.05×10^{-08}	chr7-2025096-I
7	86454144	GRM3 (intron)	5.88×10^{-9}	0.974	0.966-0.983	1.88×10^{-09}	rs12704290
7	104978523	SRPK2 (intron)	1.33×10^{-8}	0.932	0.909-0.955	1.61×10^{-07}	rs6466055
7	110936042	IMMP2L (intron)	8.06×10^{-11}	1.022	1.015-1.029	9.50×10^{-11}	rs13240464
8	**17084523**	**CNOT7 (3’UTR)**	1.82×10^{-8}	**0.924**	**0.898-0.949**	1.22×10^{-08}	**NA**

9	**101069988**	**GABBR2 (intron)**	3.17×10^{-8}	**0.968**	**0.957-0.979**	1.52×10^{-07}	**NA**
10	104639652	AS3MT (intron)	1.13×10^{-16}	1.041	1.031-1.051	1.00×10^{-15}	rs11191419
11	46372406	DGKZ (intron)	4.62×10^{-11}	1.11	1.076-1.146	7.14×10^{-10}	chr11-46350213-D
11	113370024	DRD2 (upstream)	8.52×10^{-12}	0.992	0.990-0.995	2.53×10^{-12}	rs2514218
11	124613794	NRGN (intron)	4.70×10^{-10}	0.993	0.991-0.995	1.57×10^{-09}	rs55661361
11	130719589	LINC02551 (intron)	4.50×10^{-14}	1.03	1.022-1.037	8.32×10^{-13}	rs10791097
12	2358935	CACNA1C (intron)	1.91×10^{-11}	1.085	1.060-1.112	4.27×10^{-11}	rs2007044
12	123662401	MPHOSPH9 (intron)	5.31×10^{-10}	0.921	0.897-0.945	3.28×10^{-09}	rs2851447
14	71565224	PCNX (intron)	1.18×10^{-8}	1.008	1.005-1.011	1.84×10^{-09}	rs2332700
14	104033707	KLC1 / APOPT1 (intron)	1.67×10^{-9}	0.937	0.917-0.957	5.34×10^{-10}	rs12887734

15	61879178	intergenic	4.07×10^{-8}	0.931	0.907-0.955	4.33×10^{-07}	rs12903146
15	78866721	CHRNA5 (intron)	1.84×10^{-10}	1.054	1.037-1.072	2.08×10^{-10}	rs8042374
15	85303360	ZNF592 (intron)	6.32×10^{-9}	0.922	0.897-0.948	4.32×10^{-09}	rs950169
16	29964957	TMEM219 / BOLA2 (intron)	1.17×10^{-9}	0.927	0.905-0.950	1.79×10^{-09}	rs12691307
16	58553551	SETD6 (3'UTR)	4.59×10^{-8}	0.964	0.951-0.977	1.19×10^{-07}	rs12325245
18	53226273	TCF4 (intron)	4.20×10^{-8}	0.929	0.904-0.954	1.44×10^{-07}	rs9636107
19	19519822	GATAD2A (intron)	3.08×10^{-8}	1.027	1.018-1.037	1.60×10^{-07}	rs2905426

Table S3.3: **Fine-mapping STRs and SNPs at schizophrenia-associated loci**

Locus	Finemapping method	Top Hit	Top Hit - Posterior	Top STR	Top STR - Posterior	K (# independent signals detected)
chr1-8424984	FINEMAP:dosages	chr1-8424984-D	27.04%	STR-4382	4.42%	1
chr1-8424984	FINEMAP:best-guess	chr1-8424984-D	26.05%	STR-4382	3.54%	1
chr1-8424984	fmgt:dosages	rs301797	11.90%	STR-4382	5.30%	1
chr1-8424984	fmgt:bestguess	rs301797	11.70%	STR-4386	3.20%	1
chr1-73768366	FINEMAP:dosages	rs35998080	4.46%	STR-47952	2.68%	1
chr1-73768366	FINEMAP:best-guess	STR-47952	4.52%	STR-47952	4.52%	1
chr1-73768366	fmgt:dosages	STR-47880	2.50%	STR-47880	2.50%	1
chr1-73768366	fmgt:bestguess	STR-47865	2.20%	STR-47865	2.20%	1
chr1-98501984	FINEMAP:dosages	STR-60458	6.64%	STR-60458	6.64%	1

chr1-98501984	FINEMAP:best-guess	STR-60458	10.29%	STR-60458	10.29%	1
chr1-98501984	fmgt:dosages	STR-60458	6.20%	STR-60458	6.20%	1
chr1-98501984	fmgt:bestguess	STR-60458	69.00%	STR-60458	69.00%	1
chr1-150031490	FINEMAP:dosages	rs72694943	2.39%	STR-72618	0.21%	1
chr1-150031490	FINEMAP:best-guess	rs72694943	2.40%	STR-72575	0.48%	1
chr1-150031490	fmgt:dosages	rs55802315	3.30%	0.00%	1	
chr1-150031490	fmgt:bestguess	rs55802315	3.30%	STR-72575	0.40%	1
chr1-243555105	FINEMAP:dosages	rs12748870	77.30%	STR-126189	61.88%	2
chr1-243555105	FINEMAP:best-guess	rs12748870	72.23%	STR-126189	63.77%	2
chr1-243555105	fmgt:dosages	rs12748870	61.80%	STR-126095	3.00%	2
chr1-243555105	fmgt:bestguess	rs12748870	62.50%	STR-126241	3.00%	2
chr2-57987593	FINEMAP:dosages	rs11682175	28.60%	STR-735369	6.59%	2

chr2-57987593	FINEMAP:bestguess	rs11682175	23.14%	STR-735369	19.06%	2
chr2-57987593	fmgt:dosages	rs11682175	36.60%	STR-735364	0.50%	1
chr2-57987593	fmgt:bestguess	rs11682175	36.60%	STR-735364	0.30%	1
chr2-146436222	FINEMAP:dosages	rs2381759	8.30%	STR-780716	2.74%	1
chr2-146436222	FINEMAP:bestguess	rs2381759	8.30%	STR-780719	5.29%	1
chr2-146436222	fmgt:dosages	rs2890780	9.30%	STR-780719	1.00%	1
chr2-146436222	fmgt:bestguess	rs2890780	9.00%	STR-780719	2.60%	1
chr2-198304577	FINEMAP:dosages	rs35157131	10.99%	STR-806656	0.89%	1
chr2-198304577	FINEMAP:bestguess	rs35157131	11.32%	STR-806752	1.99%	1
chr2-198304577	fmgt:dosages	rs788023	3.10%	STR-806656	1.10%	1
chr2-198304577	fmgt:bestguess	rs2565160	3.00%	STR-806593	1.90%	1
chr2-200164252	FINEMAP:dosages	rs35733345	83.60%	STR-807558	0.33%	3

chr2-200164252	FINEMAP:best-guess	rs35733345	83.51%	STR-807515	0.80%	3
chr2-200164252	fmgt:dosages	rs4673480	50.90%	0.00%	2	
chr2-200164252	fmgt:bestguess	rs4673480	50.90%	0.00%	2	
chr2-200825237	FINEMAP:dosages	rs76432012	83.61%	STR-807927	0.83%	2
chr2-200825237	FINEMAP:best-guess	rs76432012	81.63%	STR-807866	1.41%	2
chr2-200825237	fmgt:dosages	rs11693528	20.00%	STR-807927	1.40%	1
chr2-200825237	fmgt:bestguess	rs116393510	85.40%	STR-807927	1.30%	2
chr2-225391296	FINEMAP:dosages	rs4674918	58.41%	STR-822313	3.71%	2
chr2-225391296	FINEMAP:best-guess	rs4674918	58.78%	STR-822283	3.83%	2
chr2-225391296	fmgt:dosages	STR-822313	7.20%	STR-822313	7.20%	1
chr2-225391296	fmgt:bestguess	rs11686590	4.80%	STR-822313	3.60%	1
chr2-233592501	FINEMAP:dosages	STR-827401	18.74%	STR-827401	18.74%	1

chr2-233592501	FINEMAP:best-guess	rs1108252	14.27%	STR-827373	5.18%	1
chr2-233592501	fmgt:dosages	rs6704768	4.60%	0.00%	1	
chr2-233592501	fmgt:bestguess	rs6704768	4.60%	0.00%	1	
chr3-2547786	FINEMAP:dosages	rs11925117	92.22%	STR-913024	22.42%	2
chr3-2547786	FINEMAP:best-guess	rs11925117	92.16%	STR-913024	19.84%	2
chr3-2547786	fmgt:dosages	rs17194490	24.90%	STR-913024	3.20%	1
chr3-2547786	fmgt:bestguess	rs17194490	25.20%	STR-913024	2.00%	1
chr3-36858583	FINEMAP:dosages	rs75968099	20.52%	STR-930771	0.85%	1
chr3-36858583	FINEMAP:best-guess	rs75968099	20.02%	STR-930771	1.08%	1
chr3-36858583	fmgt:dosages	rs9876421	18.50%	STR-930771	0.70%	1
chr3-36858583	fmgt:bestguess	rs9876421	18.50%	STR-930771	0.70%	1
chr3-52845105	FINEMAP:dosages	rs2535629	5.33%	STR-941180	3.36%	2

chr3-52845105	FINEMAP:best-guess	rs2535629	5.23%	STR-941180	3.24%	2
chr3-52845105	fmgt:dosages	rs2710339	5.80%	STR-941159	1.80%	1
chr3-52845105	fmgt:bestguess	STR-941070	3.20%	STR-941070	3.20%	1
chr3-63833050	FINEMAP:dosages	rs832187	16.01%	STR-947312	14.73%	1
chr3-63833050	FINEMAP:best-guess	rs832187	16.10%	STR-947357	12.01%	1
chr3-63833050	fmgt:dosages	STR-947312	28.10%	STR-947312	28.10%	1
chr3-63833050	fmgt:bestguess	rs832190	21.40%	STR-947312	19.00%	1
chr3-136288405	FINEMAP:dosages	rs12488721	16.08%	STR-981628	14.94%	1
chr3-136288405	FINEMAP:best-guess	rs12488721	19.64%	STR-981845	2.40%	1
chr3-136288405	fmgt:dosages	STR-981628	12.90%	STR-981628	12.90%	1
chr3-136288405	fmgt:bestguess	STR-981628	18.30%	STR-981628	18.30%	1
chr3-180594593	FINEMAP:dosages	rs13096210	18.86%	STR-1004849	3.22%	2

chr3-180594593	FINEMAP:best-guess	STR-1005316	93.59%	STR-1005316	93.59%	3
chr3-180594593	fmgt:dosages	rs13096210	18.90%	STR-1004849	2.60%	2
chr3-180594593	fmgt:bestguess	rs13096210	18.70%	STR-1004860	3.90%	2
chr4-118731929	FINEMAP:dosages	rs139199583	11.90%	STR-1075390	2.25%	1
chr4-118731929	FINEMAP:best-guess	rs139199583	11.92%	STR-1075442	0.77%	1
chr4-118731929	fmgt:dosages	rs4446400	1.60%	STR-1075388	0.30%	1
chr4-118731929	fmgt:bestguess	rs4446400	1.60%	STR-1075388	0.50%	1
chr4-170626552	FINEMAP:dosages	rs72696665	12.96%	STR-1101256	0.18%	1
chr4-170626552	FINEMAP:best-guess	rs72696665	12.90%	STR-1101233	0.94%	1
chr4-170626552	fmgt:dosages	rs1566522	6.50%	STR-1101285	0.00%	1
chr4-170626552	fmgt:bestguess	rs1566522	6.40%	STR-1101416	0.10%	1
chr5-60598543	FINEMAP:dosages	STR-1140999	11.80%	STR-1140999	11.80%	1

chr5-60598543	FINEMAP:best-guess	STR-1140941	86.53%	STR-1140941	86.53%	2
chr5-60598543	fmgt:dosages	STR-1140946	11.20%	STR-1140946	11.20%	1
chr5-60598543	fmgt:bestguess	rs4132385	7.50%	STR-1140946	5.40%	1
chr5-152177121	FINEMAP:dosages	rs111294930	26.87%	STR-1188614	0.84%	2
chr5-152177121	FINEMAP:best-guess	rs111294930	25.96%	STR-1188654	4.76%	2
chr5-152177121	fmgt:dosages	rs2910032	19.30%	STR-1188611	0.90%	2
chr5-152177121	fmgt:bestguess	rs2910032	19.30%	STR-1188611	1.20%	2
chr5-152608619	FINEMAP:dosages	rs111294930	40.67%	STR-1188614	0.64%	2
chr5-152608619	fmgt:dosages	rs12522290	50.70%	STR-1188611	0.90%	3
chr5-152608619	fmgt:bestguess	rs12522290	50.80%	STR-1188611	1.10%	3
chr5-153680747	FINEMAP:dosages	chr5-154139507-D	66.11%	STR-1189243	35.31%	2
chr5-153680747	fmgt:dosages	rs153431	3.90%	0.00%	1	

chr5-153680747	fmgt:bestguess	rs153431	3.90%	0.00%	1	
chr6-28712247	FINEMAP:dosages	rs116591906	100.00%	STR-1222821	100.00%	5
chr6-28712247	FINEMAP:bestguess	chr6-28523687-D	100.00%	STR-1222586	100.00%	5
chr6-28712247	fmgt:dosages	STR-1222540	99.50%	STR-1222540	99.50%	2
chr6-28712247	fmgt:bestguess	rs13217619	15.60%	STR-1222643	13.70%	2
chr6-84280274	FINEMAP:dosages	STR-1250745	100.00%	STR-1250745	100.00%	5
chr6-84280274	FINEMAP:bestguess	STR-1250745	100.00%	STR-1250745	100.00%	5
chr7-2025096	FINEMAP:dosages	rs4719432	16.22%	STR-1296596	0.27%	1
chr7-2025096	FINEMAP:bestguess	rs4719432	16.35%	STR-1296596	0.18%	1
chr7-2025096	fmgt:dosages	rs4719432	37.90%	STR-1296596	0.20%	1
chr7-2025096	fmgt:bestguess	rs4719432	38.00%	STR-1296596	0.10%	1
chr7-86427626	FINEMAP:dosages	STR-1344557	31.41%	STR-1344557	31.41%	2

chr7-86427626	FINEMAP:bestguess	STR-1344557	19.99%	STR-1344557	19.99%	2
chr7-86427626	fmgt:dosages	STR-1344557	28.50%	STR-1344557	28.50%	2
chr7-86427626	fmgt:bestguess	STR-1344557	20.00%	STR-1344557	20.00%	2
chr7-104929064	FINEMAP:dosages	rs10953479	98.43%	STR-1355970	9.14%	2
chr7-104929064	FINEMAP:bestguess	rs10953479	98.36%	STR-1355947	3.18%	2
chr7-104929064	fmgt:dosages	rs10953479	100.00%	STR-1355970	10.80%	2
chr7-104929064	fmgt:bestguess	rs10953479	100.00%	STR-1355947	5.20%	2
chr7-110898915	FINEMAP:dosages	rs214475	100.00%	STR-1359239	0.00%	5
chr7-110898915	FINEMAP:bestguess	rs214475	100.00%	STR-1359239	0.00%	5
chr7-110898915	fmgt:dosages	STR-1359256	100.00%	STR-1359256	100.00%	2
chr7-110898915	fmgt:bestguess	rs13240464	15.70%	STR-1359120	3.70%	1
chr8-17084523	FINEMAP:dosages	STR-1394258	32.68%	STR-1394258	32.68%	1

chr8-17084523	FINEMAP:best-guess	STR-1394258	18.26%	STR-1394258	18.26%	1
chr8-17084523	fmgt:dosages	STR-1394258	38.50%	STR-1394258	38.50%	1
chr8-17084523	fmgt:bestguess	STR-1394258	32.80%	STR-1394258	32.80%	1
chr8-60700469	FINEMAP:dosages	rs6986251	29.01%	STR-1417379	3.88%	1
chr8-60700469	FINEMAP:best-guess	STR-1417390	100.00%	STR-1417390	100.00%	3
chr8-60700469	fmgt:dosages	rs6986251	29.00%	STR-1417379	3.60%	1
chr8-60700469	fmgt:bestguess	rs6986251	30.00%	STR-1417379	1.50%	1
chr8-111485761	FINEMAP:dosages	rs16880943	7.55%	STR-1443799	1.11%	1
chr8-111485761	FINEMAP:best-guess	rs16880943	9.09%	STR-1443892	8.92%	1
chr8-111485761	fmgt:dosages	rs34137090	4.20%	STR-1443844	0.60%	1
chr8-111485761	fmgt:bestguess	rs34137090	4.20%	STR-1443840	0.60%	1
chr9-101069988	FINEMAP:dosages	chr9-101360865-I	17.46%	STR-1501905	11.39%	2

chr9-101069988	FINEMAP:best-guess	chr9-101360865-I	17.33%	STR-1501905	10.29%	2
chr9-101069988	fmgt:dosages	STR-1501905	12.90%	STR-1501905	12.90%	1
chr9-101069988	fmgt:bestguess	STR-1501905	10.10%	STR-1501905	10.10%	1
chr10-18745105	FINEMAP:dosages	chr10-18737405-I	14.40%	STR-140824	0.89%	1
chr10-18745105	FINEMAP:best-guess	STR-140811	99.90%	STR-140811	99.90%	3
chr10-18745105	fmgt:dosages	rs12784686	22.30%	STR-140824	16.60%	1
chr10-18745105	fmgt:bestguess	rs12784686	24.00%	STR-140824	10.40%	1
chr10-104957618	FINEMAP:dosages	STR-187791	100.00%	STR-187791	100.00%	5
chr10-104957618	FINEMAP:best-guess	rs7085104	18.62%	STR-187806	16.71%	2
chr10-104957618	fmgt:dosages	STR-187806	77.10%	STR-187806	77.10%	1
chr10-104957618	fmgt:bestguess	STR-187806	67.40%	STR-187806	67.40%	1
chr11-24403620	FINEMAP:dosages	rs1579116	49.82%	STR-216044	2.73%	2

chr11-24403620	FINEMAP:best-guess	rs1579116	49.94%	STR-216044	2.49%	2
chr11-24403620	fmgt:dosages	rs12418983	23.70%	STR-216044	3.20%	2
chr11-24403620	fmgt:bestguess	rs12418983	23.70%	STR-216044	2.80%	2
chr11-46350213	FINEMAP:dosages	STR-227380	20.96%	STR-227380	20.96%	1
chr11-46350213	FINEMAP:best-guess	chr11-46350213-D	7.83%	STR-227386	4.09%	1
chr11-46350213	fmgt:dosages	STR-227380	18.50%	STR-227380	18.50%	1
chr11-46350213	fmgt:bestguess	rs7951870	6.70%	STR-227386	5.20%	1
chr11-57510294	FINEMAP:dosages	rs112614215	56.36%	STR-231316	1.40%	2
chr11-57510294	FINEMAP:best-guess	rs112614215	56.11%	STR-231404	1.11%	2
chr11-57510294	fmgt:dosages	rs112614215	97.20%	STR-231685	2.80%	2
chr11-57510294	fmgt:bestguess	rs112614215	97.70%	STR-231685	2.30%	2
chr11-109378071	FINEMAP:dosages	rs10789735	23.69%	STR-260485	1.67%	2

chr11-109378071	FINEMAP:bestguess	STR-260514	43.79%	STR-260514	43.79%	2
chr11-109378071	fmgt:dosages	rs11213112	34.40%	0.00%	1	
chr11-109378071	fmgt:bestguess	rs11213112	34.40%	0.00%	1	
chr11-113392994	FINEMAP:dosages	STR-262621	80.29%	STR-262621	80.29%	3
chr11-113392994	FINEMAP:bestguess	rs4987094	59.76%	STR-262602	0.22%	3
chr11-113392994	fmgt:dosages	rs12288145	51.80%	0.00%	3	
chr11-113392994	fmgt:bestguess	rs12288145	51.80%	0.00%	3	
chr11-124613957	FINEMAP:dosages	rs10128573	55.32%	STR-269231	39.76%	3
chr11-124613957	FINEMAP:bestguess	rs10128573	56.13%	STR-269231	11.68%	3
chr11-124613957	fmgt:dosages	rs10128573	100.00%	STR-269231	48.90%	3
chr11-124613957	fmgt:bestguess	rs10128573	100.00%	STR-269231	32.30%	3
chr11-130718630	FINEMAP:dosages	rs10894287	100.00%	STR-272427	99.34%	5

chr11-130718630	FINEMAP:bestguess	rs2324317	100.00%	STR-272411	100.00%	5
chr11-130718630	fmgt:dosages	rs4601795	29.50%	STR-272364	21.60%	2
chr11-130718630	fmgt:bestguess	rs4601795	37.40%	STR-272383	1.70%	2
chr11-133822569	FINEMAP:dosages	rs1561613	67.00%	STR-273684	0.11%	5
chr12-2344960	FINEMAP:dosages	rs1024582	19.12%	STR-275767	0.93%	1
chr12-2344960	FINEMAP:bestguess	rs1024582	18.78%	STR-275767	2.62%	1
chr12-2344960	fmgt:dosages	STR-275586	67.50%	STR-275586	67.50%	2
chr12-2344960	fmgt:bestguess	STR-275592	60.10%	STR-275592	60.10%	2
chr12-29917265	FINEMAP:bestguess	rs1874797	24.53%	STR-291274	9.12%	2
chr12-29917265	fmgt:dosages	rs1874797	25.40%	STR-291274	8.30%	1
chr12-29917265	fmgt:bestguess	rs1874797	25.10%	STR-291274	9.50%	1
chr12-123665113	FINEMAP:dosages	chr12-123742918-D	41.08%	STR-346691	1.55%	1

chr12-123665113	FINEMAP:best-guess	chr12-123618360-D	100.00%	STR-346707	100.00%	5
chr12-123665113	fmgt:dosages	rs1727319	12.80%	STR-346756	1.80%	1
chr12-123665113	fmgt:bestguess	rs1727319	11.70%	STR-346726	5.20%	1
chr14-30190316	FINEMAP:best-guess	chr14-29725144-I	89.36%	STR-408005	30.14%	4
chr14-30190316	fmgt:dosages	rs2068012	65.30%	STR-408005	1.00%	1
chr14-30190316	fmgt:bestguess	rs2068012	64.40%	STR-408005	2.50%	1
chr14-104046834	FINEMAP:dosages	rs4906364	28.90%	STR-450965	3.58%	1
chr14-104046834	FINEMAP:best-guess	rs4906364	31.38%	STR-451168	0.84%	1
chr14-104046834	fmgt:dosages	STR-450745	100.00%	STR-450745	100.00%	2
chr14-104046834	fmgt:bestguess	STR-450745	100.00%	STR-450745	100.00%	2
chr15-61854663	FINEMAP:dosages	rs2414718	18.83%	STR-476493	0.49%	1
chr15-61854663	FINEMAP:best-guess	rs2414718	17.53%	STR-476480	7.80%	1

chr15-61854663	fmgt:dosages	rs2414718	20.90%	0.00%	1	
chr15-61854663	fmgt:bestguess	rs2414718	20.90%	0.00%	1	
chr15-78859610	FINEMAP:dosages	STR-487459	18.95%	STR-487459	18.95%	2
chr15-78859610	FINEMAP:bestguess	rs147144681	10.08%	STR-487495	5.25%	2
chr15-78859610	fmgt:dosages	rs57064725	5.40%	STR-487517	2.00%	1
chr15-78859610	fmgt:bestguess	rs57064725	5.20%	STR-487517	4.30%	1
chr15-84706461	FINEMAP:bestguess	rs2002375	14.48%	STR-490798	0.39%	1
chr15-84706461	fmgt:dosages	rs11638445	4.40%	0.00%	1	
chr15-84706461	fmgt:bestguess	rs11638445	4.40%	0.00%	1	
chr15-91426560	FINEMAP:dosages	rs4702	91.19%	STR-494789	0.31%	2
chr15-91426560	FINEMAP:bestguess	rs4702	91.45%	STR-494789	0.24%	2
chr15-91426560	fmgt:dosages	rs4702	84.00%	0.00%	1	

chr15-91426560	fmgt:bestguess	rs4702	84.00%	0.00%	1	
chr16-29939877	FINEMAP:dosages	rs12691307	15.28%	STR-524508	0.62%	1
chr16-29939877	FINEMAP:bestguess	rs12691307	14.08%	STR-524484	7.29%	1
chr16-29939877	fmgt:dosages	rs4402589	21.30%	STR-524508	0.30%	1
chr16-29939877	fmgt:bestguess	rs4402589	21.30%	STR-524508	0.20%	1
chr16-58681393	FINEMAP:dosages	rs11647976	15.03%	STR-535191	0.65%	2
chr16-58681393	FINEMAP:bestguess	rs12325245	15.56%	STR-535323	0.59%	2
chr16-58681393	fmgt:dosages	rs12325003	9.80%	STR-535314	5.40%	2
chr16-58681393	fmgt:bestguess	rs12325003	10.10%	STR-535323	3.30%	2
chr16-68189340	FINEMAP:dosages	rs7193701	5.12%	STR-541443	0.61%	1
chr16-68189340	FINEMAP:bestguess	rs7193701	4.76%	STR-541292	4.46%	1
chr16-68189340	fmgt:dosages	rs10852439	7.00%	STR-541443	0.40%	1

chr16-68189340	fmgt:bestguess	rs10852439	6.70%	STR-541382	3.60%	1
chr17-2208899	FINEMAP:dosages	rs4523957	4.45%	STR-557975	1.09%	1
chr17-2208899	FINEMAP:best-guess	rs4523957	4.32%	STR-557984	1.58%	1
chr17-2208899	fmgt:dosages	rs216189	3.30%	STR-557943	1.10%	1
chr17-2208899	fmgt:bestguess	rs216189	3.30%	STR-557943	0.90%	1
chr18-52749216	FINEMAP:dosages	rs9636107	53.73%	STR-640744	1.86%	4
chr18-52749216	FINEMAP:best-guess	rs9636107	51.69%	STR-640744	6.50%	4
chr18-52749216	fmgt:dosages	rs9636107	97.80%	0.00%	4	
chr18-52749216	fmgt:bestguess	rs9636107	97.80%	0.00%	4	
chr18-53063676	FINEMAP:dosages	rs144158419	23.65%	STR-640744	2.14%	4
chr18-53063676	FINEMAP:best-guess	rs144158419	24.46%	STR-640744	3.92%	4
chr18-53063676	fmgt:dosages	rs9636107	69.60%	STR-640873	4.20%	4

chr18-53063676	fmgt:bestguess	rs9636107	69.00%	STR-640873	5.00%	4
chr18-53200117	FINEMAP:dosages	rs9636107	21.09%	STR-640744	1.73%	3
chr18-53200117	FINEMAP:best-guess	rs9636107	19.64%	STR-640744	3.83%	3
chr18-53200117	fmgt:dosages	rs9636107	69.60%	STR-640873	4.20%	4
chr18-53200117	fmgt:bestguess	rs9636107	69.00%	STR-640873	5.00%	4
chr18-53533189	FINEMAP:dosages	rs1792695	30.29%	STR-640873	2.87%	4
chr18-53533189	FINEMAP:best-guess	rs1792695	29.60%	STR-640744	4.39%	4
chr18-53533189	fmgt:dosages	rs9636107	46.10%	STR-640873	7.80%	4
chr18-53533189	fmgt:bestguess	rs9636107	45.30%	STR-640873	9.30%	4
chr18-53795514	FINEMAP:dosages	rs77882218	49.12%	STR-640873	4.34%	4
chr18-53795514	FINEMAP:best-guess	rs77882218	48.45%	STR-640873	5.52%	4
chr18-53795514	fmgt:dosages	rs77882218	100.00%	0.00%	3	

chr18-53795514	fmgt:bestguess	rs77882218	100.00%	0.00%	3	
chr19-19478022	FINEMAP:dosages	rs4808931	4.99%	STR-674033	3.20%	1
chr19-19478022	FINEMAP:bestguess	rs4808931	4.81%	STR-674033	2.17%	1
chr19-19478022	fmgt:dosages	rs7253952	5.90%	STR-674111	0.60%	1
chr19-19478022	fmgt:bestguess	rs7253952	5.90%	STR-673901	0.60%	1
chr20-37453194	FINEMAP:dosages	rs208818	18.34%	STR-853031	5.26%	2
chr20-37453194	FINEMAP:bestguess	rs208818	20.19%	STR-852980	1.04%	2
chr20-37453194	fmgt:dosages	STR-852893	100.00%	STR-852893	100.00%	2
chr20-37453194	fmgt:bestguess	STR-852893	100.00%	STR-852893	100.00%	2
chr22-39987017	FINEMAP:dosages	rs732381	15.03%	STR-904116	2.73%	2
chr22-39987017	FINEMAP:bestguess	rs732381	15.26%	STR-904116	4.52%	2
chr22-39987017	fmgt:dosages	rs732381	17.30%	STR-904139	1.10%	1

chr22-39987017	fmgt:bestguess	rs732381	16.60%	STR-904139	4.90%	1
chr22-41587556	FINEMAP:dosages	rs9607782	40.53%	STR-905118	0.23%	2
chr22-41587556	fmgt:dosages	rs5995910	13.80%	STR-905095	10.10%	2
chr22-41587556	fmgt:bestguess	rs5995910	14.30%	STR-905095	6.70%	2
chr22-42340844	FINEMAP:dosages	STR-906383	24.02%	STR-906383	24.02%	1
chr22-42340844	FINEMAP:bestguess	rs760648	23.07%	STR-906477	4.12%	1
chr22-42340844	fmgt:dosages	STR-906383	29.60%	STR-906383	29.60%	1
chr22-42340844	fmgt:bestguess	rs1023499	7.30%	STR-906383	3.00%	1

Acknowledgments

This study was supported by the Office Of The Director, National Institutes of Health under Award Number DP5OD024577. M.G. was additionally supported in part by NIH/NHGRI grants R01HG010149 and R21HG010070, and NIH/NIMH grant R01 MH113715.

Chapter 3, in full, contains material from Shubham Saini, Brittany S Leger, Jonghun Park, PGC Schizophrenia Working Group, Vineet Bafna, Alon Goren, Melissa Gymrek. ”Genome-wide analysis of the contributions of short tandem repeat variants to schizophrenia risk”, which is currently being prepared for submission for publication of the material. I was the primary investigator and author of this material.