Uncovering the variability, regulatory roles and mutation rates of short tandem repeats

by

Thomas F. Willems

B.S. in Chemical Engineering, University of California, Berkeley (2011)

Submitted to the Computational and Systems Biology Program
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Author .....
Computational and Systems Biology Program
April 26, 2016

Certified by .....
Yaniv Erlich, PhD
Assistant Professor of Computer Science, Columbia University
Thesis Supervisor

Certified by .....
Manolis Kellis, PhD
Associate Professor of Computer Science, MIT
Thesis Supervisor

Accepted by .....
Christopher B. Burge, PhD
Professor of Biology and Biological Engineering
Director, Computational and Systems Biology Graduate Program

Uncovering the variability, regulatory roles and mutation rates of short tandem repeats

by
Thomas F. Willems

Submitted to the Computational and Systems Biology Program
on April 26, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Over the past decade, the advent of next-generation DNA sequencing technologies has ushered in an exciting era of biological research. Through large-scale sequencing projects, scientists have begun to unveil the variability and function of millions of DNA mutations called single nucleotide polymorphisms. Despite this rapid growth in understanding, short tandem repeats (STRs), genomic elements consisting of a repeating pattern of 2-6 bases, have remained poorly understood. Mutating orders of magnitude more rapidly than most of the human genome, STRs have been identified as the causal variants in diseases such as Fragile X syndrome and Huntington's disease. However, in spite of their potentially profound biological consequences, STRs remain systematically understudied due to difficulties associated with obtaining accurate genotypes. To address this issue, we developed a series of bioinformatics approaches and applied them to population-scale whole-genome sequencing data sets. Using data from the 1000 Genomes Project, we performed the first genome-wide characterization of STR variability by analyzing over 700,000 loci in more than 1000 individuals. Next, we integrated these genotypes with expression data to assess the contribution of STRs to gene expression in humans, uncovering their substantial regulatory role. We then developed a state-of-the-art algorithm to genotype STRs, resulting in vastly improved accuracy and uncovering hundreds of replicable de novo mutations in a deeply sequenced trio. Lastly, we developed a novel approach to estimate mutation rates for STRs on the Y-chromosome (Y-STR), resulting in rates for hundreds of previously uncharacterized markers. Collectively, these analyses highlight the extreme variability of STRs and provide a framework for incorporating them into future studies.

Thesis Supervisor: Yaniv Erlich, PhD
Title: Assistant Professor of Computer Science, Columbia University

Thesis Supervisor: Manolis Kellis, PhD
Title: Associate Professor of Computer Science, MIT

Acknowledgments

My journey through graduate school never would have started, let alone finished, if not for an enormous number of people that have supported and inspired me along the way. I sincerely want to thank the following people that have made the past five years a fantastic experience:

- Yaniv Erlich, for believing in me, inspiring me and teaching me how to do great science, and for introducing me to the exciting world of genomics.
- Manolis Kellis, for adopting me into his lab and for his limitless positive energy and sound scientific advice.
- David Housman and Laurie Boyer, for their terrific scientific guidance throughout my PhD.
- My labmates: Melissa Gymrek, for teaching me all she knows about STRs and bioinformatics and for being a great friend and colleague. Dina Zielinski, for lending her wet lab expertise and challenging me to swimming duels. Assaf Gordon, for his never ending wise programming advice and generosity. Sophie Zaaijer, for always welcoming me and entertaining me at the NYGC. And to all the other members of the Erlich lab that have come and gone, thank you for making it a fun and exciting place to work.
- The Whitehead Institute, for providing a fantastic environment for young scientists.
- Area Four coffee, for providing the fuel that has sustained me throughout graduate school.
- Eric Zhu, for being a sage and tireless friend.
- Rosanna Lim and Aanchal Jain, for much needed relaxation during Friday night dinners.
- Lionel Lam, for providing timely musical relief and humor.
- Xiao Su, for teaching me the value of a coffee or donut break.
- Mark, Sean, Steven, William, Karthik and David, for many much needed nights out in Boston.
- Jacquie Carota, for her tireless work as our program administrator.
- My brother Nick, my sister Julie and my brother-in-law Laurens, for their love and support throughout graduate school.
- My girlfriend Su Luo, whose love has carried me through difficult times and has inspired me to aim for new heights.

- My parents Paul and Carine, for molding me into the person I am today and prioritizing my education. Mom and Dad, without your endless advice and support, none of this would ever have been possible.

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

1.1 Overview

Tremendous advances in DNA sequencing technology have revolutionized the field of human genetics in the past 15 years. While the original human genome project required over 10 years and $3 billion to complete, rapidly developing next-generation technologies can now sequence a human genome for less than $1,000 and in only a few days. Empowered by these remarkable breakthroughs, scientists have begun to sequence increasingly large human populations in an effort to understand human genetic variation. The early phases of the International Hap Map Project (2005), 1000 Genomes Project (2013) and UK10K Project (2015) have provided more detailed views of human genetic variation by characterizing the genomes of approximately 270, 1000 and 3700 individuals, respectively [1, 2, 3]. Further underscoring the rapid progression in DNA sequencing applications, Genomics England has recently begun to sequence a cohort of 100,000 individuals in an effort to integrate genomic information into personalized medicine.

In addition to breakthroughs in sequencing technology, developments in statistical genetics have enabled novel insights into the contributions of genetic variants to complex traits. Genome-wide association studies (GWAS) have identified tens of thousands of variants linked to hundreds of unique traits, ranging from Alzheimer's disease to blood pressure and height [4]. In addition, techniques that estimate the proportion of a trait attributable to genetic components have revealed that many complex traits and diseases are highly polygenic, involving hundreds of loci [5, 6].

Despite this progress, most of these advances have focused on single nucleotide polymorphisms (SNPs), mutations involving a change to a single DNA base pair. This limitation largely stems from the fact that SNPs are easier to characterize, both experimentally and computationally. Other variant types, such as copy number variants, insertions, and deletions, involve much larger segments of DNA. As a result, accurately detecting them requires higher coverage datasets,

longer sequencing reads and more complex algorithms that have only become readily available in the last few years.

In this thesis, I describe our efforts to expand our understanding of a class of genetic variant called short tandem repeats (STRs). Comprised of a repeating 2-6 base pair motif, STRs occupy ∼1% of the human genome. Their repetitive sequence causes frequent errors during DNA replication, resulting in mutation rates that are orders of magnitude higher than most other types of variants. These elevated mutation rates result in extremely diverse allelic spectra, making STRs information-rich markers that are valuable for applications in forensics, genealogy and ancestry inference. While STRs are widely used in genetics, their highly repetitive nature has also made them difficult to profile using next-generation sequencing technologies. As a result, recent large-scale efforts to characterize genetic variation have largely ignored STRs. **Chapter 2** describes how we address this limitation using data from the 1000 Genomes Project [2]. We utilize a novel bioinformatics tool to characterize over half a million STR loci in more than 1,000 individuals, substantially expanding our knowledge about the number of polymorphic STRs, the sequence features that govern their mutability and the levels of linkage disequilibrium between SNPs and STRs. In **Chapter 3**, we integrate this rich set of STR genotypes with gene expression data to identify STRs that modulate gene expression (eSTRs). This analysis demonstrates that STRs have a large regulatory role and underscores their putative contribution to more complex traits. **Chapter 4** details our efforts to develop and benchmark HipSTR, a state-of-the-art variant caller for STRs. Using several high coverage sequencing datasets, we demonstrate that HipSTR offers unprecedented accuracy and that we can phase and impute STRs using slight modifications to existing statistical methods. Finally, **Chapter 5** outlines our approach to estimating the mutation rates of STRs on the Y-chromosome (Y-STRs). Through both simulations and external validation, we demonstrate that our approach is highly accurate and we use it to estimate the mutation rates of hundreds of previously uncharacterized Y-STRs. We then leverage these estimates to predict the number of de novo STR mutations genome-wide and demonstrate that it rivals that of any other variant class.

In the sections that follow, I begin by providing a brief overview of the definition and applications of STRs. I then lay the framework for each of the chapters that follow by describing what was previously known about STRs and how our contributions have advanced this knowledge.

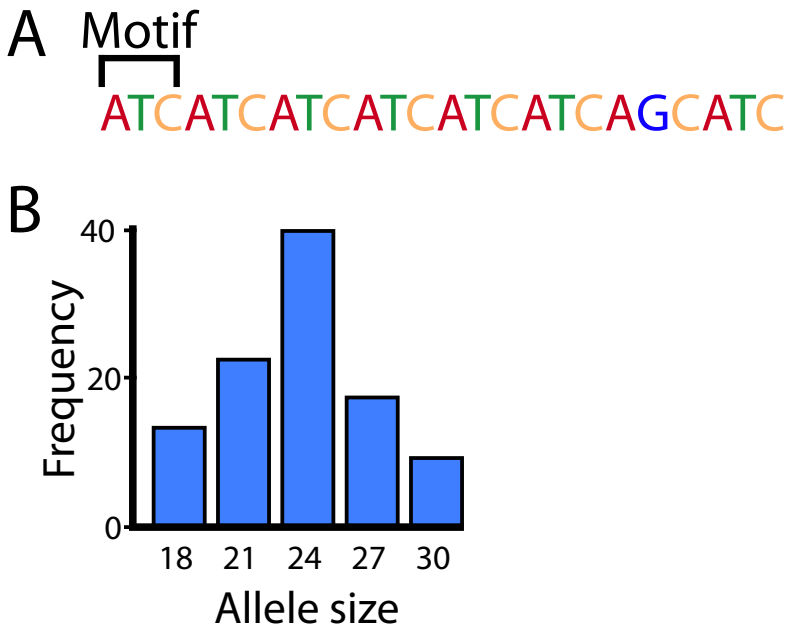

Figure 1-1: **Example of an STR's (a) sequence and (b) allele frequency distribution.**

1.2 Defining an STR

Short tandem repeats are one of the must abundant types of repetitive elements in the human genome. Each STR is made up of a 2-6 base pair (bp) sequence that is repeated end-to-end a particular number of times. The sequence of the repeat unit is called the *motif*, while the length of the repeat unit is known as the *period*. **Figure 1-1a** outlines a simple example of a 24 bp STR whose motif and period are *ATC* and three, respectively. In contrast to SNPs, which mostly only have two alleles, STRs frequently have three or more alleles. In addition, because STR mutations typically add or remove copies of the motif, their alleles usually occur in increments of the period. As a result, STR variations are thought of in terms of the lengths and frequencies of the alleles at each locus. **Figure 1-1b** contains a hypothetical frequency distribution for the example STR described earlier.

Despite the relatively simple definition of an STR, it remains unclear how many repeats distinguish an STR from background DNA. For instance, should an *ATCATC* sequence be classified as an STR, or is it not sufficiently long? This question is further complicated by interruptions

to the repeat structure, such as the *G* interruption in **Figure 1-1a**. Part of **Chapter 2** outlines the quantitative approach we used to address this question. Motivated by the notion that true STRs should be substantially more repetitive than randomly generated sequences, we ultimately identified nearly 700,000 STRs across the human genome.

1.3 STR applications

In the 1980's, scientists were still searching for DNA markers that displayed variation between human individuals. Early work revealed that dinucleotide STR repeats amplified using PCR displayed abundant length polymorphisms [7, 8]. This discovery immediately highlighted STRs as useful genetic markers, because in addition to their diverse allelic spectra, dinucleotide repeats were known to be abundant and distributed across the human genome. Scientists therefore began to use STRs to construct improved linkage maps of the human genome [9], largely supplanting those that had been based on restriction fragment length polymorphisms [10, 11]. These maps provided both the ordering and distance between markers on each chromosome, enabling studies to pinpoint the genes and mutations involved in a wide range of heritable diseases.

STRs have also been leveraged to obtain insights into human population structure. By counting the fraction of alleles shared between pairs of individuals at 30 microsatellites, Bowcock et al. demonstrated that samples from distinct populations clustered together by geographic origin [12]. These findings ultimately led to the development of a wide range of STR-specific distance measures [13, 14], which in turn have uncovered the relatedness between human populations and their divergence times.

The high mutability of STRs has led to widespread applications in DNA identification. Preliminary studies in the early 1990's were the first to suggest that the extreme variability of STRs makes them ideal candidates for developing DNA profiles [15]. These early studies were followed by governmental efforts to develop DNA identification systems. In the United States, 13 autosomal STRs and one sex-specific STR are used to generate a unique DNA signature for each individual [16]. Profiles collected from crime scene samples can then be matched to a suspect's profile or a database of over 12 million profiles to provide valuable information in criminal cases.

STRs on the Y-chromosome have also been used widely due to their unique properties. Because of the lack of recombination on the Y-chromosome, father-son Y-STR haplotypes remain unchanged from generation to generation in the absence of mutations. As a result, genealogists

have used Y-STRs to ascertain the relatedness of families [17] and even resolve historical debates such as the contentious paternal relationship between Thomas Jefferson and Sally Hemings' children [18]. More recently, Gymrek et al. employed these markers to demonstrate that one can infer the surname of an anonymous genome, a finding that stimulated important conversations related to genetic privacy [19]. Y-STR markers are also useful in forensics settings because they are male specific. As a result, Y-STR profiles analogous to CODIS profiles are often used to help resolve sex crime cases.

1.4 Population-scale STR variation

Despite the widespread applicability of STRs, little is known about most STRs in the human genome. Of the nearly 700,000 markers, only a few thousand have been thoroughly characterized. Most of these characterizations were performed during the development of STR linkage panels [20]. However, these studies used only a few families and therefore provide little information about the worldwide variability of these markers. Other large-scale studies have characterized a few thousand markers using ∼1000 unrelated individuals [21] or ∼25,000 trios [22]. Nonetheless, it remains unclear whether the trends observed for these small subsets of markers apply to the rest of the genome.

In addition to their limited scope, previous studies of STR variability have resulted in discordant findings. While many studies have suggested that STRs with tetranucleotide motifs are more variable than those with dinucleotide motifs [23, 22], others have reported the opposite [24, 21]. Similarly, contrasting reports have suggested that the *AT* motif is both the least [25, 21] and most [26, 22] variable dinucleotide repeat unit. These discrepancies largely stem from the fact that each study considered different subsets of markers, resulting in varying degrees of ascertainment biases.

To address these issues, we used an unbiased approach to characterize STRs genome-wide and in over 1000 individuals. **Chapter 2** outlines the resulting analysis, in which we analyzed high-throughput sequencing data from the 1000 Genomes Project [2] using lobSTR[27], a STR-specific variant caller. After assessing the reliability of our calls, we were able to dissect the sequence drivers of STR variability. In particular, we observed that dinucleotide repeats were the most variable class of STR, even after accounting for allele length, and that the *AT* motif is the most variable dinucleotide repeat unit. In addition to these analyses, we also characterized

the levels of linkage disequilibrium between STRs and SNPs and described characteristics of the prototypical STR.

1.5 STRs and complex traits

Although STRs were originally regarded as "junk DNA" and were believed to be nonfunctional, studies continue to uncover instances in which they contribute to complex traits or disease. Many of the first such insights were obtained from dissecting human diseases, over 30 of which have been attributed to STR variations [28]. These diseases primarily involve trinucleotide repeat expansions in exonic regions, the most well known of which is Huntington's disease [29]. However, a subset arise from STR expansions within promoter, intronic and untranslated regions, such as Fragile X syndrome [30]. Many STR-associated diseases are characterized by a switch-like behavior in which individuals with repeat lengths below a threshold are unaffected, but individuals above the threshold manifest the disease. In addition, once above the threshold, these repeats typically expand from generation to generation, resulting in earlier disease onset. While the exact mechanisms behind these repeat disorders remains unclear, increasing evidence suggests that repeat expansions result in protein aggregates [28, 31] or toxic gain-of-function changes to RNA [32, 33]. In addition, a recent study demonstrated that the repeat in Huntington's disease may act by causing splicing errors, highlighting the diversity of STR regulatory mechanisms [34].

While most STR-related diseases are caused by large expansions, minute STR variations can also result in profound phenotypic effects. In Gilbert's syndrome, an increase from 6 to 7 copies of a TA dinucleotide repeat can result in the onset of the syndrome [35, 36]. This STR lies within the TATA box of the promoter for an enzyme that aids in the conjugation of bilirubin. As a result, expansion of the STR impairs transcription of the enzyme, thereby reducing its activity and resulting in elevated levels of unconjugated bilirubin in the bloodstream.

Model organisms have provided a number of examples in which STRs lead to phenotypic variation. In yeast, changes to the STR within the *FLO1* gene modulate the strength of cell surface adhesion [37], while varying the number of repeats in promoters has been directly linked to changes in gene expression [38]. STRs have also been implicated in canine phenotypic variation, as changes to the STR sizes in exons of *Alx-4* and *Runx-2* result in an additional rear claw and different facial morphologies [39]. Finally, an intronic STR in *IIT1* has been shown to impair

Arabidopsis growth at high temperatures [40].

Single-gene studies in humans have also uncovered a myriad of STR regulatory mechanisms. Contente et al. discovered an STR within the promoter of *PIG3* that acts as a transcription factor binding site [41]. As a result, changes in its size drive concomitant changes in gene expression. Hefferon et al. identified a dinucleotide repeat in the intron of *CFTR* that modulates splicing efficiency and linked these changes to the stability of the STR's hairpin structure [42]. More recently, Grunewald et al. uncovered an STR in Ewing Sarcoma that acts as an enhancer, leading to upregulation of a gene implicated in the disease [43].

Despite the diversity of these known regulatory mechanisms, there has been no genome-wide assessment of STR's regulatory impact. Instead, such analyses have mostly focused on SNPs [44] and to a lesser extent CNVs [45]. In **Chapter 3**, we address this limitation by identifying STRs that modulate gene expression (eSTRs). Using STR genotypes from **Chapter 2** and gene expression data for lymphoblastoid cell lines from the gEUVADIS project [44], we identify over 2,000 such markers. We then use heritability analyses to disentangle the relative contributions of SNPs and STR. These analyses suggest that STRs contribute 10-15% of the cis-heritability and therefore play a substantial role in modulating gene expression. Finally, we perform an association analysis using eSTRs and phenotypes from the UK10K project, uncovering 11 novel associations and demonstrating that eSTRs are enriched in clinically-relevant conditions. These analyses constitute the first genome-wide assessment of STR contributions to complex traits.

1.6 STR genotyping methods

The widespread adoption of STRs in the forensics community has led to the development of extremely accurate genotyping methods. To characterize an STR of interest, fluorescently labeled PCR primers are first used to amplify a region containing the locus (**Figure 1-2**). The resulting PCR products are then run through a capillary electrophoresis (CE) machine to determine the size of both STR alleles. CE-based approaches have an estimated accuracy of over 99% [46], but they are limited by their low throughput as individual primers must be designed for every STR. They also have a relatively high per-locus cost, with current kits for characterizing forensic markers costing roughly $0.50-$1 per marker per sample.

In addition to these issues, CE-based approaches have several technical limitations. i) They provide no information about the underlying sequence of the STR, rendering alleles with identical

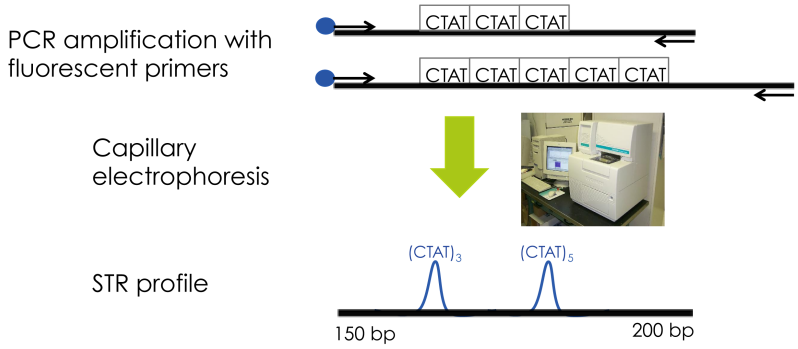

Figure 1-2: **Traditional approach used to genotypes STRs.** For the past two decades, STRs have been characterized using an approach in which fluorescently labeled PCR primers are used to amplify the STR region of interest and some flanking sequence. These amplified DNA segments are then separated by size using a capillary electrophoresis machine, which also measures the fluorescence profile as a function of the fragment length. In the example depicted here, this approach determines that an individual has 3 and 5 copies of a CTAT repeat unit at one STR in the genome.

lengths but different sequences indistinguishable. ii) The PCR primers amplify not only the STR but also the region surrounding it. Insertions and deletions within these regions therefore result in a measured size that is not reflective of the STR's true length. In addition, SNPs within the primer binding sites can result in failed amplification and cause genotyping errors. iii) The PCR amplification step can produce *PCR stutter* products that differ from the true STR's size by one or more repeat units (**Figure 1-3**). These artifacts are particularly prevalent for loci with 2 bp motifs and make it difficult to accurately genotype these markers [46], resulting in their exclusion from most forensic panels [47].

Figure 1-3: **PCR stutter artifacts.** During PCR amplification, a phenomenon known as PCR stutter can insert or delete copies of the repeat unit from the STR region. In this example, although the true STR allele has 3 copies of an ATC repeat, stutter deletions result in reads with less than 3 copies while stutter insertions result in reads with 4 or more copies.

Whole-genome high-throughput sequencing data offers a unique opportunity to address each of

these issues. By providing reads that are informative for hundreds of thousands of STRs, these datasets enable the simultaneous characterization of substantially more markers. In addition, the sequence information contained in each read can be used to determine an STR's sequence in addition to its length. Lastly, the recent development of PCR-free sequencing protocols may mitigate the effects of PCR stutter and enable accurate characterization of STRs with 2 bp periods as well as homopolymer repeats [48].

Despite these numerous advantages, WGS datasets also present unique challenges. To date, most high quality next-generation sequencing data has been generated using the Illumina sequencing platform. This technology has been able to provide increasingly large sequencing datasets with relatively low cost and has been the technology of choice for most population-scale sequencing projects [2, 49, 3]. As a result, many limitations of existing WGS datasets stem from limitations of this platform itself. **Figure 1-4a** provides a very coarse overview of the Illumina sequencing platform. Briefly, to sequence a DNA segment of interest, bridge amplification is first used to generate a cluster of DNA molecules with the same sequence as the original DNA molecule. Each base in the DNA segment is then determined using an iterative process. First, nucleotides with a fluorescent dye are added and incorporated into synthesized DNA fragments by DNA polymerase. These nucleotides contain a blocking group such that additional nucleotides cannot be incorporated. Next, the unincorporated nucleotides are washed away and the fluorescence of each cluster is measured. As each nucleotide is labeled with a unique dye, the fluorescence signal is used to infer the underlying base in the original DNA segment. Lastly, the fluorescent dyes and blocking groups are removed from the incorporated nucleotides and the process is then repeated for the next DNA base [50, 51]. For short DNA segments, the individual DNA fragments within each cluster are typically synthesized in-sync, resulting in clear and robust fluorescent signals that result in accurate base calls (**Figure 1-4b**). However, as longer stretches of DNA are synthesized, individual fragments within each cluster increasingly become out-of-sync, resulting in discordant fluorescent signals and eventually unusable base calls (**Figure 1-4c**).

Due to these issues, Illumina sequencing datasets have primarily been restricted to 100 base pair reads. These short reads make it difficult to characterize STRs that are longer than 60 base pairs, as sufficient sequence flanking the STR is required to map reads to the human genome. Short reads also require sensitive alignment algorithms capable of handling large insertions and deletions. Furthermore, variant callers downstream of alignment tools must be able to account for the unique error profiles present at STR loci and the high frequency of locally misaligned

reads. Although improvements to the Illumina technology have resulted in 150-250 bp reads, these datasets have only recently become available and were not available for most of the analyses in this thesis.

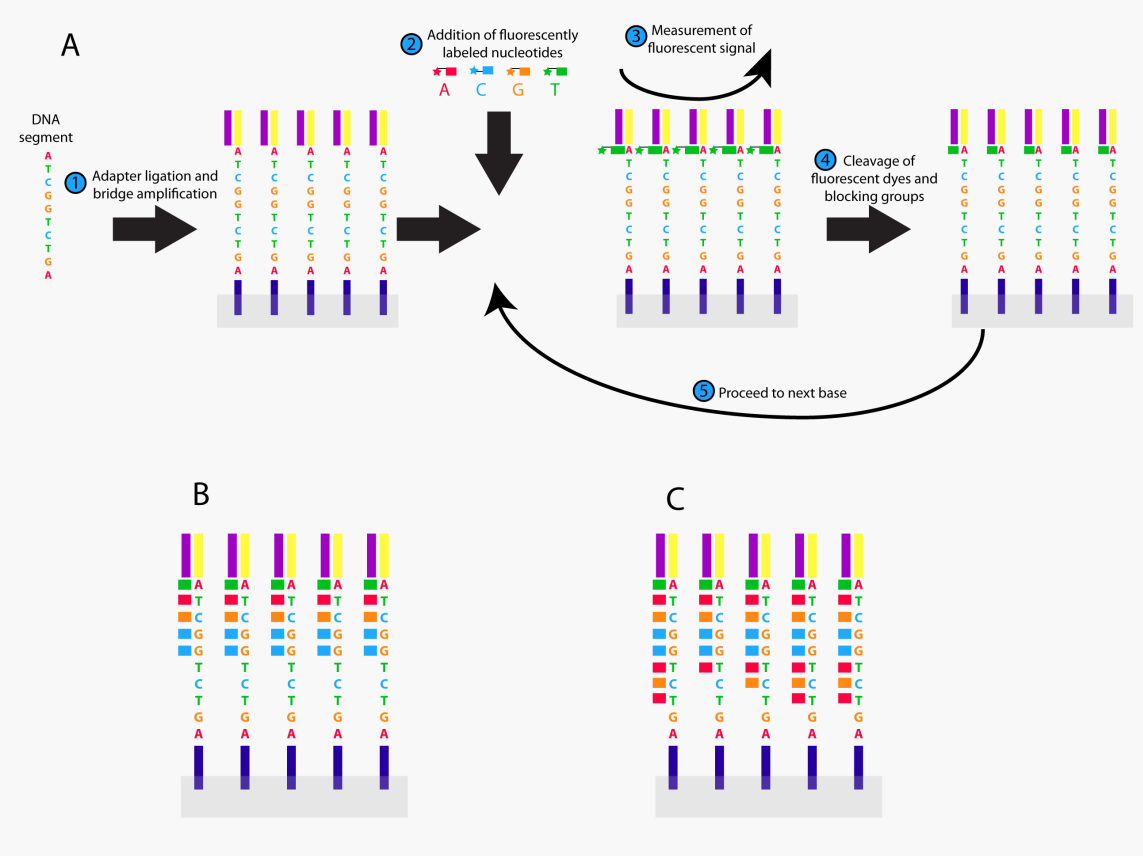

Figure 1-4: **Illumina sequencing technology. (a)** Overview of the steps involved in Illumina sequencing. **(b-c)** Synthesized fragments within each cluster are initially in-phase **(b)** but become increasingly out-of-phase **(c)** as longer DNA stretches are synthesized.

Despite the difficulties associated with WGS data, several STR-specific tools have recently been developed to leverage these rich datasets. lobSTR [27] and RepeatSeq [52], two of the most widely used tools, use the sequence features of STRs to develop error profiles that are incorporated into maximum likelihood or Bayesian genotyping frameworks. In **Chapter 2**, we assessed the accuracy of both of these tools by comparing their STR genotypes to those obtained from capillary electrophoresis. The resulting analyses suggested that they are capable of characterizing the allelic spectra of STRs but have low accuracy for individual genotypes. In addition,

these tools only assess the length of STRs, masking valuable sequence information. Collectively, these limitations require the development of new algorithms before WGS-based STR calls can be used to identify de novo mutations.

Recent studies have also developed state-of-the-art algorithms for genotyping SNPs and indels. Many of these approaches, such as GATK [53] and Platypus [54], use hidden Markov models and local haplotype assembly to mitigate the effects of alignment errors. These tools are widely used in WGS studies to generate call sets and have demonstrated markedly improved accuracy compared to original methods that lacked these features. However, it remains unclear whether they are suitable for characterizing STRs.

Chapter 4 outlines our efforts to develop state-of-the-art algorithms for STR genotyping. We begin by describing HipSTR, a novel haplotype-based caller that determines both the sequence and length of an STR. Using various WGS and capillary datasets, we demonstrate that HipSTR dramatically improves STR genotyping accuracy and is 5 times faster than the next best method. We also demonstrate that STR-specific callers perform poorly while GATK results in robust STR calls. In addition to genotyping methods, we also assess the feasibility of phasing and imputing STRs using HipSTR and slightly modified version of existing algorithms. Finally, we demonstrate the added value of HipSTR by identifying hundreds of de novo mutations in a deeply sequenced trio of individuals. To our knowledge, such an analysis has never been previously performed and emphasizes the importance of considering STR variations in all genetic studies.

1.7 Y-STR mutation rates

Short tandem repeats on the Y-chromosome (Y-STRs) are widely used in genetic genealogy and forensics. The utility of these markers stems from the fact that they are male specific and are transmitted along the paternal inheritance line. As a result, Y-STRs are used to resolve sex crimes [55] and to uncover or confirm familial relationships [17, 18]. The lack of recombination on the Y-chromosome is crucial to these applications, but it also creates problems in forensic scenarios. Because father-son Y-STR haplotypes are identical in the absence of de novo mutations, Y-STRs can rarely differentiate between close patrilineal relatives.

Motivated by these limitations, studies have sought to improve the fidelity and power of Y-STR applications by identifying loci with exceptionally high mutation rates. By genotyping thousands of father-son pairs and assessing the frequency of discordant genotypes, large-scale efforts have

estimated the mutations rate of ∼200 Y-STRs [56, 57]. However, these studies utilize capillary electrophoresis and require enormous samples for accurate inference, making them cumbersome and costly. Moreover, their relatively low throughput has left the mutation rates of hundreds of other polymorphic Y-STRs uncharacterized.

Obtaining a more comprehensive assessment of Y-STR mutation rates is of paramount importance for several reasons. Firstly, as high-throughput sequencing technologies continue to mature, it may become cost-effective to replace capillary-based Y-STR analyses with high-throughput sequencing alternatives. As these approaches would target markers across the Y-chromosome, they could provide much higher discrimination capability than current Y-STR panels of 20-50 markers. To assess the potentials gains of such an approach, more Y-STR mutation rates are needed. Analyzing mutation rates of STRs on the Y-chromosome can also provide insight into genome-wide trends. Though various studies have estimated the number of de novo mutations for SNPs, indels, [58, 59, 60, 61], no such estimate is available for STRs. One approach to obtaining such an estimate is to characterize hundreds of Y-STR mutation rates, build sequence-based predictors of their mutability and apply them to STRs across the genome.

Chapter 5 describes our efforts to estimate mutation rates for all Y-STRs accessible to Illumina technology. We begin by outlining MUTEA, a novel method that uses high-throughput sequencing data from unrelated individuals to obtain these estimates. Through extensive simulations and comparisons to existing estimates, we demonstrate that MUTEA is extremely accurate and obviates the need for current capillary approaches. We then apply MUTEA to over 700 Y-STRs and build sequence-based predictors of mutability. These predictors suggest that STR mutation rates are largely explained by the longest uninterrupted tract length and the period. We also apply these predictors genome-wide, highlighting STRs as a rich source of de novo variation that contribute at least 75 mutations per generation. Finally, we demonstrate that a modified version of MUTEA can also be used to accurately impute many Y-STRs.

Chapter 2

The landscape of human STR variation

Much of this chapter was first published as:

> **Willems TF**, Gymrek M, Highnam G, The 1000 Genomes Project, Mittelman D, Erlich Y. The Landscape of Human STR Variation. *Genome Res.* (August 2014).

Abstract: Short tandem repeats are among the most polymorphic loci in the human genome. These loci play a role in the etiology of a range of genetic diseases and have been frequently utilized in forensics, population genetics, and genetic genealogy. Despite this plethora of applications, little is known about the variation of most STRs in the human population. Here, we report the largest-scale analysis of human STR variation to date. We collected information for nearly 700,000 STR loci across over 1,000 individuals in phase 1 of the 1000 Genomes Project. Extensive quality controls show that reliable allelic spectra can be obtained for close to 90% of the STR loci in the genome. We utilize this call set to analyze determinants of STR variation, assess the human reference genome's representation of STR alleles, find STR loci with common loss-of-function alleles, and obtain initial estimates of the linkage disequilibrium between STRs and common SNPs. Overall, these analyses further elucidate the scale of genetic variation beyond classical point mutations.

2.1 Introduction

STRs are abundant repetitive elements that are comprised of recurring DNA motifs of 2-6 bases. These loci are highly prone to mutations due to their susceptibility to slippage events during DNA replication [62]. To date, STR mutations have been linked to at least 40 monogenic disorders [63, 28], including a range of neurological conditions such as Huntington's disease,

amyotrophic lateral sclerosis, and certain types of ataxia. Some disorders, such as Huntington's disease, are triggered by the expansion of a large number of repeat units. In other cases, such as oculopharyngeal muscular dystrophy, the pathogenic allele is only two repeat units away from the wild-type allele [64, 65]. In addition to Mendelian conditions, multiple studies have suggested that STR variations contribute to an array of complex traits [66], ranging from the period of the circadian clock in Drosophila [67] to gene expression in yeast [38] and splicing in humans [42, 34].

Beyond their importance to medical genetics, STR variations convey high information content due to their rapid mutations and multi-allelic spectra. Population genetics studies have utilized STRs in a wide-range of methods to find signatures of selection and to elucidate mutation patterns in nearby SNPs [68, 22]. In DNA forensics, STRs play a significant role as both the US and the European forensic DNA databases rely solely on these loci to create genetic fingerprints [69]. Finally, the vibrant genetic genealogy community extensively uses these loci to develop impressive databases containing lineages for hundreds of thousands of individuals [70].

Despite their utility, systematic data about the landscape of STR variations in the human population is far from comprehensive. Currently, most of the genetic information concerns a few thousand loci that were part of historical STR linkage and association panels in the pre SNP-array era [20, 71] and several hundred loci involved in forensic analysis, genetic genealogy, or genetic diseases [72, 63]. In total, there are only 5,500 loci under the microsatellite category in dbSNP139. For the vast majority of STR loci, little is known about their normal allelic ranges, frequency spectra, and population differences. This knowledge gap largely stems from the absence of high-throughput genotyping techniques for these loci [73]. Capillary electrophoresis offers the most reliable method to probe these loci, but this technology scales poorly. More recently, several studies have begun to genotype STR loci with whole-genome sequencing datasets obtained from long read platforms such as Sanger sequencing [74] and 454 [75, 76]. However, due to the relatively low throughput of these platforms, these studies analyzed STR variations in only a few genomes.

Illumina sequencing has the potential to profile STR variations on a population-scale. However, STR variations present significant challenges for standard sequence analysis frameworks [77]. In order to reduce computation time, most alignment algorithms employ heuristics that reduce their tolerance to large indels, hampering alignment of STRs with large contractions or expansions. In addition, due to the repetitive nature of STRs, the PCR steps involved in sample preparation induce in vitro slippage events [78]. These events, called stutter noise, generate erroneous

reads that mask the true genotypes. Because of these issues, previous large-scale efforts to catalog genetic variations have omitted STRs from their analyses [2, 79, 80] and early attempts to analyze STRs using the 1000 Genomes data were mainly focused on exonic regions [81] or extremely short STR regions with a relatively small number of individuals based on the native indel callset [82].

In our previous studies, we created publicly available programs that specialize in STR profiling using Illumina whole-genome sequencing data [27, 52]. Recently, we deployed one of these tools (lobSTR) to accurately genotype STRs on the Y chromosome of anonymous individuals in the 1000 Genomes Project to infer their surnames [19], demonstrating the potential utility of STR analysis from Illumina sequencing. Here, we used these tools to conduct a genome-wide analysis of STR variation in the human population using the sequencing data of the Phase 1 of the 1000 Genome Project.

2.2 Results

2.2.1 Identifying STR loci in the human genome

The first task in creating a catalog of STR variation is to determine the loci in the human reference that should be considered as STRs. This problem primarily stems from the lack of consensus in the literature as to how many copies of a repeat distinguish an STR from genomic background [83, 84, 85]. For example, is $(AC)_2$ an STR? What about $(AC)_3$ or $(AC)_10$? Furthermore, as sporadic bases can interrupt repetitive DNA sequences, purity must also be taken into account when deciding whether a locus is a true STR.

We employed a quantitative approach to identify STR loci in the reference genome. Multiple lines of study have proposed that the birth of an STR is a relatively rare event with complex biology [62, 86, 87, 66, 88, 82]. The transition from a proto-STR to a mature STR requires non-trivial mutations such as the arrival of a transposable element, slippage-induced expansion of the proto-STR, or precise point mutations that destroy non-repetitive gaps between two short repeat stretches. Based on these observations, it was suggested that randomly-shuffled DNA sequence should rarely produce mature STR sequences and therefore can be used as negative controls for STR discovery algorithms [66, 85]. We utilized this approach to identify STR loci in the human genome while controlling the false positive rate **(Supplemental Figure 2-**

6; Supplemental Text 2.6.2). We first integrated the purity, composition, and length of putative STRs in the genome into a score using Tandem Repeats Finder [TRF] [89]. Then, we generated random DNA sequences using a second-order Markov chain with similar properties to the human genome (i.e. nucleotide composition and transition frequencies). We tuned the TRF score threshold such that only 1% of the identified STR loci in our collection were expected to be false positives. The resulting score thresholds were in good qualitative agreement with those previously produced using a variety of alternative experimental and analytical methods [90, 91, 84] **(Supplemental Text 2.6.2)**. We then evaluated the false negative rate of our catalog using two methods. First, we collected a preliminary call set of repeat number variability across the human population with a highly permissive definition of STR loci. We found that our catalog misses only ∼1% of loci that exhibited repeat variability in the permissive call set **(Supplemental Table 2.1)**. Second, we also collated a set of about 850 annotated bona-fide STR loci, mainly from the CODIS forensic panel and Marshfield linkage panel. Only 12 (1.4%) of these markers were not included in the catalog based on the TRF score threshold. The results of the two validation methods suggest that our catalog includes ∼99% of the true STRs in the genome and has a false negative rate of about 1%.

Overall, our STR reference includes approximately 700,000 loci in the human genome. About 75% of these loci are di and tetra-nucleotide STRs, while the remaining loci are tri, penta and hexa-nucleotide STRs **(Supplemental Table 2.2)**. Approximately 4,500 loci overlap coding regions, 80% of which have either trimeric or hexameric repeat units. In addition, the reference contains a roughly equal proportion of interrupted and uninterrupted microsatellites.

2.2.2 Profiling STRs in 1000 Genomes samples

We collected variations for these 700,000 STR loci using 1,009 individuals from phase 1 of the 1000 Genomes Project (**Methods 2.4**). These samples span populations from five continents and were subject to low coverage (∼5x) whole-genome sequencing using 76bp and 100bp Illumina paired-end reads. In addition, high coverage exome sequencing data was available for 975 of these samples and was integrated with the whole-genome raw sequencing files.

We tested two distinct STR genotyping pipelines designed to analyze high-throughput sequencing data, namely lobSTR [27] and RepeatSeq [52]. Briefly, lobSTR utilizes the non-repetitive flanking regions surrounding STRs to align reads and assess their allele lengths, while RepeatSeq utilizes Bayesian model selection to genotype previously aligned STR-containing reads. Despite

significant methodological differences, the STR genotypes from the two tools were quite concordant and matched for 133,375,900 (93%) out of the 143,428,544 calls that were reported by both tools. We tested multiple methods to unify the two call sets in order to further improve the quality **(Supplemental Text 2.6.5, Supplemental Figure 2-7)**. However, none of these integration methods improved the accuracy. Since the lobSTR calls showed better quality for highly polymorphic STRs, we proceeded to analyze STR variations using only this call set.

On average, we collected STR genotypes for approximately 530 individuals per locus (**Figure 2-1a**) and 350,000 STR loci per individual (**Figure 2-1b**), accumulating a total of about 350 million STR genotypes in the catalog. We examined the marginal increase in the number of covered STR loci as a function of sample size (**Methods 2.4, Figure 2-1**). Our analysis shows that after analyzing 100 samples, there is a negligible increase in the number of genotyped STRs. However, even with all of the data, 3% of STR loci are persistently absent from the catalog. The average reference allele length of the missing STR loci was 182bp compared to 31bp for the rest of the reference, suggesting that the missing STR loci have allele lengths beyond the read length of Illumina sequencing. We also examined the marginal increase of polymorphic STR loci with minor allele frequencies (MAF) greater than 1%. Again, we observed an asymptote after approximately 100 samples. These saturation analyses suggest that with the current sample size, the STR variation catalog virtually exhausted all loci with MAF>1% that can be observed with 100bp Illumina reads and our analysis pipeline.

The full catalog of STR variations is publicly available at `http://strcat.teamerlich.org` in VCF format. In addition, the website provides a series of graphical interfaces to search for STR loci with specific biological properties such as distance to splice sites, obtain summary statistics such as allelic spectrum and heterozygosity rates, and view the supporting raw sequencing reads.

2.2.3 Quality assessment

To initially assess the accuracy of our STR calls, we first examined patterns of Mendelian inheritance (MI) of STR alleles for three low-coverage trios present in the sample set. In total, we accumulated half a million genotypes calls. Without any read depth threshold, 94% of the STR loci followed MI (**Figure 2-2a**). The MI rates increased monotonically with read depth and restricting the analysis to loci with at least ten reads increased the Mendelian inheritance to over 97%.

Next, we compared the concordance of the calls in our catalog to those obtained using capillary

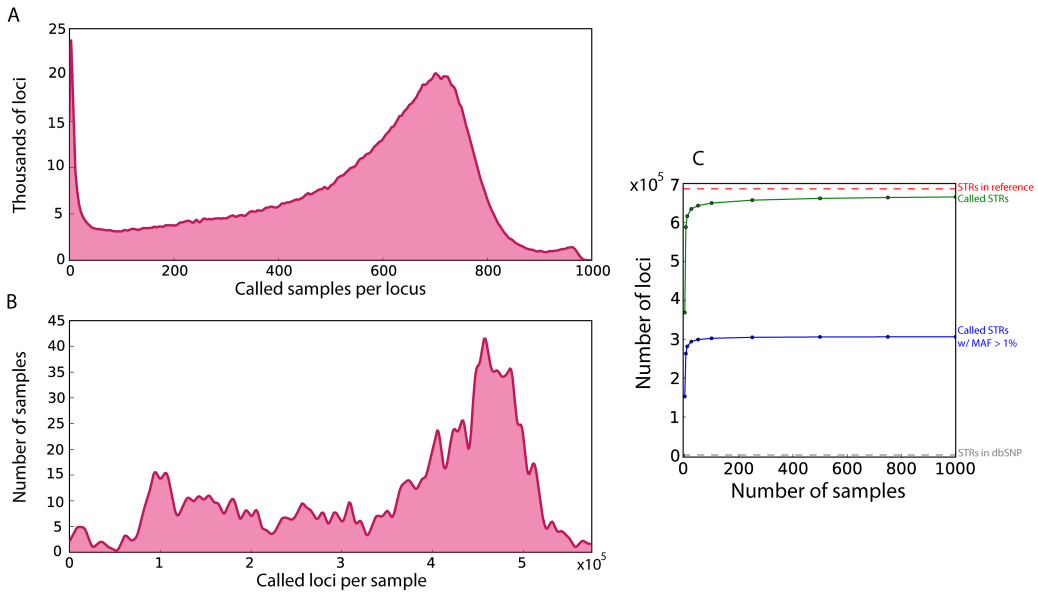

Figure 2-1: **Call set statistics (A)** Distribution of the number of called samples per locus. The average is 528 samples per STR with a standard deviation of 231 **(B)** Distribution of the number of called loci per sample. The average is 349,892 STR per sample with a standard deviation of 145,135 **(C)** Saturation curves for the catalog. The number of called loci (green) rapidly approaches the total number of STRs in the genome (red line). The number of called loci with a MAF>1% (blue) saturates after 100 samples and far exceeds the number of STR variants in dbSNP (grey line close to the Xaxis).

electrophoresis, the gold standard for STR calling (**Methods** 2.4). We focused on datasets containing Marshfield and PowerPlex Y chromosome panel genotypes that are available for a subset of the 1000 Genomes individuals. These panels ascertain some of the most polymorphic STR loci, testing our pipeline in a challenging scenario. The Marshfield capillary panel [92] reported 5,164 genotypes that overlapped with the lobSTR calls and pertained to 157 autosomal STRs and 140 individuals, while the PowerPlex capillary panel reported 784 genotypes that overlapped with the lobSTR calls and pertained to 17 Y-STRs and 228 individuals.

One key question is finding an adequate cost function to assess the concordance between the STR calls. In SNPs, the proportion of mismatches is a natural measure of concordance due to their binary nature. However, for STRs, this approach assigns the same penalty for missing one repeat unit and ten repeat units. As an alternative, we focused on measuring the goodness-of-fit (R^2) between the STR dosages. The dosage of an STR was defined as the sum of the number of

base pairs after subtracting the reference allele. For example, if the genotype was 16bp/18bp and the reference allele was 14bp, the dosage of the locus was set to 2+4=6, while for hemizygous loci the dosage was the difference from the reference allele. We focused on assessing dosage concordance because of the growing body of studies suggesting that the phenotypic impact of STRs is strongly correlated with length [93, 94, 41, 42]. R^2 confers the property that the cost is proportional to the (squared) magnitude of the error in terms of length. In addition, the R^2 of the dosages measures the amount of genetic variance that was recovered by lobSTR under strict additivity, which might be important for downstream association studies.

After regressing the lobSTR dosages with the capillary dosages, the resulting goodness of fit estimators (R^2) were 0.71 for the autosomal genotypes and 0.94 for the Y chromosome genotypes (**Figure 2-2b; Supplemental Figure 2-8**). By further stratifying the autosomal calls by the capillary genotype, we found that lobSTR correctly reported 89.5% of all homozygous loci and recovered one or more alleles for 91.5% of all heterozygous loci, but only correctly reported both alleles for 12.8% of all heterozygous loci **(Supplemental Table 2.4)**. For the Y chromosome, 95% of the lobSTR genotypes exactly matched the capillary genotypes for the PowerPlex Y panel **(Supplemental Table 2.5)**.

Collectively, these results suggest that the individual allele lengths are relatively accurate and that the primary source of noise is the recovery of only one STR allele for heterozygous loci, an issue known as allelic dropout. This statement is supported by the relatively good accuracy achieved for the homozygous autosomal loci and hemizygous Y chromosome loci, and the monotonically increasing relationship between heterozygote accuracy and read depth, with a heterozygote accuracy of nearly 80% achieved for loci covered by 6 or more reads **(Supplemental Figure 2-9)**. In general, allelic dropouts are quite expected given the relatively low sequencing coverage but are also known to be an issue in genotyping STRs with capillary electrophoresis [95].

We performed various analyses that demonstrate that allelic dropouts do not hamper the ability to deduce population-scale patterns of human STR variation. First, we examined the concordance of heterozygosity rates obtained from the lobSTR and the capillary calls for Marshfield STRs in three European subpopulations (CEU, GBR and FIN). The heterozygosity rate is based on the frequency spectrum of a locus (**Methods 2.4**) and should be unaffected by random allelic dropout. As expected, we found that the heterozygosity rates were highly similar between the capillary and the lobSTR results (**Figure 2-2c**). The regression slope was 0.996 and the root mean squared error (RMSE) was 0.044 based on over 200 STRs. This analysis shows that the heterozygosity estimates obtained from our call set are relatively unbiased.

We also benchmarked the quality of population-scale patterns by comparing the allelic spectra for the Marshfield loci **(Supplemental Figure 2-10)**. We found that in most cases, the lobSTR and capillary spectra matched in the median and interdecile range of the reported allelic lengths. We also inspected the frequency spectra of STRs that are part of the forensic CODIS test panel using a similar procedure (**Figure 2-2d; Supplemental Figure 2-11**). A previous study reported the spectra of these loci by genotyping ∼200 Caucasians in the United States using capillary electrophoresis [96]. Again, these comparisons resulted in similar patterns for eight of the ten analyzed markers. We found marked biases only for FGA and D18S51, with lobSTR reporting systematically shorter alleles. As the maximal allele sizes of these two loci are over 80bp, the long alleles are seldom spanned by the mixture of 76 and 100 bp Illumina reads in Phase 1, creating a bias toward shorter alleles.

We sought to further characterize potential biases towards ascertaining shorter alleles with lobSTR and the 76bp/100bp Illumina reads. To that end, we inspected the concordance between the lobSTR calls and the NCBI reference (**Figure 2-2e**). The NCBI reference was generated by long Sanger reads and therefore should be an unbiased estimator of the most common allele in the population. In the absence of any systematic bias towards shorter alleles, the expected deviation of a lobSTR allele from the NCBI reference should be zero. On the other hand, in the presence of such a bias, the lobSTR calls should be systematically smaller than the NCBI reference and generate a negative deviation. We found that the median deviation of lobSTR was around zero for STRs with reference alleles up to 45bp. Above this threshold, we started to observe systematic deviations towards shorter alleles. The deviation did not monotonically decrease but exhibited a local maximum around 65bp, which presumably stems from the heterogeneity of the sequencing read lengths and the exhaustion of alleles that can be spanned by 76bp reads. Importantly, only 10% of all loci in our catalog have a reference allele greater than 45bp. This implies that for the vast majority of the loci, the allelic spectra are expected to be unbiased.

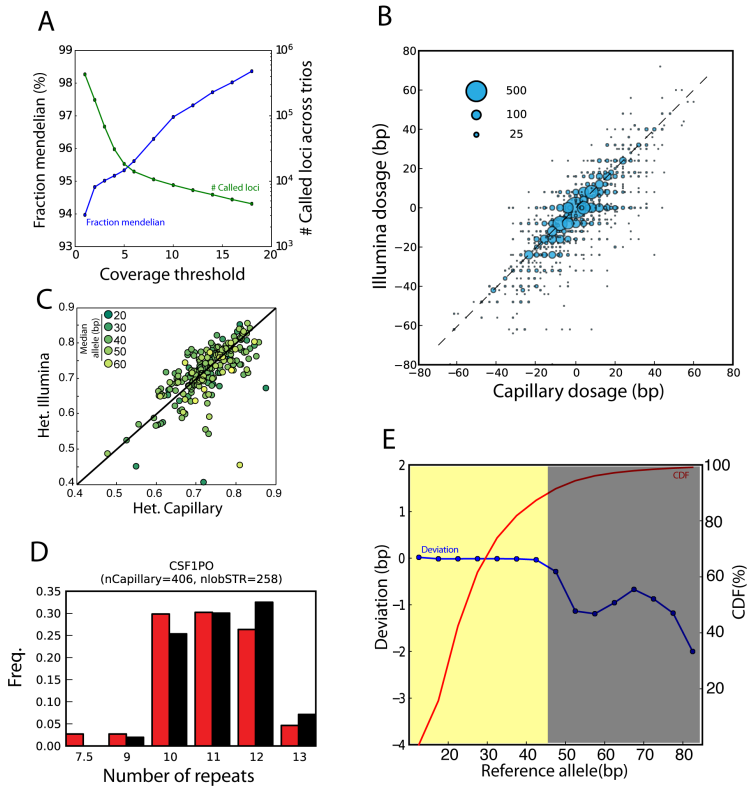

Figure 2-2: **Quality assessments of the STR catalog (A)** Consistency of lobSTR calls with Mendelian inheritance. The blue line denotes the fraction of STR loci that followed Mendelian inheritance as a function of the read coverage threshold. The green line denotes the total number of calls in the three trios that passed the coverage threshold **(B)** Concordance between lobSTR and capillary electrophoresis genotypes. The STR calls were taken from the highly polymorphic Marshfield panel. The dosage is reported as the sum of base pair differences from the NCBI reference. The area of each bubble is proportional to the number of calls of the dosage combination and the broken line indicates the diagonal **(C)** Comparison of heterozygosity rates for Marshfield panel STRs. The color denotes the length of the median allele of the STR (dark-short; bright-long) **(D)** A comparison of allelic spectra obtained by lobSTR and capillary electrophoresis for a CODIS marker in European individuals. Red: lobSTR, black: capillary electrophoresis. nlobSTR and nCapillary indicate the number of alleles called in the respective call sets. **(E)** The reliable range of lobSTR allelic spectra. The figure presents the median deviation of the lobSTR calls from the NCBI reference as a function of the NCBI reference alleles (blue curve). Negative deviations indicate a potential preference towards ascertaining shorter alleles. STRs with reference alleles of up to ∼45bp show very minimal deviations (yellow region) and are expected to display unbiased frequency spectra with the current read lengths. These STR loci comprise close to 90% of the total genotyped STRs in our catalog (red curve).

2.2.4 Validation using population genetics trends

To further assess the utility of our catalog, we tested its ability to replicate known population genetics trends. We specifically wondered about the quality of the most variable STR loci in the catalog. One hypothesis is that these loci are just extreme cases of genotyping errors; an alternative hypothesis is that these loci are truly polymorphic and can provide useful observations about the underlying populations. We first compared the heterozygosities of the 10% most variable autosomal loci across ten different subpopulations from Africa, East Asia, and Europe. Consistent with the Out-of-Africa bottleneck [97], we found that the genetic diversity of the African subpopulations significantly exceeded those of Europe and East Asia (sign test; $p < 10^{-50}$ for any African non-African pair) (**Figure 2-3a; Supplemental Table 2.6**). Second, we focused on the 100 most heterozygous autosomal loci in our catalog and inspected the ability of STRUCTURE [98] to cluster a subset of the samples into three main ancestries in an unsupervised manner. Our results show that all of these samples clustered distinctly by geographical region (**Figure 2-3b**). These analyses demonstrate that even the most variable loci in the catalog still convey valid genetic information that can be useful for population genetic analyses. Finally, we also analyzed the genetic variability of all STRs with MAF>1% on the autosome, X chromosome, and Y chromosome (**Figure 2-3c**). Autosomal STRs showed the highest variability, followed by STRs on the X and the Y chromosomes. This result is consistent with the differences in the effective population sizes of these three types of chromosomes, providing an additional sanity check.

In summary, the multiple lines of quality assessment suggest that our catalog can be used to infer patterns of human STR variations such as heterozygosity, allelic spectra, and population structure. The most notable shortcoming of the catalog is allelic dropouts stemming from the low sequencing coverage of the 1000 Genomes. However, the experiments above suggest that valuable summary statistics can be extracted from the call set despite this caveat.

2.2.5 Patterns of STR variation

Despite a plethora of STR studies, there is no consensus in the literature regarding the effect of motif characteristics on STR variability. The classical study by Weber and Wong [23] originally suggested that tetranucleotide STRs mutate more rapidly than those with dinucleotide motifs based on the analysis of de-novo mutation in trios for 50 STRs. This finding was recently

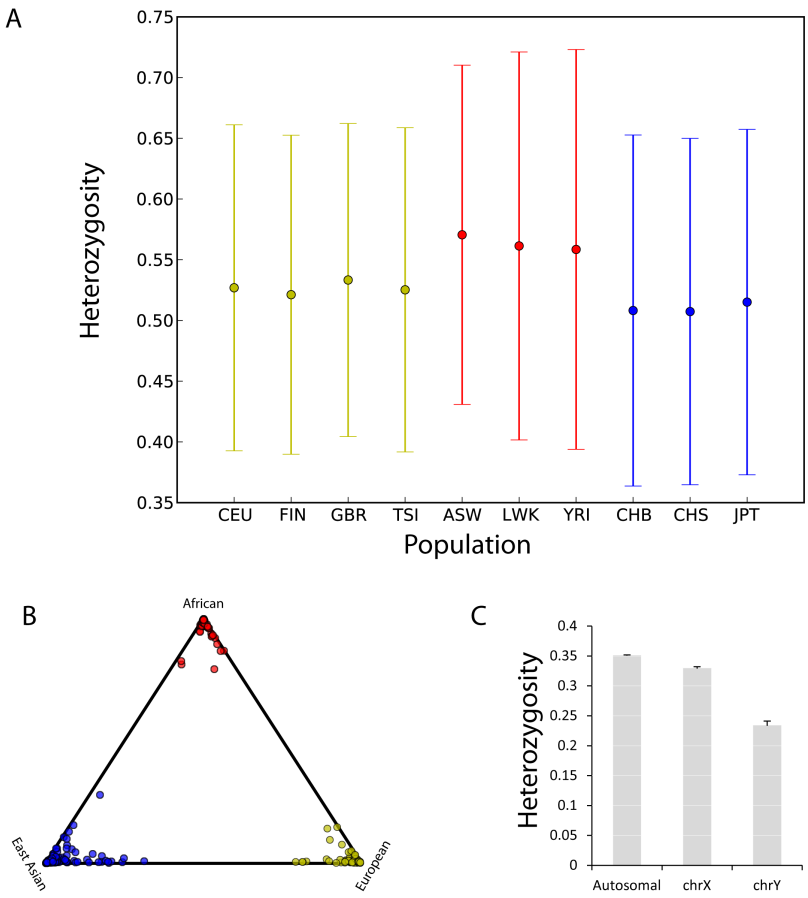

Figure 2-3: **Evaluation of the STR catalog for population genetics**. **(A)** Genetic diversity of the 10% most heterozygous autosomal loci in different populations. Yellow: European, Red: African, Blue: East Asian. The mean heterozygosities (dot) of the African subpopulations consistently exceed those of the non-African subpopulations. The whiskers extend to $\pm$ one standard deviation. **(B)** STRUCTURE clustering based on the 100 most polymorphic autosomal STR loci. Each subpopulation clusters tightly by geographic origin. Color labels as in (A). **(C)** Average STR heterozygosity as a function of chromosome type. Bars denote the standard error.

supported by a much larger trio-based study of nearly 2500 STRs [22]. However, various other studies have suggested that dinucleotides have higher mutation rates [24, 21]. These disagreements may largely stem from the fact that many of these studies considered very small panels of STRs that are subject to ascertainment biases.

To address this open question, we analyzed the sequence determinants of STR variation in

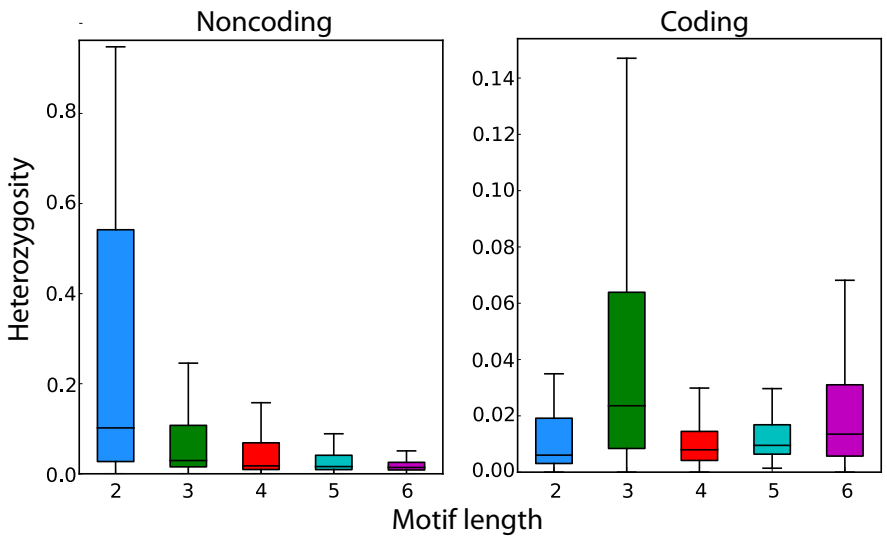

Figure 2-4: **Motif length and coding capabilities as determinants of STR variability.** STR heterozygosity monotonically decreases with motif length for noncoding loci and is generally reduced in non-coding (left) versus coding regions (right). The box extends from the lower to upper quartiles of the heterozygosity distribution and the interior line indicates the median. The whiskers extend to the most extreme points within 1.5×IQR of the quartiles.

our catalog. We found that for noncoding STRs, variability monotonically decreased with motif length (**Figure 2-4**). In contrast, loci with trimeric and hexameric motifs were the most polymorphic among coding STRs. These STR loci can vary without introducing frameshift mutations and therefore may be exposed to weaker purifying selection. In addition, coding STRs demonstrated significantly reduced heterozygosity compared to noncoding STRs for periods 2-5bp (Mann-Whitney U test; $p < 0.01$, **Supplemental Table 2.7**) while hexameric STRs showed no statistically significant difference in variability between these two classes. To ensure that the dependence between motif length and heterozygosity was not confounded by length or purity biases, we stratified STR heterozygosity for pure STRs based on major allele length and motif length. This analysis still showed an inverse correlation between motif length and STR variability after stratification based on the length of the most common allele **(Supplemental Figure 2-12)**. In addition, this analysis showed a monotonic increase in STR variability as a function of the major allele length. Similar trends also applied for STRs with various levels of impurities, albeit with a reduced magnitude of effect and slight deviations from monotonicity **(Supplemental Figure 2-13)**. This observation is concordant with previous studies [99, 100, 90, 101].

Next, we explored the effect of nucleotide composition on STR variability, another issue for which the literature has not yet reached a consensus. Previous studies have suggested that AT repeats are the least variable motif for dinucleotide STRs [25, 21], whereas other studies claimed that AT repeats are the most variable motif [26, 22]. We repeated our analysis by stratifying the STRs based on motif sequence and major allele length **(Supplemental Figure 2-14)**. The resulting per-motif variability results were remarkably similar with those generated using a comparison of orthologous STRs in humans and chimps [26]. Our analysis shows that AT repeats are in general more variable that AC repeats after controlling for length of the most major allele. Similarly, for most motif lengths, STRs with an $[A]_nT$ motif tend to be more variable with long major allele lengths. However, we could not find a clear pattern across motif lengths, which is similar to the result of a previous analysis of a few dozens Y-STRs [56].

2.2.6 The prototypical STR

We also wondered about the prototypical pattern of variation of an STR locus in terms of the number of alleles and their distribution. We found that 30% of STRs have a common polymorphism with at least two alleles with frequencies above 5%. Dinucleotide STRs have the highest rate, with 48% of these loci displaying a common polymorphism. Moreover, 30% of all dinucleotide STRs have more than 3 alleles with a frequency above 5%. On the other hand, hexanucleotide STRs have the lowest common polymorphism rate, with only 13% of these loci displaying a common polymorphism **(Supplemental Figure 2-15A, Supplemental Table 2.8)**.

Next, we turned to finding the prototypical allelic spectra of STRs. For each STR, we normalized the reported alleles such that they reflected the distance in number of repeats from the locus' most common allele. Then, we generated histograms that show the allelic spectra by aggregating all the alleles of STRs with the same motif length. This coarse-grained picture was similar across repeat lengths **(Supplemental Figure 2-15B)**. The allelic spectrum of an STR is unimodal and relatively symmetric. There is one, highly prevalent major allele, two less common alleles one repeat above and one repeat below the most common allele, and a range of rare alleles with monotonically decreasing frequency that reach over Âś5 repeats from the most common allele.

We also wondered about the population differentiation of autosomal STRs. We analyzed the R_{st} [13] for each STR between African, Asian, and European populations for STRs with heterozygosity above 5% **(Supplemental Table 2.9)**. The average Rst was between 4.5-6% across the

motif lengths and the median was around 2-3%. In coding regions, when compared to noncoding STRs, the average Rst was less than half for trimeric STRs but the same for hexameric STRs. Our results regarding population differentiation using STRs are reminiscent of a classical study that found similar levels of differentiation by analyzing close to 800 STR markers [102].

2.2.7 STRs in the NCBI reference and LoF analysis

We were interested in assessing how well the most common alleles are represented in the NCBI reference (**Figure 2-5a**). We found that for over 69,000 loci (10% of our reference set), the most common allele across the 1000 Genomes populations was at least one repeat away from the NCBI hg19 reference allele. Furthermore, the length of the most common allele only matched the length of the orthologous chimp STR 50% of the time, reflecting the high mutability of these loci. In addition, 15,581 loci (2.25%) in the reference genome were 10bp or more away from the most common allele in our dataset.

For STRs in coding regions, the most common allele for 48 loci (1.1% of coding STRs) did not match the allele present in the NCBI reference **(Supplemental Table 2.10)**. In 46 out of 48 of these cases, these differences occurred for loci with trinucleotide or hexanucleotide repeats and conserved the reading frame. Moreover, for the two loci whose most common alleles were frame-shifted, these variations are unlikely to trigger the non-sense mediated decay pathway. The deletion of one 4bp unit in *DCHS2* occurs a few nucleotides before the annotated RefSeq stop codon. This variation slightly alters the location of the stop codon and affects only five amino acids in the C-terminus of the protein. The 14bp deletion in *ANKLE1* occurs in the last exon of the gene and introduces about 20 new amino acids into the tail of the protein.

We also sought to identify a confident set of STR loci with relatively common loss of function (LoF) alleles. To accomplish this goal, we considered only alleles supported by at least two reads and 30% of the total reads per called genotype. We further required that alleles be carried by 10 or more samples. Seven common LoF alleles across five genes passed this criterion: *DCHS2, FAM166B, GP6, SLC9A8*, and *TMEM254* **(Supplemental Table 2.11)**. Out of these 5 genes, only *GP6* has known implications for a Mendelian condition: a mild platelet-type bleeding disorder [103, 104]. However, the LoF mutation in this gene resides in the last exon and is unlikely to induce the non-sense mediated decay pathway. In conclusion, the LoF analysis indicates that common STR polymorphisms rarely disrupt the reading frame.

2.2.8 Linkage disequilibrium between STRs and SNPs

The linkage disequilibrium (LD) structure of STRs and SNPs is largely unknown. On top of recombination events, the SNP-STR LD structure also absorbs STR back mutations that could further shift these pairs of loci towards equilibrium. However, there is minimal empirical data in the literature about the pattern of this LD structure, most of which pertains to a few hundred autosomal Marshfield markers [105]. To get a chromosome-wide estimate, we inspected STR loci on the hemizygous X chromosomes in male samples. Similar to the Y chromosome data, these calls do not suffer from allelic dropouts and are already phased with SNP alleles, conferring a technically reliable dataset for a chromosome-wide analysis.

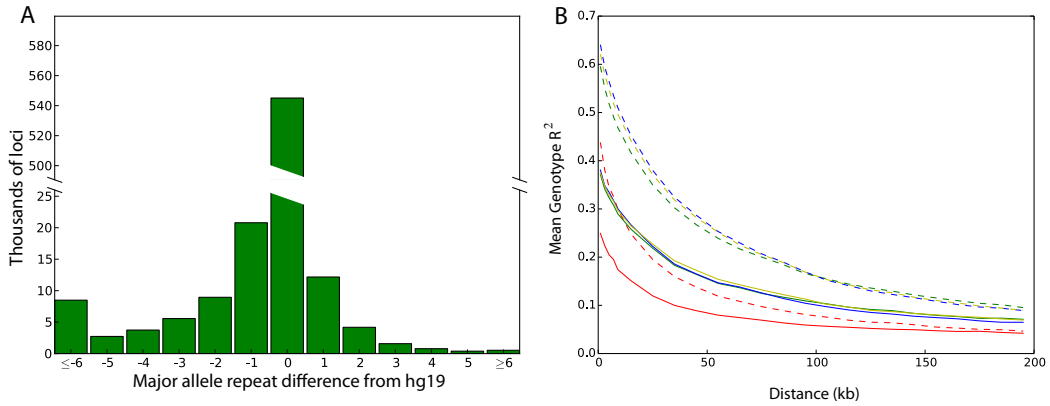

Figure 2-5: **Population-scale analyses of STR variation. (A)** Distribution of base-pair differences between each locus' most common allele and the NCBI reference allele. **(B)** Patterns of linkage disequilibrium for SNPs and STRs on the X chromosome. SNP-SNP LD (dashed lines) generally exceeds SNP-STR LD (solid lines) across a range of distances and for Africans (red), Admixed Americans (green), Europeans (yellow) and East Asians (blue).

We determined the LD in terms of the R^2 between SNPs and STRs as a function of the distance between these markers. Only STRs and SNPs with common polymorphisms were used for the analysis. Hexameric STRs were not included due to the small sample size of 24 sites; for the other repeat motifs, we obtained hundreds to thousands of polymorphic markers. We stratified the STR-SNP LD based on the four major continental populations (Africa, Asia, Europe, and America) and contrasted them to the patterns for classical SNP-SNP LD (**Figure 2-5b**). In all cases, the SNP-SNP LD consistently exceeded mean STR-SNP LD. In addition, the African population demonstrated markedly reduced levels of SNP-STR LD and SNP-SNP LD, consistent

with its larger effective population size. In general, dinucleotide STRs showed the weakest LD with nearby SNPs, which likely stems from their higher mutation rates **(Supplemental Figure 2-16)**. To ensure that the reduction in STR-SNP LD did not stem from comparing R^2 values for multiallelic and biallelic makers, we converted the STR alleles to binary markers, where the two states corresponded to the most common allele and all alternative alleles grouped together. The resulting levels of mean SNP-STR LD using these binary genotypes were nearly identical to those obtained using the multiallelic STR genotypes, indicating that this potential issue had little effect **(Supplemental Figure 2-17)**.

Overall, this analysis shows that the average SNP-STR LD is approximately half of the SNP-SNP LD for variations with the same distance on the X chromosome. Since the effective population size of the X chromosome is smaller than that of the autosome, the STR-SNP LD should be even smaller on the autosome. These results suggest that association studies with tagging SNPs might be considerably underpowered to detect loci with causal STRs, specifically dinucleotide loci.

2.3 Discussion

In the last few years, population-scale sequencing projects have made tremendous progress in documenting genetic variation across human populations. The 1000 Genomes Project has already reported approximately 40 million SNPs, 1.4 million insertion and deletions, and over 10,000 structural variants [2]. Similar catalogs, albeit to lesser degrees of completeness, have been produced for other types of variations, such as LINE-1 insertions [106] and Alu repeat variations [107]. Here, we presented a population-scale analysis of STR variation, adding another layer of genetic variation to existing catalogs.

Our analysis significantly augments the level of knowledge of STR variation. Currently, dbSNP reports data for only 5,500 STR loci. Our catalog provides data on close to 700,000 STR loci, which encompasses 97% of the STRs with motifs of 2-6bp in the genome, and contains over 300,000 STR loci with a MAF of over 1%. One caveat of our catalog is the low reliability of individual genotypes due to allelic dropout. Nonetheless, we showed using multiple lines of analysis that summary statistic results such as frequency spectra and variation trends can be extracted from the catalog for most of the STRs. Another caveat of our catalog is that with the mixture of 76bp and 100bp sequencing reads, we could only unbiasedly ascertain the allelic

spectra of about 90% of the STRs, those with NCBI alleles of up to 45bp. To indicate this caveat, our website alerts users about a potential bias in the allelic spectrum when inspecting STRs with reference allele length beyond this range. However, we expect caveat will be alleviated in the near future with the public release of the Phase 3 data that re-sequenced a large number of Phase 1 samples with 100bp Illumina reads. We expect that this dataset will enable the generation of unbiased allelic spectra for longer STRs.

Despite these limitations, our data provides several biological insights about STR variation. Shorter repeat motif, longer major allele, higher purity of the repeat motif, and residing outside of a coding region are all associated with an increase in STR variability. Most of the STR loci display a unimodal distribution with one very common allele and series of minor alleles with rapidly declining frequencies. This picture suggests that the stepwise mutation model largely describes the creation of new alleles in most of these loci. An open question is the exact mutation rate per generation for each locus in the genome. This question is theoretically addressable with sufficiently large number of samples by analyzing the distribution of squared differences in the repeat size between two alleles of the same locus [13]. However, this question cannot be addressed by our call set due to the large number of allelic dropouts that might confound such an analysis and should be addressed with datasets obtained from deeply covered genomes.

The landscape of STR variations in the apparently healthy 1000 Genomes individuals suggests several rules of thumbs for analyzing STR variations for medical sequencing. Previous work found that membrane proteins of several pathogens contain STR loci with non-triplet motifs whose variations can be beneficial to the organism [66]. These STRs confer high evolvability and adaptability of these proteins by dynamically changing the reading frame. In contrast, our data suggests that for the vast majority of human proteins, frame-shift mutations in their STR regions are not favorable. Only a handful of STRs harbor common frame-shift polymorphisms and half of the LoF alleles create a very small change in the C-terminus tail of the protein. Based on these observations, we hypothesize that most of the non-triplet coding STRs are not well tolerated and are exposed to negative selection similar to regular indels in the same region. Therefore, it is advisable for medical sequencing projects to also analyze these loci and treat them as regular LoF alleles rather than filtering them. This rule of thumb is well-echoed in a recent study of medullary cystic kidney disease type 1 that implicated the genetic pathology in a frame-shift mutation caused by a length change of a homopolymer run [108]. For in-frame STR variations, our call set contains deep allelic spectra of most of these loci, providing reference

distributions of apparently healthy alleles. These spectra can be used to identify atypical STR alleles and might serve as an indicator for pathogenicity.

Although STR alleles within our call set rarely induced frame-shifts, they may introduce premature stop codons by modulating the splicing machinery. Several prior studies have observed a direct dependence of splicing efficiency on STR repeat number for *CFTR* [42], *HTT* [34] and *NOS3* [109]. To facilitate the analysis of such cases, we created a dedicated table on the catalog website that specifies all of the 2,237 STRs that reside within 20 base pairs of an exon-intron boundary.

Another issue raised by our findings is the potential contribution of STRs to complex traits. Using the prototypical allelic spectra, we estimate that the average variance of STR repeat dosage is 3, 0.7, 0.4, 0.25 and 0.1 for 2-6mer STRs, respectively. Interestingly, the theoretical maximum variance for a bi-allelic SNP dosage is 0.5, six times smaller than the observed variance of dinucleotide STRs. From a theoretical statistical genetics perspective, this suggests that causal dinucleotide STR loci could explain a considerable fraction of phenotypic variance even with a relatively modest effect size. Therefore, if each STR allele in a locus slightly changes a quantitative trait in a gradual manner, the net effect on the phenotypic variance could be quite large due to the wide range of these alleles and their relatively high frequencies. Interestingly, we found that loci with dinucleotide motifs show relatively weak LD with SNPs, suggesting that GWAS studies with SNP arrays are prone to miss causal STR loci. Given the theoretical potential of STRs to contribute to phenotypic variance on one hand and their weaker LD to tagging SNPs on the other hand, one intriguing possibility is that STRs contribute to the missing heritability phenomenon of complex traits [110, 111]. Our hope is that this catalog can be a reference point to test this hypothesis in future studies.

2.4 Methods

2.4.1 Call set generation

The raw sequencing files for Phase 1 of the 1000 Genomes Project were analyzed.

The lobSTR calls were generated using computing resources hosted by Amazon Web Services, GitHub version 8a6aeb9 of the lobSTR genotyper and Github version a85bb7f of the lobSTR allelotyper (`https://github.com/mgymrek/lobstr-code`). In particular, the lobSTR genotyper

was run using the options fft-window-size=16, fft-window-step=4 and bwaq=15 and a default minimum flanking region of 8bp on both sizes of the STR region. Reads that were aligned to multiple locations were excluded from the analysis. PCR duplicates were removed from the resulting BAM files for each experiment using SAMtools [112]. The individual BAMs were merged by population and the lobSTR allelotyper was run using all population BAMs concurrently, the include-flank option and version 2.0.3 of lobSTR's Illumina PCR stutter model.

RepeatSeq (available http://github.com/adaptivegenome/repeatseq) was run using default parameters on the read alignments produced by the 1000 Genomes project.

For both programs, we used the set of 700,000 STRs that was constructed using the second-order Markov framework **Supplemental Text 2.6.1**.

2.4.2 Estimating the number of samples per locus and number of loci per sample

The distributions of the call set parameters were smoothed using the gaussian_kde function in the scipy.stats python package. Covariance factors of 0.01 and 0.025 were used to smooth the samples per locus and loci per sample distributions, respectively.

2.4.3 Saturation analysis

We determined the number of loci with calls for sample subsets containing 1, 5, 10, 25, 50, 100, 250, 500, 750 and 1000 individuals. In particular, we began by randomly selecting 1 individual. To create a subset of 5 individuals, we then added 4 more random individuals and so on. For each of these sample subsets, we determined the number of loci with one or more STR calls across all samples in the subset. We repeated this whole process 10 times and used the median number of called loci across each of the 10 repetitions to create the saturation profile for all loci.

We also determined whether loci had a MAF $>$1% using all 1009 samples. We then used a procedure analogous to the one described above to select subsets of samples and determine whether or not each of these loci had a corresponding call in each subset. This procedure resulted in the saturation profile for loci with MAF $>$1%.

2.4.4 Mendelian inheritance

The three low-coverage trios contained within the dataset consisted of the following sample sets: HG00656, HG00657, HG00702 (trio 1), NA19661, NA19660, NA19685 (trio 2) and NA19679, NA19678, NA19675 (trio 3). To assess the consistency with Mendelian inheritance for a given trio, only loci for which all three samples had calls were analyzed. The coverage assigned to each trio of calls corresponded to the minimum coverage across the three samples.

2.4.5 Capillary electrophoresis comparison

We downloaded capillary electrophoresis Marshfield genotypes generated as part of a prior study [92]. Prior to comparing genotypes, offsets were calculated to match the lobSTR calls to the length of the Marshfield PCR products. For each locus, all observed offsets were considered and scored and the optimally scoring offset across all samples was selected. In particular, for each sample, an offset was scored as a 1, 0.5, 0.25 or 0 if the lobSTR calls matched exactly, were homozygous and recovered one Marshfield allele, were heterozygous and recovered one Marshfield allele or did not match at all, respectively. Only loci with at least 20 calls were considered in the comparison. Finally, the Pearson correlation coefficient was calculated using the sum of the allele length differences from hg19 for each locus in each sample.

Y-chromosome PowerPlex genotypes were downloaded from the 1000 Genomes Y chromosome working group FTP site. Offsets were once again calculated to match the length of the PCR products to the lobSTR calls. For each locus, the offset was calculated as the most common difference between the lobSTR and PowerPlex genotypes across samples. Only loci with at least 5 calls were considered in the comparison and the R^2 was calculated between the allele length differences from hg19 for each locus in each sample. In addition, the 15 heterozygous lobSTR calls were ignored.

Slopes and R^2 values for STR dosage comparisons were calculated using the linregress function in the scipy.stats package. To mitigate the effects of outliers, we explored using regular linear regression, regression with a zero intercept and L1 penalized regression. The resulting slopes were essentially invariant to the calculation method and so statistics were reported based on traditional linear regression.

2.4.6 Heterozygosity calculations

For each analysis, heterozygosity was calculated using the aggregated frequency spectra according to the formula $H_E = 1 - \sum_i f_i^2$ where f_i denotes the frequency of the ith allele at the locus.

2.4.7 Summary statistic comparisons

We downloaded the allelic spectra for the Marshfield panel from the Marshfield clinic and parsed them using a custom Perl script. Samples from the CEU, GBR, TSI, and FIN subpopulations were analyzed, and only markers with more than 50 calls were included.

We utilized all of the lobSTR calls for the CEU, GBR and FIN subpopulations to generate the lobSTR frequency spectra for each CODIS marker. Spectra were not available for 3 of the CODIS markers (D21S11, VWa, TPOX). D21S11 is too long to be spanned by Illumina reads; we had annotation difficulties for VWa and TPOX (assigning the correct STR in hg19 to the NIST STR). We then compared the available frequency spectra to those published for a Caucasian population in the United States [96] . Because of some annotation differences between the capillary data and our reference locations, we shifted the lobSTR spectra for the D8S1179 marker by +2 repeat units. Finally, repeat lengths for which the maximum frequency was less than 2% were not displayed.

2.4.8 Comparison of population heterozygosity

To obtain accurate measures of heterozygosity, autosomal STR loci with less than 30 calls in any of the 10 subpopulations considered were ignored. Of the remaining loci, the 10% most heterozygous (24,637 loci) were selected and their means and standard deviations were calculated. To determine whether a pair of populations had systematically different heterozygosity at these loci, we paired the heterozygosities for each locus and counted the number of pairs in which population A had a larger heterozygosity than population B. Ignoring the relatively small number of loci in which heterozygosities were identical, the p-value for this over/underrepresentation was then calculated using the cdf function in the scipy.stats.binom python package.

2.4.9 Deviation of lobSTR calls from the NCBI reference

For each locus with one or more genotyped samples, we calculated the mean deviation of all samples' genotypes from the NCBI reference allele. We then pooled these per-locus deviations by reference allele length using 5bp intervals. The median within each length bin resulted in the corresponding plot of deviation vs. reference allele length.

2.4.10 Sample clustering

STRUCTURE version 2.3.4 was utilized to perform the MCMC-based clustering [113]. The program was run using MAXPOPS=3, BURNIN=500000, NUMREPS=1000000, no prior population information, unphased genotypes, the admixture model and no linkage disequilibrium. All 321 samples from the JPT, CHB, YRI and CEU subpopulations present in the data were clustered based on the 100 most heterozygous autosomal STRs with at least 750 called samples. Samples for which at least 75% of the selected makers were missing calls were not including in the resulting visualization. The final triangle plot therefore contained data for 71, 80, 81, and 82 samples from the CEU, CHB, JPT and YRI populations, respectively.

2.4.11 STR variability trends

Analysis was restricted to STRs with at least 100 called samples. STRs that overlapped an annotated RefSeq translated region were regarded as coding and these annotations were downloaded from the UCSC table browser on 2/11/2014. The mannwhitneyu function in the scipy.stats python package was used to test for significant differences between coding and non-coding STR heterozygosity. For analyses related to allele length or purity, STRs were further restricted to those whose most common allele matched the hg19 reference to enable calculation of the locus' purity. In particular, the purity of each of these STRs was calculated as the fraction of possible positions within the STR region where the subsequent bases corresponded to a cyclic permutation of the STR's motif. The pearsonr function in the scipy.stats python package was used to calculate the Pearson correlation coefficients and their associated p-values, where each STR's length and heterozygosity represented an individual point. Finally, to generate the plots of heterozygosity vs. length, the heterozygosity for each length was calculated as the mean variability of loci within 2bp.

2.4.12 Extraction of orthologous chimp STR lengths

Tandem Repeats Finder was run on the panTro4 assembly of the chimp genome using the default parameters and a minimum score threshold of 5. To resolve overlapping repeats, we discarded repeats with period greater than six and scanned from low to high coordinates and selected the highest scoring repeat for each overlap conflict. The chimp coordinates were mapped to hg19 coordinates using liftOver and a minimum mapping fraction of 50%. We then intersected these coordinates with those of our reference panel and retained those loci within our panel that had a single intersecting chimp repeat whose motif matched. This resulted in orthologous chimp repeats for ∼83% of our reference set of STRs.

2.4.13 Rst levels

The Rst was calculated according to Slatkin [13] using a custom Python script (code available on https://github.com/erlichya/str_catalog_supplemental_scripts). The African, European and Asian populations were comprised of the same subpopulations used throughout this study, except that the ASW population was omitted due to potential admixture. Only loci with heterozygosity above 5% and at least 100 genotyped samples were considered.

2.4.14 Assessing linkage disequilibrium

In order to avoid phasing SNPs and STRs, we only analyzed X chromosome genotypes in male samples. SNP calls for the corresponding samples were obtained from the 1000 Genomes Phase 1 11/23/2010 release and any pseudoautosomal loci were ignored. Analysis of STR-SNP LD was restricted to STR loci with both a heterozygosity of at least 9.5% and at least 20 genotypes for each super population (African, East Asian, European and Ad Mixed American). For each STR that met this requirement, we identified all SNPs within 200 KB of the STR start coordinate. After filtering out SNPs with a MAF below 5% in any of the four super populations, we calculated the level of LD for the remaining STR-SNP pairs. In particular, the R^2 was calculated between the SNP genotype indicator variable and the base pair difference of the STR from the reference. We also recalculated the STR-SNP LD after converting the STR alleles to binary variables, where the most common allele and all alternative alleles were mapped to 0 and 1, respectively. This binary mapping was applied to each super population individually.

For SNP-SNP LD calculations, a seed SNP was identified for each STR meeting the aforementioned requirements. In particular, the SNP closest to the STR's start coordinate with MAF $>$5% for each super population was selected. If no such SNP existed within 1Kb, no SNP was selected and the STR was omitted from the STR-SNP LD analysis. Otherwise, we identified all SNPs within 200 KB of the seed SNP and once again removed SNPs with a MAF $<$5% in any of the super populations. The LD between the seed SNP and each of these remaining SNPs was then assessed as the R^2 between the two SNP genotype indicator variables.

2.5 Acknowledgments

M.G. is supported by the National Defense Science and Engineering Graduate Fellowship. Y.E. is an Andria and Paul Heafy Family Fellow and holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. This study was funded by a gift from Cathy and Jim Stone and an AWS Education Grant award. The authors thank Chris Taylor Smith, Wei Wei, Qasim Ayub, and Yali Xue for providing the results of the Y-STR panel for the 1000 Genomes individuals and the 1000 Genomes Project members for useful discussions. Y.E. dedicates this manuscript to Lia Erlich that was born during the last revision of this work.

2.6 Supplemental Text

2.6.1 Finding putative STR loci in the human genome

To identify putative STRs, we relied on the heuristic that these sequences should rarely occur in randomly generated sequences. To accomplish this aim, we determined the dinucleotide transition frequencies for each of the 22 autosomal chromosomes and 2 sex chromosomes. Then, for each chromosome, we used a second order Markov process to generate a random DNA sequence with the same length and transition frequencies. We repeated this process ten times for each chromosome. Next, Tandem Repeats Finder (TRF) [89] was run on the random and real human chromosomes with a match weight of 2, a mismatch and indel penalty of 7, an 80% probability of matching and a 10% probability of an indel. In cases where TRF reported two overlapping STRs, we selected the locus with the highest score. For each chromosome, the repeats for periods 2-6bp were then analyzed separately to calculate the minimum TRF score

where the number of repeats in the hg19 chromosomal sequence was at least 100 times greater than the mean number in the corresponding ten simulated chromosomes **(Supplemental Figure 2-61a)**. These per-chromosome thresholds were combined into five genome-wide thresholds by taking the maximum threshold across all chromosomes **(Supplemental Table 2.12)**. The identified repeats in hg19 were then filtered to only include those loci with a score above the threshold for its period. Finally, loci that originally overlapped other TRF results were removed, as these loci were generally very impure.

2.6.2 Comparison of STR thresholds to prior studies

For STRs without any indels or SNPs, our empirically determined TRF score cutoffs are equivalent to length thresholds of 11, 14, 14, 16 and 17 base pairs for STRs with 2-6 base pair motifs. A host of previous studies have attempted to determine length thresholds at which a repetitive locus can first be regarded as an STR. Fondon et al. [84] examined WGS data from inbred Drosophila samples and quantified the proportion of variation at repetitive loci attributable to unit step changes. This study suggested that this form of variation, which is consistent with classical STR mutational models, begins to dominate when alleles are 13, 20, 23 and 27 base pairs in length for STR periods 2-5, cutoffs that are stricter than those that we determined. Another study by Kelkar et al. [91] examined polymorphism levels and in-vitro polymerase slippage errors to determine a length threshold where slippage errors dominate. Both analyses suggested that 10 base pairs was the transition threshold for STRs with dinucleotide motifs, in good agreement with our 11 base pair threshold. Lai et al. [90] utilized Markov and multi-type branching processes to match the distribution of repeats produced by a mutation model to those in the human genome, resulting in length thresholds of 8, 12, 16, 20 and 24 base pairs for STRs with 2-6 base pair motifs, respectively. Finally, Ananda et al. [82] focused on uninterrupted STRs with motifs equal or smaller than 4bp. Their definition for an STR was loci with at least two consecutive repeat elements and that are shorter than ∼10 repeat elements). Overall, previous thresholds are less conservative for low periods and more conservative for high periods, but agree qualitatively with our thresholds.

2.6.3 Incorporation of annotated STRs

To collate a set of annotated STR markers, we downloaded published PCR primers for the Marshfield markers(http://www.stanford.edu/group/rosenberglab/repeatsDownload.

html) [21]. We then utilized in silico PCR to determine the genomic coordinates of these primers. Locations for Y-STRs were obtained as described in our previous work [19] and locations for the CODIS markers were determined using PCR primers contained in the NIST database and the method outlined above. To integrate these markers into our reference set, we utilized BEDTools [114] to remove any empirical loci that overlapped annotated loci. The remaining empirical loci combined with all annotated markers comprised our final STR reference.

2.6.4 Assessment of empirical score thresholds with a permissive call set

Having assembled a genome-wide STR reference, we sought to ensure that the chosen TRF score thresholds were appropriate. As dynamic expansions and contractions are hallmarks of STR loci, we examined the rates of polymorphism as a function of TRF score in a permissive call set generated using data from the 1000 Genomes.

To this end, we created a permissive reference by running TRF using relaxed parameters. In particular, we ran the program using a much lower score cutoff of 14 (instead of a score greater than equal to the empirical TRF thresholds) to identify a candidate set of STRs with repeat unit sizes of 2 to 5bp. The other parameters used to run TRF were match=2, indel=5, mismatch=5, match probability=80, indel probability=10, and minimum score=14. We removed STRs that localized to areas that might preclude unique mapping, such as large repeats or transposable elements. Transposons and other repetitive elements were identified using RepeatMasker and the TRF results in or within 20 bases of these regions were removed. We further removed any STRs that were located next to or within 20 bases of another STR. Finally, we pruned the list of STRs on the basis of empirically derived tract length thresholds and purity thresholds, developed in a previous study that characterized the minimum requirements of a sequence to mutate as an STR [84], namely minimum tract lengths of 13, 20, 23, and 27bp for 2-5mers, respectively.

We then loaded this permissive reference to RepeatSeq and generated a preliminary STR call set using data from Phase 1 of the 1000 Genomes project. Encouragingly, we found that loci around the cutoffs identified by running TRF on the random chromosomes were close to fixation. The mean heterozygosity rapidly increased shortly after the threshold **(Supplemental Figure 2-6b)**. This phenomenon matches the hallmark of mature STRs that dynamically expand and contract.

We further quantified the number of polymorphic loci (heterozygosity > 2%) omitted and included by these score thresholds **(Supplemental Table 2.1)**. These analyses revealed that

less than 1% of loci omitted by the thresholds were polymorphic, while roughly 40% of included loci were polymorphic. In addition, our thresholds only omitted 1% of all polymorphic loci. Collectively, these analyses strongly suggest that our score thresholds are well calibrated and have a low false negative rate.

2.6.5 Call set integration

RepeatSeq [52] and lobSTR [27] are currently the two primary tools utilized to genotype STRs in high-throughput sequencing data. As a result, to create the most accurate call set, we sought to assess their individual performance and potentially integrate their calls if it improved accuracy. lobSTR calls were generated as previously described **(Methods)** while RepeatSeq was run using default parameters on the read alignments produced by the 1000 Genomes project.

To assess call set performance, we utilized the Marshfield capillary electrophoresis (CE) genotypes generated for a subset of the 1000 Genomes samples [92]. The loci in this panel are highly polymorphic and provide a challenging test set to compare between lobSTR and RepeatSeq. We regressed the dosage produced by the method of interest (the sum of the predicted base pair differences from the reference allele) versus the dosage obtained from CE **(Supplemental Figure 2-7A-B)**. Comparing the calls produced by the individual methods indicated that lobSTR outperformed RepeatSeq as the R^2 values were 0.71 and 0.4, respectively. While RepeatSeq produced more calls, they were in general strongly biased towards the reference allele genotype.

To integrate calls from RepeatSeq and lobSTR, we first had to combine the genotype likelihoods of lobSTR and RepeatSeq. While RepeatSeq reports $P(genotype|data)$, lobSTR reports $P(data|genotype)$. Therefore, we generated comparable posteriors for the lobSTR calls by using the population-wide occurrences of each STR allele in the lobSTR VCF GT field as a prior.

In total, we explored three different simplistic integration strategies. The first strategy selected the genotype with the highest posterior likelihood across both methods. This method assumes that the lobSTR and RepeatSeq posteriors are well calibrated and therefore defers to the method with the most confidence. We explored two variants of this strategy by considering a) only calls for which both methods produced genotypes and b) calls for which either method produced genotypes. The resulting R^2 values of 0.52 and 0.53 **(Supplemental Figure 2-7C-D)** indicated that the integrated calls outperformed those of RepeatSeq but were still inferior to those of lobSTR alone. The next integration strategy selected the genotype with the highest mean posterior. When we applied this strategy in combination with the two aforementioned vari-

ants, the R^2 values were nearly identical to those obtained by selecting the maximum posterior **(Supplemental Figure 2-7E-F)**. Finally, we employed a simple strategy in which we only considered concordant calls. In addition to greatly limiting the number of calls, this strategy was surprisingly worse than the previous two integration strategies as it resulted in an R^2 of 0.46 **(Supplemental Figure 2-7G)**.

In summary, despite various attempts to integrate RepeatSeq and lobSTR calls, our efforts were ultimately unsuccessful. Though the integrated calls were superior to those of RepeatSeq alone, they were ultimately inferior to lobSTR's calls. The results of these efforts suggested that RepeatSeq calls had systematic biases and that these biases, when integrated with lobSTR calls, persisted. We therefore chose to proceed using only the lobSTR calls.

2.6.6 Homopolymer STRs

There is no consensus in the literature whether homopolymers are considered STRs or not ([62] vs. [74]). Our original call set included homopolymer repeats. Similar to the other STR periods, we created a reference panel of homopolymers and genotyped them in all of the individuals. However, we found that the PCR stutter artifacts within the raw sequencing reads were too noisy to reliably recover STR alleles with such low sequencing coverage. For example, we examined the STR calls on the X chromosome of males. Ideally, we should observe only a single allele in this callset. However, we found that reads for homopolymer STRs have a $\sim$20% chance of containing a stutter product rather than the original allele. This contrasts greatly with the estimated stutter rates for the other periods, which we estimated to be between $\sim$2 and 4%. As a vast fraction of the calls use only 1 read, we were concerned about the quality of the data. In addition, mononucleotides showed low Mendelian inheritance patterns in the two trios (70% for homopolymers vs. >90% for other repeats). The low Mendelian inheritance for homopolymers never reached more than 85% even after analyzing STRs with more than 20 reads. Finally, there was no significant reduction in heterozygosity in exonic regions for homopolymer STRs. Based on these analyses, we decided to omit the homopolymer calls from out catalog.

2.7 Supplemental Tables

Table 2.1: Polymorphism levels of loci omitted and included by TRF score cutoffs

	Loci with score below threshold			Loci with score $\geq$ threshold		
Period	Polymorphic	Total	% Polymorphic	Polymorphic	Total	% Polymorphic
2	227	40803	0.56%	25431	47446	53.60%
3	28	4675	0.60%	1178	7316	16.10%
4	17	9171	0.19%	500	10500	4.76%
5	0	2237	0.00%	81	2405	3.37%
All	272	56886	0.48%	27190	67667	40.18%

Table 2.2: The distribution of STRs in the lobSTR reference

Period	Number of STRs
2	277822
3	77327
4	220859
5	72637
6	40867
Total	689512

Table 2.3: Population breakdown of genotyped samples

Please see the original publication

Table 2.4: Marshfield concordance statistics

	lobSTR A/A call	lobSTR C/C call	lobSTR A/B call	lobSTR A/C call	lobSTR C/D call
Homozygous Marshfield Sites (A/A)	1465 (89.5%)	119 (7.3%)	N/A	52 (3.2%)	1 (< .1%)
Heterozygous Marshfield Sites (A/B)	2681 (76.0%)	290 (8.2%)	452 (12.8%)	95 (2.7%)	9 (.3%)

Marshfield genotypes are always A/A or A/B.
lobSTR calls with C and/or D alleles indicate alleles that were not observed in the Marshfield genotype.

Table 2.5: Y-STR PowerPlex concordance statistics

Locus	Correct Calls	Total Calls	% Correct
DYS481	14	19	73.68
DYS458	10	12	83.33
DYS533	60	66	90.91
DYS389I	53	58	91.38
DYS392	36	39	92.31
DYS391	98	105	93.33
DYS456	47	50	94.00
DYS438	51	53	96.23
DYS393	57	59	96.61
DYS439	83	85	97.65
Y-GATA-H4	49	50	98.00
DYS549	66	67	98.51
DYS19	15	15	100.00
DYS437	14	14	100.00
DYS570	15	15	100.00
DYS576	26	26	100.00
DYS643	51	51	100.00
Total	745	784	95.03

Table 2.6: Comparing heterozygosities between populations

	CEU	FIN	GBR	TSI	ASW	LWK	YRI	CHB	CHS	JPT
CEU		0.53	0.46	0.51	0.36	0.39	0.40	0.55	0.55	0.53
FIN	0.47		0.43	0.48	0.35	0.38	0.38	0.53	0.53	0.50
GBR	0.54	0.57		0.55	0.38	0.41	0.41	0.57	0.57	0.55
TSI	0.49	0.52	0.45		0.36	0.39	0.39	0.54	0.54	0.52
ASW	0.64	0.65	0.62	0.64		0.53	0.54	0.68	0.67	0.66
LWK	0.61	0.62	0.59	0.61	0.47		0.51	0.65	0.64	0.63
YRI	0.60	0.62	0.59	0.61	0.46	0.49		0.64	0.64	0.62
CHB	0.45	0.47	0.43	0.46	0.32	0.35	0.36		0.5	0.46
CHS	0.45	0.47	0.43	0.46	0.33	0.36	0.36	0.5		0.46
JPT	0.47	0.50	0.45	0.48	0.34	0.37	0.38	0.54	0.54	

Only the 10% most variable STRs were used.
Each cell represents the fraction of loci for which population A (row) had higher heterozygosity than population B (column).

Table 2.7: Comparison of heterozygosity in noncoding and coding STRs

	Mean Heterozygosity		Median Heterozygosity			
Period	Noncoding	coding	Noncoding	Coding	p-value	Adj. p-value
2	0.274	0.033	0.102	0.005	5.78E-38	1.74E-37
3	0.129	0.065	0.029	0.023	2.01E-46	1.21E-45
4	0.109	0.027	0.018	0.007	1.07E-09	1.60E-09
5	0.083	0.023	0.016	0.011	1.76E-03	2.11E-03
6	0.050	0.044	0.014	0.014	5.37E-02	5.37E-02

Table 2.8: Allele frequency distributions for STR type subsets

Please see the original publication

Table 2.9: Rst levels between Africans, Asians, and Europeans

Period	#	Max	Mean	Median	Std
2	150786	0.940	0.048	0.022	0.065
2	159	0.303	0.037	0.0158	0.054
3	23800	0.598	0.050	0.024	0.065
3	578	0.329	0.021	0.0055	0.041
4	52990	1.000	0.059	0.034	0.072
4	43	0.252	0.064	0.0354	0.069
5	14051	0.613	0.058	0.032	0.071
5	25	0.307	0.063	0.0461	0.080
6	4388	0.515	0.044	0.019	0.062
6	47	0.199	0.043	0.0141	0.056

STRs in noncoding regions
STRs in coding regions

Table 2.10: Coding STRs with non-reference major alleles

Please see the original publication

Table 2.11: Common LoF alleles

chr	STR (start-stop)	Diff from hg19 (bp)	Motif	# Samples	Exon (start-stop)	Gene
4	155244402 -155244432	-4	AAAC	39	155244389 -155244481	DCHS2
9	35561913 -35561938	-8	ACCC	22	35561864 -35562092	FAM166B
10	81841429 -81841443	-4	AAAG	25	81841395 -81841490	TMEM254
10	81841429 -81841443	1	AAAG	111	81841395 -81841490	TMEM254
10	81841429 -81841443	2	AAAG	31	81841395 -81841490	TMEM254
19	55526092 -55526121	4	ACAG	82	55525449 -55526533	GP6
20	48467310 -48467334	-1	AC	94	48467298 -48467381	SLC9A8

Table 2.12: TRF score cutoffs

Period	Score Cutoff
2	22
3	28
4	28
5	32
6	34

2.8 Supplemental Figures

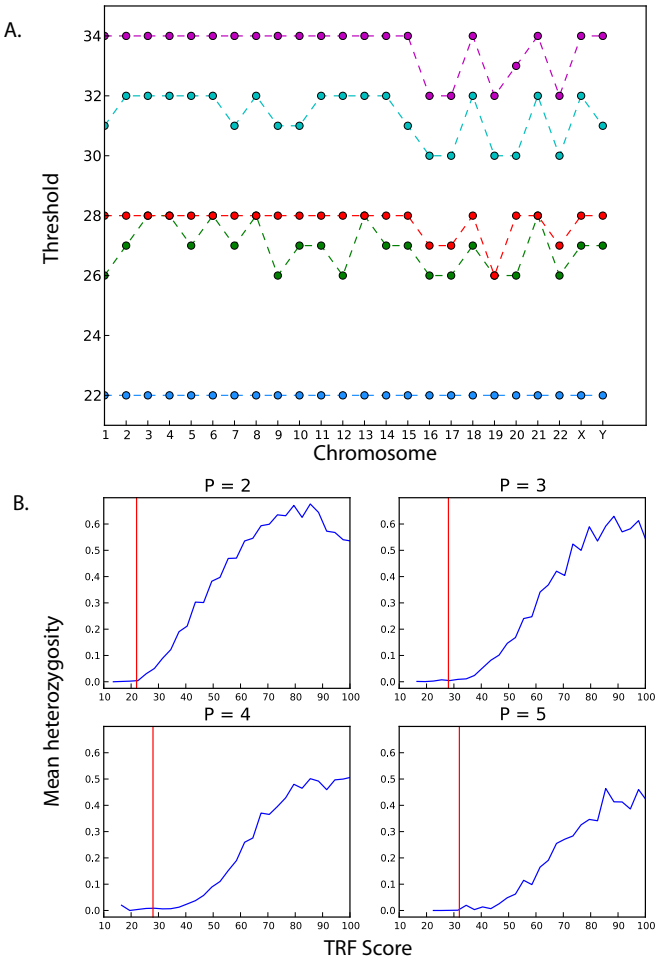

Figure 2-6: **Evidence-based criteria for STR loci. (A)** The TRF score threshold for each chromosome and motif length. These thresholds were calculated based on an FDR of 1% using the second-order Markov chain simulations (blue: 2mer, green: 3mer, red: 4mer, cyan: 5mer, purple: 6mer). **(B)** Validating the thresholds with a preliminary call set. The plots show the average heterozygosity levels (y-axis) for STRs in the permissive catalog as a function of their TRF scores (x-axis). P denotes the motif length in bp. The red line shows the thresholds that were selected for the final definition based on the Markov chain simulations. The putative STRs around the thresholds are close to fixation and STRs with TRF score above the threshold show a rapid increase in their heterozygosity. This indicates that the thresholds are well calibrated and include most of the STRs that are subject to contractions and expansions.

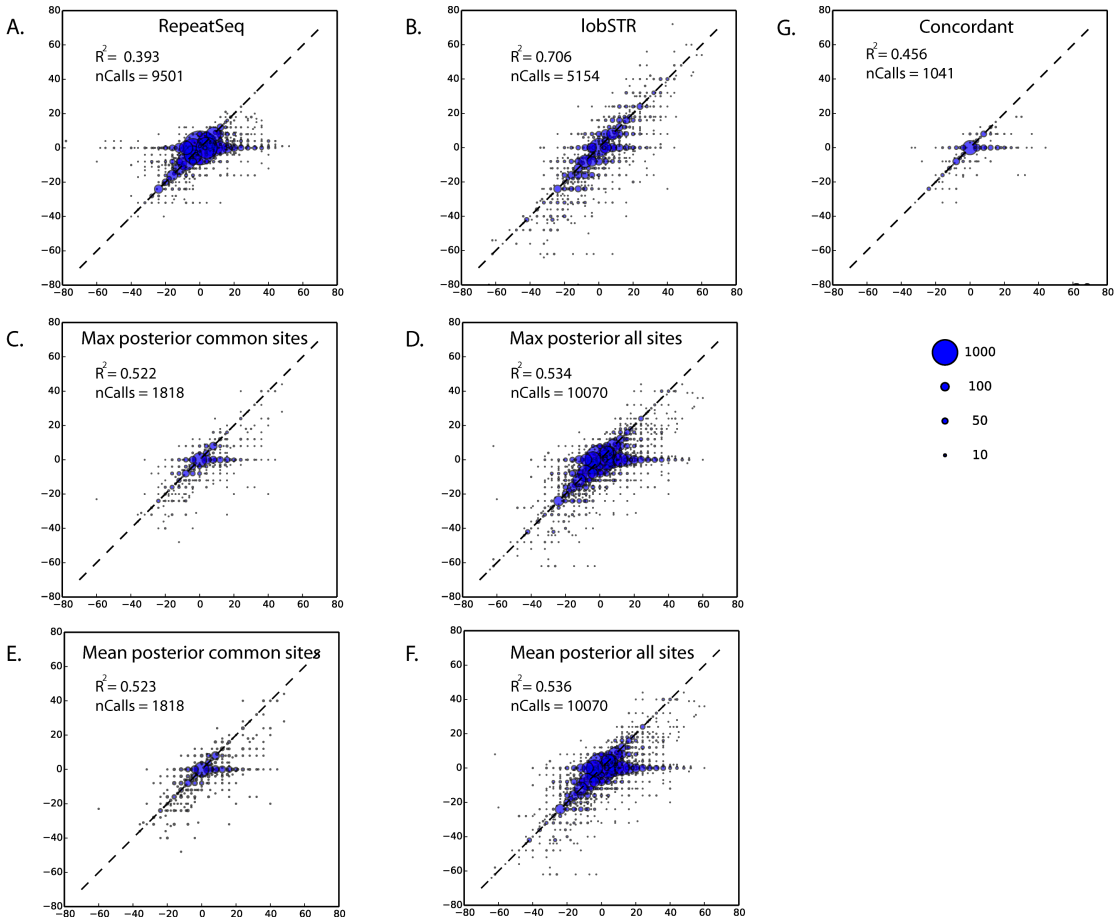

Figure 2-7: **Integration efforts for RepeatSeq and lobSTR.** Each bubble plot shows the regression of the Marshfield capillary dosages (x-axis) with a different method to obtain STR calls from the 1000 Genomes (y-axis). The R^2 and number of calls (nCalls) are reported **(A)** RepeatSeq alone **(B)** lobSTR alone **(C)** RepeatSeq+lobSTR integration based on maximum posterior for sites that appeared in both datasets **(D)** RepeatSeq+lobSTR integration based on maximum posterior for sites that appeared in at least one dataset **(E)** RepeatSeq+lobSTR based on mean posterior for sites that appeared in both datasets **(F)** RepeatSeq+lobSTR based on mean posterior for sites that appeared in at least one dataset **(G)** RepeatSeq+lobSTR integration by reporting only genotypes concordant between the two methods. The best R^2 was obtained by lobSTR alone.

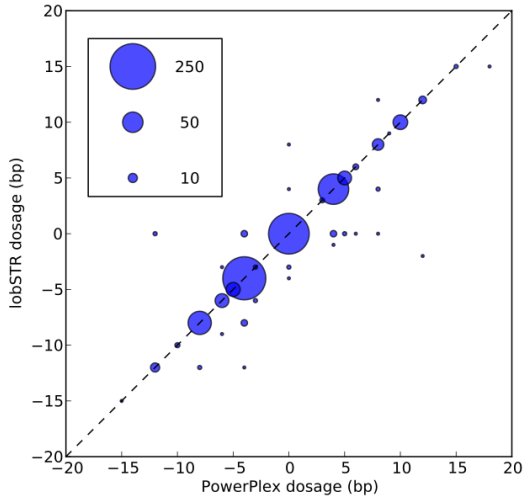

Figure 2-8: **lobSTR dosage concordance with capillary electrophoresis for hemizygous Y-STRs.** The dosage is reported as the base pair difference from the NCBI reference. The area of each bubble is proportional to the number of calls of the dosage combination and the broken line indicates the diagonal.

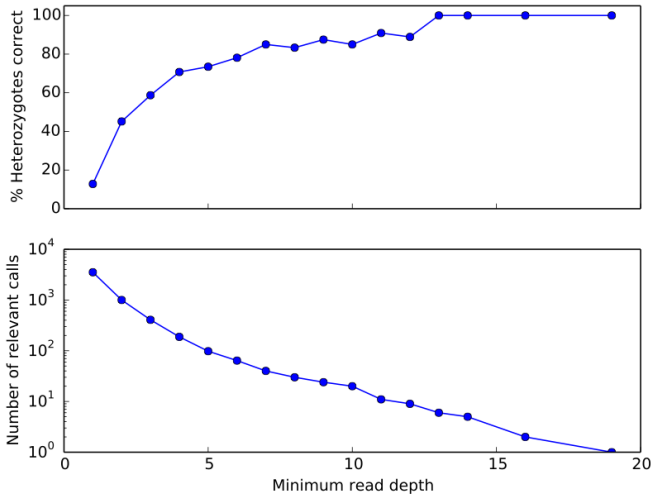

Figure 2-9: **Influence of read depth on the accuracy of lobSTR genotypes for heterozygous sites in the Marshfield panel.** The fraction of heterozygous individuals correctly genotyped increases monotonically with the minimum number of spanning reads.

Figure 2-10: Please see the original publication.

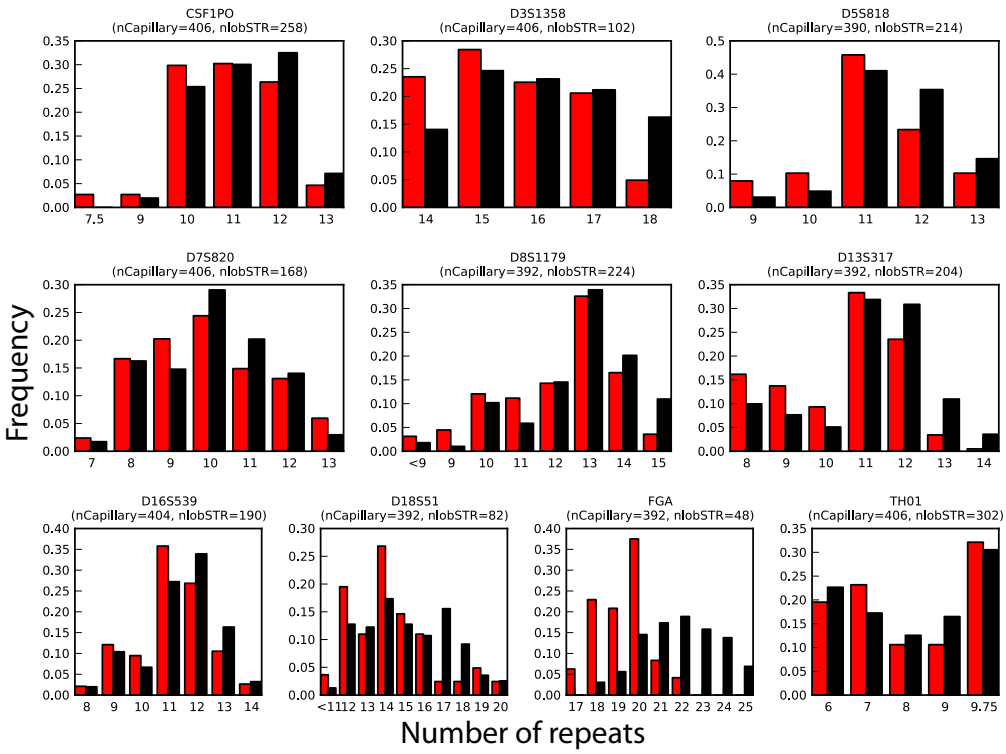

Figure 2-11: **Allelic spectra of the CODIS markers**. Red: lobSTR, black: capillary electrophoresis. nlobSTR and nCapillary indicate the number of allele called in the respective call sets.

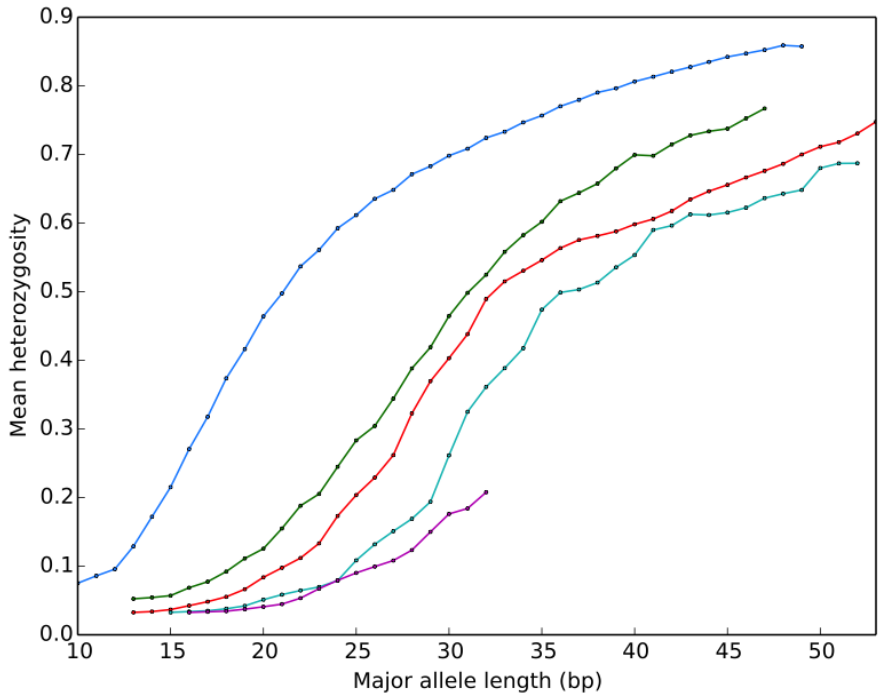

Figure 2-12: **STR variability as a function of the length of the most common allele (base pairs).** The mean heterozygosity of STRs increases monotonically with allele length for each of the five STR periods. Analysis was restricted to STRs whose most common allele matched the reference and with no indels or SNPs interrupting the STR motif (blue: 2mer, green: 3mer, red: 4mer, cyan: 5mer, purple: 6mer). The curves were smoothed by averaging the data points by a sliding window of $\pm$ 2bp

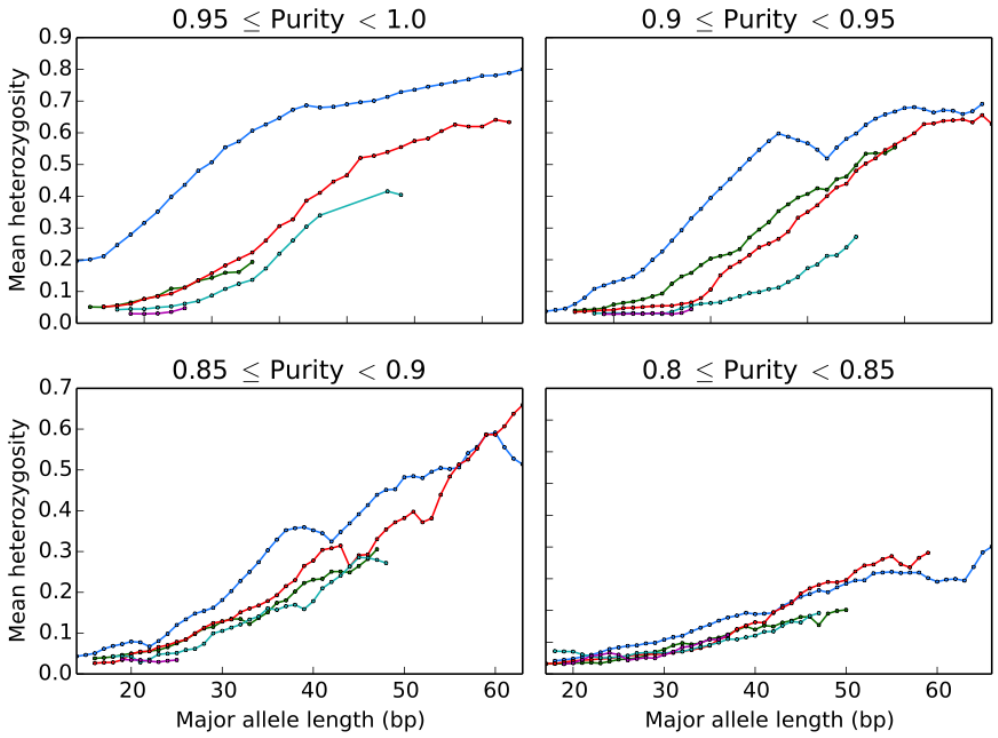

Figure 2-13: **STR variability as a function of major allele allele length (bp) for impure STR loci.** Analysis is stratified based on motif length (blue: 2mer, green: 3mer, red: 4mer, cyan: 5mer, purple: 6mer) and the purity of the STR (see methods) and is restricted to STRs whose major allele matches the reference. The curves were smoothed by averaging the data points by a sliding window of ± 2bp.

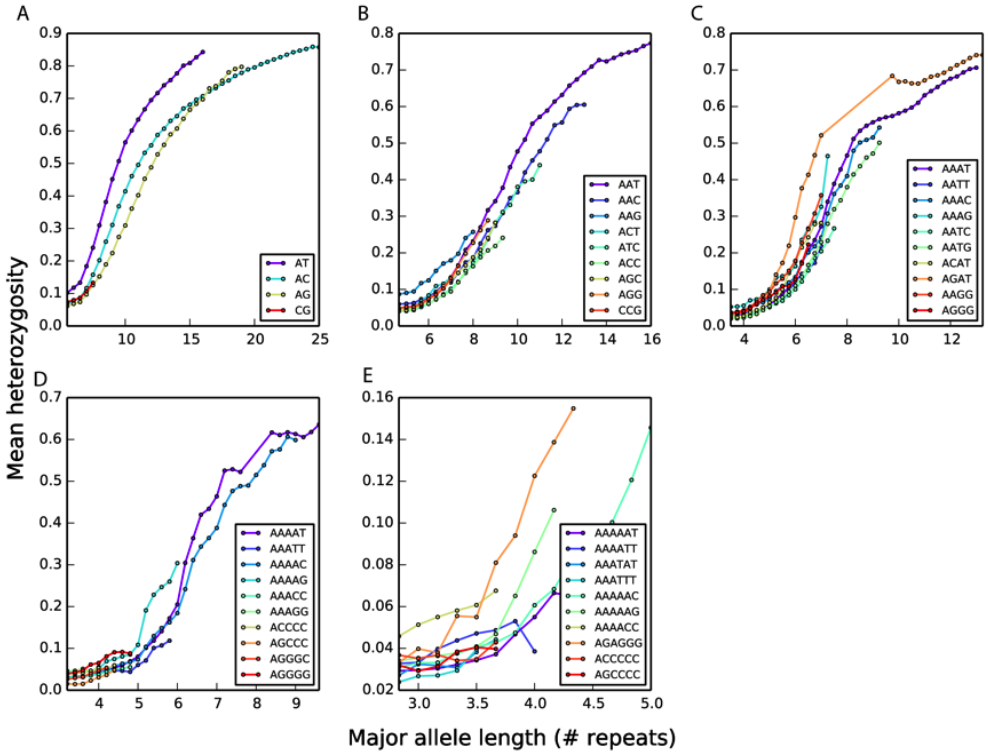

Figure 2-14: **STR variability as a function of length for (A) Dinucleotide, (B) Trinucleotide, (C) Tetranucleotide, (D) Pentanucleotide and (E) Hexanucleotide motifs.** The motif sequences were converted to their canonical form, namely the sequence with the highest lexicographic order among all cyclic permutations of the STR motif from both strands (e.g. AATA repeat was converted to AAAT). For the same period, different motifs can have substantially different mean levels of variability. Analysis was restricted to STRs whose major allele matched the reference allele and to loci without any SNPs or indels that disrupted the motif sequence. The curves were smoothed by averaging the data points by a sliding window of $\pm$ 2bp.

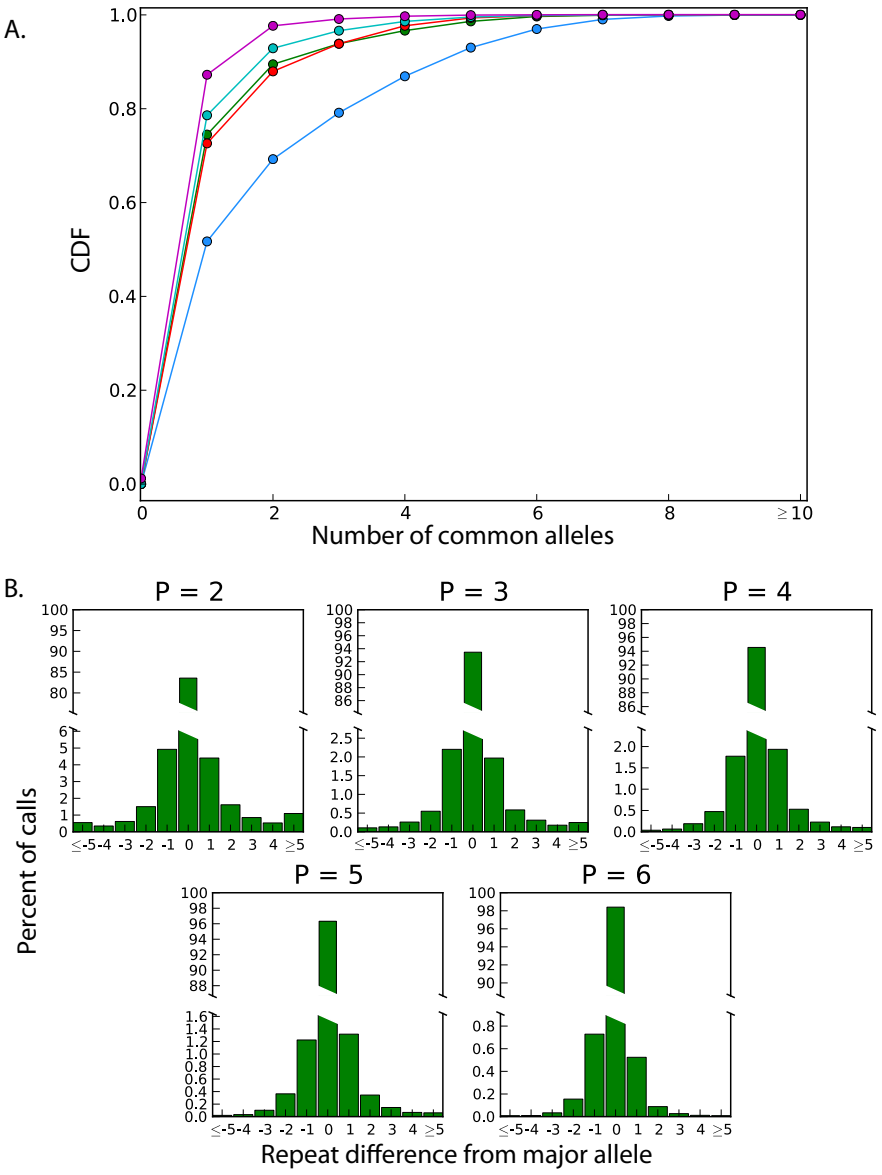

Figure 2-15: **Patterns of STR variation (A)** The cumulative distribution function of the number of alleles with MAF>5% stratified by motif length (blue: 2mer, green: 3mer, red: 4mer, cyan: 5mer, purple: 6mer) **(B)** The averaged allelic spectra of STRs. P denotes the motif length in bp. The 0 allele is the most common allele.

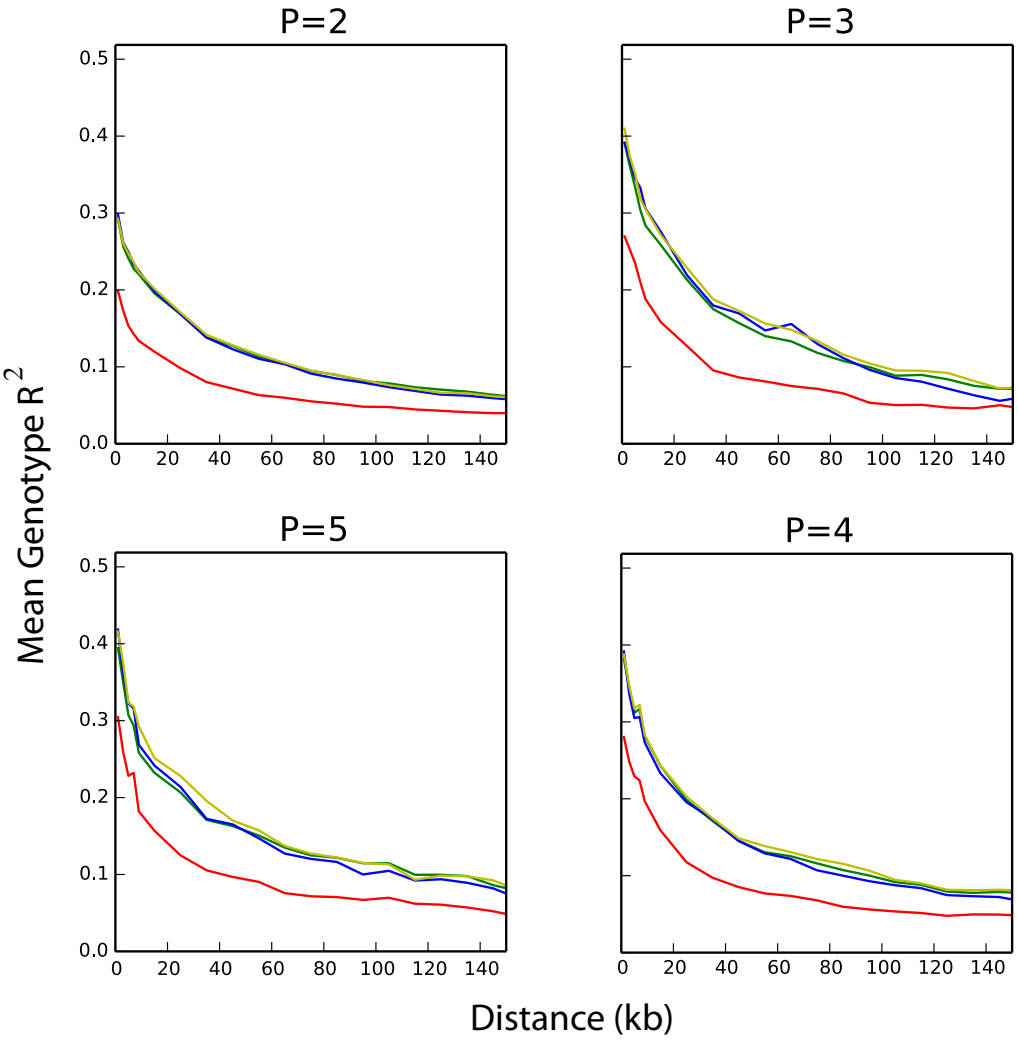

Figure 2-16: **STR-SNP linkage disequilibrium on chromosome X stratified by motif length.** P denotes the motif length in bp. Africans (red), Admixed Americans (green), Europeans (yellow) and East Asians (blue). Longer repeat motifs show an increase in the level of STR-SNP LD.

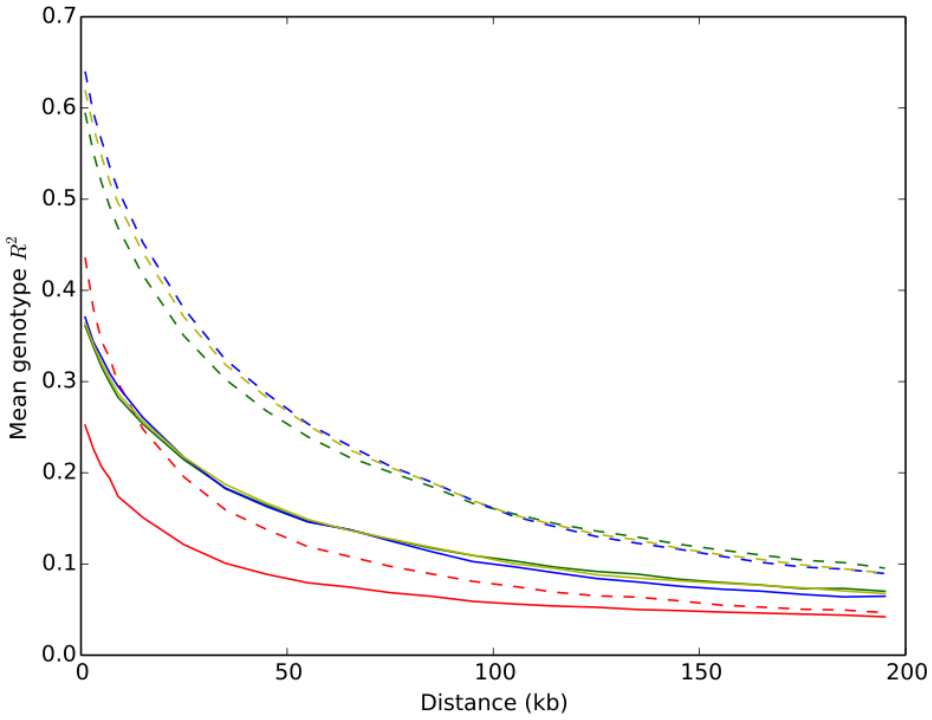

Figure 2-17: **Linkage disequilibrium for SNPs and STRs on the X chromosome after STR allele binarization.** Patterns of STR-SNP LD are invariant to STR allele binarization. As in the non-binarized case, SNP-SNP LD (dashed lines) generally exceeds SNP-STR LD (solid lines) across a range of distances and for Africans (red), Admixed Americans (green), Europeans (yellow) and East Asians (blue). Binarization was performed by mapping the most common STR allele and all alternate alleles to 0 and 1, respectively.

Chapter 3

Abundant contribution of short tandem repeats to gene expression variation in humans

---

Most of this chapter was first published as:

> Gymrek M, **Willems TF**, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, Erlich Y. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature Genetics*. (2015).

Melissa Gymrek was the lead author for this paper and ultimately performed most of the analyses outlined in this chapter.

Abstract: The contribution of repetitive elements to quantitative human traits is largely unknown. Here, we report a genome-wide survey of the contribution of Short Tandem Repeats (STRs), one of the most polymorphic and abundant repeat classes, to gene expression in humans. Our survey identified 2,060 significant expression STRs (eSTRs). These eSTRs were replicable in orthogonal populations and expression assays. We used variance partitioning to disentangle the contribution of eSTRs from linked SNPs and indels and found that eSTRs contribute 10%-15% of the cis-heritability mediated by all common variants. Further functional genomic analyses showed that eSTRs are enriched in conserved regions, co-localize with regulatory elements, and can modulate certain histone modifications. By analyzing known GWAS hits and searching for new associations in 1,685 deeply-phenotyped whole-genomes, we found that eSTRs are enriched in various clinically-relevant conditions. These results highlight the contribution of short tandem repeats to the genetic architecture of quantitative human traits.

3.1 Introduction

In recent years, there has been tremendous progress in identifying genetic variants that affect expression of nearby genes, termed cis expression quantitative trait loci (cis-eQTLs). Multiple studies have shown that disease-associated variants often overlap cis-eQTLs in the affected tissue [115, 116, 117]. These observations suggest that understanding the genetic architecture of the transcriptome may provide insights into the cellular-level mediators underlying complex traits [118, 119, 120]. So far, eQTL-mapping studies have mainly focused on SNPs and to a lesser extent on bi-allelic indels and CNVs as determinants of gene expression [121, 122, 44, 45, 80]. However, these variants do not account for all of the heritability of gene expression attributable to cis-regulatory elements as measured by twin studies, leaving on average about 20-30% unexplained [122, 123]. It has been speculated that such heritability gaps could indicate the involvement of repetitive elements that are not well tagged by common SNPs [110, 111].

To augment the repertoire of eQTL classes, we focused on Short Tandem Repeats (STRs), one of the most polymorphic and abundant types of repetitive elements in the human genome [62, 66]. These loci consist of periodic DNA motifs of 2-6bp spanning a median length of around 25bp. There are about 700,000 STR loci covering almost 1% of the human genome. Their repetitive structure induces DNA-polymerase slippage events that add or delete repeat units, creating mutation rates that are orders of magnitude higher than those of most other variant types [62, 23]. Over 40 Mendelian disorders, such as Huntington's Disease, are attributed to STR mutations, most of which are caused by large expansions of trinucleotide coding repeats [28].

Several properties of STRs suggest they may play a regulatory role. In vitro studies have shown that STR variations can modulate the binding of transcription factors [41, 124], change the distance between promoter elements [125, 126], alter splicing efficiency [42, 127], and induce irregular DNA structures that may modulate transcription [128]. In vivo experiments have reported specific examples of STR variations that control gene expression across a wide range of taxa, including Haemophilus influenza [129], Saccharomyces cerevisiae [38], Arabidopsis thaliana [40], and vole [130]. Recent studies reported that dinucleotide repeats are a hallmark of enhancers in Drosophila and are enriched in predicted enhancers in humans [131]. Human promoters also disproportionately harbor STRs [132] and the presence of STRs in promoters or transcribed regions greatly increases the divergence of gene expression profiles across great apes [133], suggesting that STRs play a key role in the evolution of expression. Several candidate-

gene studies in human indeed reported that STR variations modulate gene expression [41, 134, 93, 135, 94, 136] and alternative splicing [42, 109, 34]. In one example, a recent study found that the underlying mechanism behind a GWAS signal for Ewing Sarcoma is a sequence variant in an AAGG repeat that increases the binding of the EWSR1-FLI1 oncoprotein resulting in EGF2 overexpression [43]. Despite these accumulating lines of evidence, there has been no systematic evaluation of the contribution of STRs to gene expression in humans.

To this end, we conducted a genome-wide analysis of STRs that affect expression of nearby genes, termed expression STRs (eSTRs), in lymphoblastoid cell lines (LCLs), a central ex-vivo model for eQTL studies. Next, we used a multitude of statistical genetic and functional genomics analyses to show that hundreds of these eSTRs are predicted to be functional. Finally, we tested the involvement of eSTRs in clinically relevant phenotypes.

3.2 Results

3.2.1 Initial genome-wide discovery of eSTRs

The initial genome-wide discovery of potential eSTRs relied on finding associations between STR length and expression of nearby genes. We focused on 311 European individuals whose LCL expression profiles were measured using RNA-sequencing by the gEUVADIS [44] project and whose whole genomes were sequenced by the 1000 Genomes Project [137]. The STR genotypes were obtained in our previous study [138] in which we created a catalog of STR variation as part of the 1000 Genomes Project using lobSTR, a specialized algorithm for profiling STR variations from high throughput sequencing data [27]. Briefly, lobSTR identifies reads with repetitive sequences that are flanked by non-repetitive segments. It then aligns the non-repetitive regions to the genome using the STR motif to narrow the search, thereby overcoming the gapped alignment problem and conferring alignment specificity. Finally, lobSTR aggregates aligned reads and employs a model of STR-specific sequencing errors to report the maximum likelihood genotype at each locus. lobSTR recovered most (r^2=0.71) of the variation in STR locus lengths in the 1000 Genomes datasets based on large-scale validation using 5,000 STR genotype calls obtained by capillary electrophoresis, the gold standard for STR genotyping [138]. The majority of genotype errors were from dropout of one allele at heterozygote sites due to low sequencing coverage. We simulated the performance of STR associations using lobSTR calls compared to the capillary calls. This process showed that STR genotype errors reduce the power to detect

eSTRs by 30-50% but importantly do not create spurious associations (**Supplementary Note 3.7** and **Supplementary Figure 3.9**).

To detect eSTR associations, we regressed gene expression on STR dosage, defined as the sum of the two STR allele lengths in each individual. We opted to use this measure based on previous findings that reported a linear trend between STR length and gene expression [41, 93, 94] or disease phenotypes [139, 140]. As covariates, we included sex, population structure, and other technical parameters (**Figure 3-1a** and **Supplementary Note 3.7**). We employed this process on 15,000 coding genes whose expression profiles were detected in the RNA-sequencing data. For each gene, we considered all polymorphic STR variations that passed our quality criteria (**Methods 3.6**) and were within 100kb of the transcription start and end sites of the gene transcripts as annotated by Ensembl [141]. On average, 13 STR loci were tested for each gene (**Supplementary Figure 3.9**), yielding a total of 190,016 STR×gene tests.

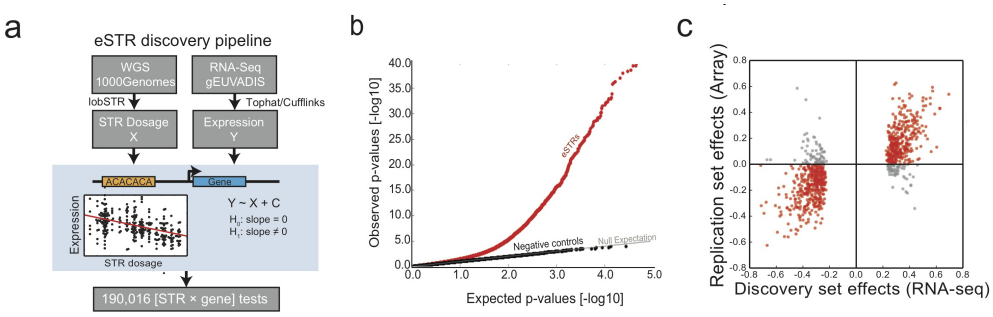

Figure 3-1: **eSTR discovery and replication. (a)** eSTR discovery pipeline. An association test using linear regression was performed between STR dosage and expression level for every STR within 100kb of a gene **(b)** Quantile-quantile plot showing results of association tests. The gray line gives the expected p-value distribution under the null hypothesis of no association. Black dots give p-values for permuted controls. Red dots give the results of the observed association tests **(c)** Comparison of eSTR effect sizes as Pearson correlations in the discovery dataset vs. the replication dataset. Red points denote eSTRs whose directions of effect were concordant in both datasets and gray points denote eSTRs with discordant directions.

Our analysis identified 2,060 unique protein-coding genes with a significant eSTR (gene level FDR $\leq$5%) (**Figure 3-1b** and **Supplementary Data Set 1** (see *Nature Genetics* website)). The majority of these were di- and tetra-nucleotide STRs (**Supplementary Tables 3.8** and **3.8**). Only 13 eSTRs fall in coding exons, but eSTRs were nonetheless strongly enriched in 5'UTRs ($p = 1.0 \times 10^{-8}$), 3'UTRs ($p = 1.7 \times 10^{-9}$) and regions near genes ($p < 10^{-28}$) compared

to all STRs analyzed (**Supplementary Table 3.8**). Overall, there was no bias in direction of effect (**Supplementary Table 3.8**). We also repeated the association tests with two negative control conditions by regressing expression on (i) STR dosages permuted between samples and (ii) STR dosages from randomly chosen unlinked loci (**Figure 3-1b** and **Supplementary Figure 3.9**). Both negative controls produced uniform p-value distributions expected under the null hypothesis. This provides support for the absence of spurious associations due to inflation of the test statistic or the presence of uncorrected population structure. To assess the effect of low sequencing coverage on our results, we generated high coverage targeted sequencing of 2,472 promoter STRs and repeated the eSTR analysis (**Methods 3.6**). We found that association results were largely reproducible across datasets, with 80% of tested eSTRs showing the same direction of effect ($p = 9.9 \times 10^{-12}$; $n = 126$) (**Supplementary Note 3.7** and **Supplementary Figure 3.9**). Three previous studies described candidate gene studies of expression STRs and involved STRs that were tested in our framework [41, 94, 93]. Our genome-wide approach was able to replicate the association between *PIG3* and the pentanucleotide STR in the 5'UTR of the gene and showed the same direction of effect. However, the other two candidate genes did not meet the multiple hypothesis p-value threshold (**Supplementary Table 3.8**).

The initial discovery set of eSTRs was largely reproducible in an independent set of individuals using an orthogonal expression assay technology. We obtained an additional set of over 200 individuals whose genomes were also sequenced as part of the 1000 Genomes Project and whose LCL expression profiles were measured by Illumina expression array [142]. These individuals belong to cohorts with African, Asian, European, and Mexican ancestry, enabling testing of the associations in a largely distinct set of populations. The Illumina expression array allowed us to test 882 eSTRs out of the 2,060 identified above. The association signals of 734 of the 882 (83%) tested eSTRs showed the same direction of effect in both datasets (sign test $p = 2.7 \times 10^{-94}$) and the effect sizes were strongly correlated ($R = 0.73$, $p = 1.4 \times 10^{-149}$) (**Figure 3-1c**), despite only moderate reproducibility of expression profiles across platforms (**Supplementary Note 3.7** and **Supplementary Figure 3.9**). For comparison, only 54% of non-eSTRs showed the same direction of effect, close to the expected value of 50% for null associations. Overall, these results show that eSTR association signals are robust and reproducible across populations and expression assay technologies.

3.2.2 Partitioning the contribution of eSTR and nearby variants

An important question is whether eSTR association signals stem from causal STR loci or are merely due to tagging SNPs or other variants in linkage disequilibrium (LD). Previous results reported that the average STR-SNP LD is approximately half of the traditional SNP-SNP LD [138, 105, 143] but there are known examples of STRs tagging GWAS SNPs [144].

To address this question, we partitioned the relative contributions of eSTRs versus all common (MAF≥1%) bi-allelic SNPs, indels, and structural variants (SV) in the cis region of each gene using a linear mixed model (LMM) (**Figure 3-2a**). Multiple studies have used this approach to measure the total contribution of common variants to the heritability of quantitative traits and to partition the contribution of different classes of variants [145, 146]. Taking a similar approach, we included two types of effects for each gene: a random effect (h^2_b) that captures all common bi-allelic loci detected within 100kb of the gene and a fixed effect (h^2_{STR}) that captures the lead STR. To test whether other causal variants in the local region could inflate the estimate of the STR contribution, we simulated gene expression with one or two causal SNP eQTLs per gene while preserving the local haplotype structure. In this negative control scenario, the LMM correctly reported a median $(h^2_{STR})/(h^2_{cis}) \approx 0$ across all conditions (**Supplementary Note 3.7** and **Supplementary Figures 3.9-3.9**), where $h^2_{cis} = h^2_b + h^2_{STR}$. This suggests that other causal variants in LD do not inflate the estimate of the relative contribution of STRs. However, simulations based on capillary electrophoresis data suggest that the variance explained by STRs is downwardly biased in the presence of genotyping errors (**Supplementary Note 3.7** and **Supplementary Figure 3.9**), suggesting that the reported h^2_{STR} is likely to be conservative.

The LMM results showed that eSTRs contribute about 12% of the genetic variance attributed to common *cis* polymorphisms. For genes with a significant eSTR, the median h^2_{STR} was 1.80%, whereas the median h^2_b was 12.0% (**Figure 3-2b**), with a median ratio of (h^2_{STR})âĄĎ(h^2_{cis}) of 12.3% ($CI_{95\%}$ 11.1%-14.2%; $n = 1,928$) (**Supplementary Table 3.8**). We repeated the same analysis for genes with at least moderate ($\geq 5\%$) *cis*-heritability (**Methods 3.6**) regardless of the presence of a significant eSTR in the discovery set. The motivation for this analysis was to avoid potential winner's curse [147] and to obtain a transcriptome-wide perspective on the role of STRs in gene expression (**Figure 3-2c**). In this set of genes, eSTRs contribute about 13% ($CI_{95\%}$ 12.2%-13.5%; $n = 6,272$) of the genetic variance attributed to *cis* common polymorphisms. The median h^2_{STR} was 1.45% of the total expression variance, whereas the median h^2_b was 9.10% (**Supplementary Table 3.8**). Repeating the analysis while considering STRs as a random

effect showed highly similar results (**Supplementary Note 3.7**, **Supplementary Table 3.8**, and **Supplementary Figure 3.9**). Taken together, this analysis shows that STR variations explain a sizeable component of gene expression variation after controlling for all variants that are well tagged by common bi-allelic markers in the *cis* region.

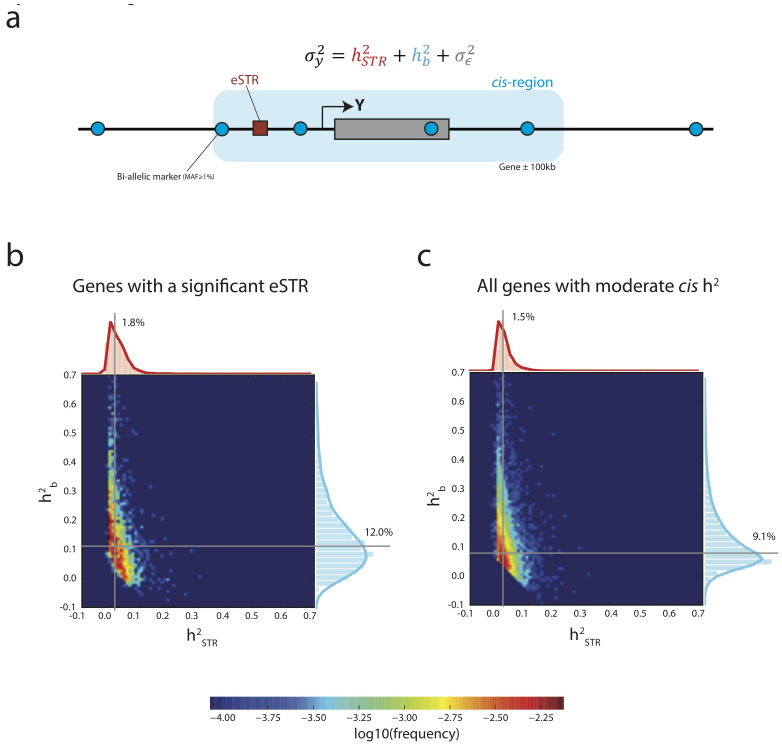

Figure 3-2: **Variance partitioning using linear mixed models (a)** The normalized variance of the expression of gene Y was modeled as the contribution of the best eSTR and all common bi-allelic markers in the cis region (±100kb from the gene boundaries) **(b-c)** Heatmaps show the joint distributions of variance explained by eSTRs and by the cis region. Gray lines denote the median variance explained **(b)** Variance partitioning across genes with a significant eSTR in the discovery set and **(c)** Variance partitioning across genes with moderate cis heritability.

3.2.3 The effect of eSTRs in the context of individual SNP eQTLs

To further assess the contribution of eSTRs in the context of other variants, we also inspected the relationship between eSTRs and individual cis-SNP eQTLs (eSNPs). We performed a traditional eQTL analysis using the whole genome sequencing data for 311 individuals that were part of

the discovery set to identify common eSNPs [minor allele frequency (MAF) $\geq$5%] within 100kb of each gene. This process identified 4,290 genes with an eSNP (gene-level FDR$\leq$5%). We then re-analyzed the eSTR association signals while conditioning on the genotype of the most significant eSNP (**Figure 3-3a**). For each eSTR, we ascertained the subset of individuals that were homozygous for the major allele of the lead eSNP in the region. If the eSTR simply tags this eSNP, its conditioned effect should be randomly distributed compared to the unconditioned effect. Alternatively, if the eSTR is causal, the direction of the conditioned effect should match that of the original effect. We conducted this analysis for eSTR loci with at least 25 individuals homozygous for the lead eSNP and for which these individuals had at least two unique STR genotypes (1,856 loci). After conditioning on the lead eSNP, the direction of effect for 1,395 loci (75%) was identical to that in the original analysis (sign test $p < 4.2 \times 10^{-109}$) and the effect sizes were significantly correlated ($R = 0.52$; $p = 3.2 \times 10^{-130}$) (**Figure 3-3b**). This further supports the additional role of eSTRs beyond traditional cis-eQTLs.

We also found that hundreds of eSTRs in the discovery set provide additional explanatory value for gene expression beyond the lead eSNP. ANOVA model comparison showed that for 23% of the cases, a model with an eSTR significantly improved the explained variance of gene expression over considering only the lead eSNP (FDR<5%) (**Figure 3-3c-e** and **Methods 3.6**). Combined with the 183 genes with an eSTR but no significant eSNP, these results show that at least 30% of the eSTRs identified by our initial scan cannot be fully attributed to tagging of the lead eSNP. Given the reduced quality of STR compared to SNP genotypes, this analysis is likely to underestimate the true contribution of STRs. Nonetheless, our results show concrete examples for hundreds of associations in which the eSTR increases the variance explained by the lead eSNP.

3.2.4 Integrative genomic evidence for a functional role of eSTRs

To provide further evidence of their regulatory role, we analyzed eSTRs in the context of functional genomics data. First, we assessed the potential functionality of STR regions by measuring signatures of purifying selection, since previous studies reported that putatively causal eSNPs are slightly enriched in conserved regions [148]. We inspected the sequence conservation [149] across 46 vertebrates in the sequence upstream and downstream of the eSTRs in our discovery dataset (**Figure 3-4a**). To tune the null expectation, we matched each tested eSTR to a random STR that did not reach significance in the association analysis but had a similar distance

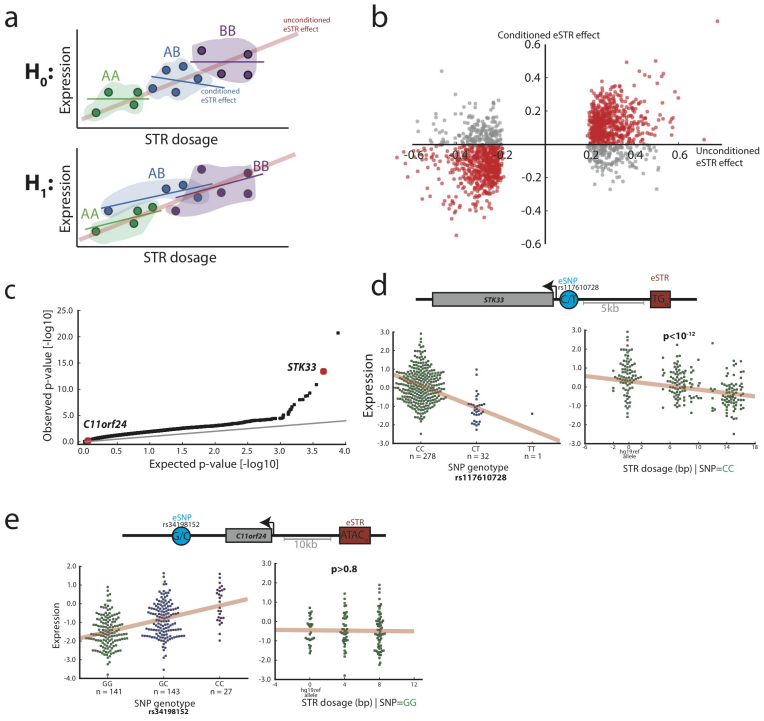

Figure 3-3: **eSTR associations in the context of eSNPs (a)** Schematic of the eSTR effect versus the effect conditioned on the lead eSNP genotype. Under the null expectation, the original association (red line) comes from mere tagging of eSNPs. Thus, the eSTR effect disappears when restricting to a group of individuals (dots) with the same eSNP genotype (colored patches). Under the alternative hypothesis, the effect is concordant between the original and conditioned associations **(b)** The original eSTR effect versus the conditioned eSTR effect. Red points denote eSTRs whose direction of effect was concordant in both datasets and gray points denote eSTRs with discordant directions **(c)** Quantile-quantile plot of p-values from ANOVA testing of the explanatory value of eSTRs beyond that of eSNPs **(d)** *STK33* is an example of a gene for which the eSTR (red rectangle) has a strong explanatory value beyond the lead eSNP (blue circle) based on ANOVA. When conditioning on individuals that are homozygous for the "C" eSNP allele (bottom left, green dots), the STR dosage still shows a significant effect (bottom right) **(e)** *C11orf24* is an example of a gene for which the eSTR was part of the discovery set but did not pass the ANOVA threshold. After conditioning on individuals that are homozygous for the "G" eSNP allele (bottom left, green dots), the STR effect is lost (bottom right).

to the nearest transcription start site (TSS). The average conservation level of a $\pm$500bp window around eSTRs was slightly but significantly higher ($p<0.03$) compared to control STRs.

Tightening the window size to shorter stretches of ±50bp showed a more significant contrast in the conservation scores of the eSTRs versus the control STRs ($p<0.01$) (**Figure 3-4a** inset), indicating that the excess in conservation comes from the vicinity of the eSTR loci. Taken together, these results show that eSTRs discovered by our association pipeline reside in regions exposed to relatively higher purifying selection, further suggesting a functional role.

eSTRs substantially co-localize with functional elements. They show the strongest enrichment closest to transcription start sites (**Figure 3-4b**) and to a lesser extent in or near predicted enhancers (**Supplementary Figure 3.9**). We also inspected the co-localization of eSTRs with histone modifications as annotated by the Encode Consortium [121] in LCLs. eSTRs were strongly enriched in peaks of histone modifications associated with regulatory regions (H3K4me3, H3K27ac, H3K9ac) and transcribed regions (H3K36me3) and were depleted in repressed regions (H3K27me3) (**Figure 3-4b**). To test the significance of these signals, we constructed a null distribution for each histone modification by measuring the co-localization of eSTRs with randomly shifted histone peaks similar to the procedure used by Trynka et al [150]. This null distribution controls for the co-occurrence of eSTRs and histone peaks due to their proximity to other causal variants. We found eSTR/histone co-localizations were significant (weakest $p<0.01$) after the peak shifting procedure, suggesting that these results stem from the eSTRs themselves (**Supplementary Table 3.8**). We also performed a peak-shifting analysis using ChromHMM annotations [151] (**Figure 3-4c**) which indicated that eSTRs are most strongly enriched in weak-promoters ($p<0.002$) and weak-enhancers ($p<0.004$). Again, this analysis shows overlap of eSTRs with elements that are predicted to regulate gene expression.

We also found that eSTR length variations are more likely to modulate the presence of certain histone marks (**Methods 3.6** and **Supplementary Figure 3.9**). We introduced different eSTR alleles to GERV [152], a machine learning approach that examines the effect of DNA sequence on histone marks. This process found that eSTRs have significantly greater effects than control STRs on predicted regulatory regions (H3K4me3 $p=0.00109$, DNAseI hypersensitivity $p=0.00045$, H3K9ac $p=0.00462$) and transcribed regions (H3K36me3 $p=0.01336$). These results are consistent with the analysis of chromatin modifications above. Importantly, since the input material for this analysis is solely STR variations that are independent of any linked variants, these results provide an orthogonal piece of evidence for the functionality of eSTRs and suggest histone mark modulation as a potential mechanism.

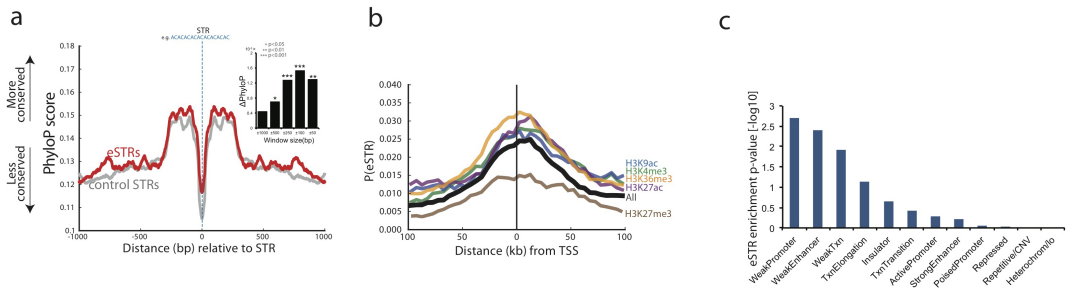

Figure 3-4: **Conservation and epigenetic analysis of eSTR loci (a)** Median PhyloP conservation score as a function of distance from the STR. Red: eSTR loci, gray: matched control STRs. Inset: the difference in the PhyloP conservation score between eSTRs and matched control STRs as a function of window size around the STR. **(b)** The probability that an STR scores as an eSTR in the discovery set as a function of distance from the transcription start site (TSS). eSTRs show clustering around the TSS (black line). Conditioning on the presence of a histone mark (colored lines) significantly modulated the probability that an STR is an eSTR **(c)** The enrichment of eSTRs in different chromatin states.

3.2.5 The potential role of eSTRs in human conditions

Encouraged by the evidence for the regulatory role of eSTRs, we wondered about their potential involvement in clinically-relevant conditions. First, we tested whether genes implicated by previous GWAS scans listed in the NHGRI GWAS catalog [4] are enriched for eSTR genes. We focused on seven complex disorders: rheumatoid arthritis, Crohn's disease, type 1 diabetes, type 2 diabetes, blood pressure, bi-polar disorder, and coronary artery disease. The first three conditions have a strong autoimmune component, rendering them more relevant to the LCL data used for eSTR discovery. To create a proper null, we compared the overlap of eSTR genes to randomly chosen sets of genes matched to the tested GWAS genes on both gene expression level in LCLs and on cis heritability.

We found that GWAS genes for Crohn's disease are significantly ($p<0.001$) enriched for eSTR hits (**Figure 3-5a** and **Supplementary Figure 3.9**). Moderate enrichment for eSTRs ($p=0.074$) was found in GWAS genes for rheumatoid arthritis, consistent with the known role of immune function in these traits. Enrichments were 2-3 times higher for autoimmune diseases than for the other conditions (average overlap: 6%). Interestingly, for seven overlapping genes, the eSTRs explained more variance in gene expression than the lead eSNP of the gene. Furthermore, for

close to thirty genes, a joint model of the lead eSTR and eSNP explained significantly more variance in gene expression than the eSNP alone, raising the possibility of an etiological role.

Next, we performed an association study using eSTRs to further test the hypothesis that eSTRs underlie clinically relevant phenotypes. For this, we turned to ∼1,700 unrelated individuals that were sequenced to medium coverage (6x) with 100bp paired-end reads using Illumina as part of the TwinsUK cohort of the UK10K project [3] and were phenotyped for a wide array of quantitative traits, primarily blood metabolites and anthropometric traits. While most of these conditions are not directly related to the immune system, we hypothesized that similar to other eQTLs [117], some of the discovered eSTRs are shared across tissues and could play a role in additional tissues. After genotyping STRs with lobSTR, we tested for association between eSTRs and each of the 38 reported phenotypes, while controlling for sex, age, and population structure. To enrich for STR loci that are likely to be causal for gene expression variation, we restricted analysis to eSTRs that significantly improved the explained variance of gene expression over a model with the lead eSNP alone. In total, we obtained 499 eSTRs after applying this condition and excluding eSTRs that were genotyped in <1000 individuals.

We identified 12 significant associations (FDR per phenotype<10%) between eSTRs and the clinical phenotypes in the TwinsUK data (**Figure 3-5b** and **Supplementary Table 3.8**). Only one association overlapped a known GWAS hit: an AAAC repeat on 4p16 was associated with decreased expression of *SLC2A9* and increased uric acid in serum samples of the TwinsUK, which matches previous studies with SNPs [153, 154, 155, 156]. The other 11 associations involved changes in blood metabolites such as albumin and C-reactive protein and physical traits such as diastolic blood pressure and FEV1 lung function and have yet to be described before in GWAS catalogs, suggesting novel loci. We caution that full validation of each of these associations will require replication in additional cohorts. Nonetheless, as we were mainly interested in the overall trend for eSTRs, we repeated the association of the 38 phenotypes in the TwinsUK cohort with a similar number of random STR loci matched on distance to transcription start sites, repeat motif, and number of genotyped samples. One hundred rounds of bootstrapping showed that eSTRs produced significantly more associations than the matched STR controls (mean for controls: 6.8 associations at FDR<10%, z-test, $p < 1.8 \times 10^{-16}$). Repeating this test with a more stringent FDR of 5% revealed a similar picture: the eSTRs produced 6 associations passing this threshold (Supplementary Table 9), significantly more that the matched STR controls (mean for controls: 3.2 associations at FDR<5%, $p < 1.1 \times 10^{-5}$). Taken together, our results show that eSTR signals are enriched in clinical phenotypes both in known and potentially novel GWAS hits.

These results could inform future efforts for disease mapping studies.

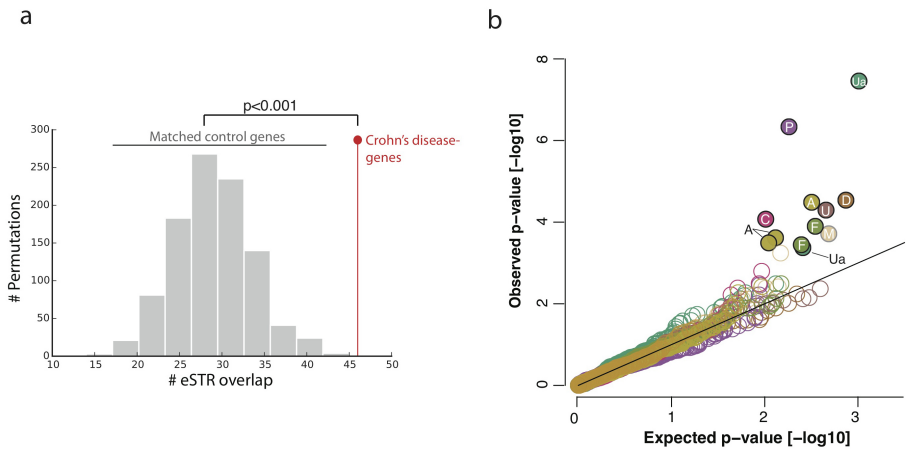

Figure 3-5: **Association of eSTRs with clinical phenotypes (a)** The overlap between eSTRs and Crohn's disease GWAS genes (red) versus random subsets of genes (gray) matched on expression and heritability profiles in LCLs **(b)** Quantile-quantile plots of eSTR associations in the TwinsUK data. Only traits with significant (FDR<0.1) associations are plotted. Closed circles: significant, open circles: non-significant. A: albumin; C: C-reactive protein; D: diastolic blood pressure, F: FVC, M: mean corpuscular volume, P: phosphate, U: Urea, Ua: Uric acid.

3.3 Discussion

Repetitive elements have often been considered as neutral with no phenotypic consequences [66]. This coupled with the technical difficulties in analyzing these regions has led large-scale genetic studies to largely overlook the putative contribution of repeats to human phenotypes. Our study focused on short tandem repeats, one of the most polymorphic classes of loci that comprise 1% of the human genome. Despite being less abundant than SNPs, previous studies have shown that STRs are enriched in promoters and enhancers, where they frequently induce multiple base-pair variations, increasing the prior expectation of their ability to explain gene expression variation. Following these observations, we conducted a genome-wide scan for the contribution of STRs to gene expression. Our scan identified over 2,000 potential eSTRs and found that eSTRs contribute on average about 10-15% of the cis-heritability of gene expression attributed to common (MAF$\geq$1%) polymorphisms. Functional genomics analyses provided further support

for the predicted causal role of eSTRs. Finally, we found that eSTRs are enriched in clinically relevant phenotypes.

We hypothesize that there are more eSTRs to find in the genome as our analysis had several technical limitations. First, the higher genotyping error rates for STRs compared to SNPs limited our power to detect eSTRs and likely downwardly biased their estimated contribution in the LMM and ANOVA analyses. In addition, about 10% of STR loci in the genome could not be analyzed because they are too long to be spanned by current sequencing read lengths[138]. Second, based on previous findings in humans [41, 93, 94], our association tests focused on a linear relationship between STR length and gene expression. However, experimental work in yeast reported that certain loci exhibit non-linear relationships between STR length and expression [38], which are unlikely to be captured in our current analysis. Finally, our association pipeline takes into account only the length polymorphisms of STRs and cannot distinguish the effect of sequence variations inside STR alleles with identical lengths (dubbed homoplastic alleles [46]). Addressing these technical complexities would likely require phased STR haplotypes and longer sequence reads that are currently unavailable for large sample sizes. We envision that recent advancements in sequencing technologies [157] will further expand the catalog of eSTRs.

Despite these technical limitations, our findings show that repetitive elements in the human genome extensively contribute to expression variation and are enriched in clinically relevant phenotypes. Our results are consistent with a recent study that reported that haplotypes of common SNPs, which capture genetic variants poorly tagged by current genotype panels, can explain substantially more heritability than common SNPs alone [158]. We anticipate that integrating the analysis of repetitive elements, specifically STR variations, will explain additional heritability and will lead to the discovery of new genetic variants relevant to human conditions.

3.4 Acknowledgements

M.G. was supported by the National Defense Science and Engineering Graduate Fellowship. Y.E. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. This study was supported by a gift from Andria and Paul Heafy (Y.E), NIJ grant 2014-DN-BX-K089 (Y.E, T.W), and NIH grants 1U01HG007037 (H.Z), R01MH084703(J.P), R01HG006399 (A.L.P), HG006696 (A.J.S), DA033660 (A.J.S), and MH097018 (A.J.S), and a research grant 6-FY13-92 from the March of Dimes Foundation (A.J.S). We thank Tuuli Lappalainen, Alon

Goren, Tatsu Hashimoto, and Dina Zielinksi for useful comments and discussions.

3.5 Author contributions

M.G. and Y.E. conceived the study. M.G., T.W., H.Z., B.M., and Y.E. performed analyses. A.G. performed experimental work to generate high coverage sequencing data for promoter STRs. S.G., M.J.D., A.L.P., and J.K.P. provided statistical input. A.J.S. contributed data and analyses. M.G., T.W., and Y.E. authored the manuscript.

3.6 Methods

3.6.1 Genotype datasets

lobSTR genotypes were generated for the phase 1 individuals from the 1000 Genomes Project as described in [138]. Variants from the 1000 Genomes Project phase 1 release were downloaded in VCF format from the project website. HapMap genotypes were used to correct association tests for population structure. Genotypes for 1.3 million SNPs were downloaded for draft release 3 from the HapMap Consortium. SNPs were converted to hg19 coordinates using the liftOver tool and filtered using Plink [159] to contain only the individuals for which both expression array data and STR calls were available. Throughout this manuscript, all coordinates and genomic data are referenced according to hg19.

3.6.2 Targeted sequencing of promoter region STRs

We used a previously published method using capture and high-throughput sequencing [160] to sequence 2,472 STRs located in gene promoters (TSS $\pm$ 1kb) in 120 HapMap individuals of European (58 CEU individuals) and African (62 YRI individuals) ancestry. Briefly, the method uses a custom Nimblegen EZ Capture system to enrich the genomic sequence flanking, and sometimes including, the target STRs to be genotyped prior to sequencing using an Illumina Hiseq2000 instrument. We multiplexed 24 individuals per sequencing lane and utilized 100bp single-end reads. We used lobSTR version 3.0.3 to genotype STRs in these samples.

3.6.3 Expression datasets

RNA-sequencing datasets from 311 HapMap lymphoblastoid cell lines for which STR and SNP genotypes were also available were obtained from the gEUVADIS Consortium. Raw FASTQ files containing paired end 100bp Illumina reads were downloaded from EBI. The hg19 Ensembl transcriptome annotation was downloaded as a GTF file from the UCSC Genome Browser [161, 162] ensGene table. The RNA-sequencing reads were mapped to the Ensembl transcriptome using Tophat v2.0.7 [163] with default parameters. Gene expression levels were quantified using Cufflinks v2.0.2 [164] with default parameters and supplied with the GTF file for the Ensembl reference version 71. Genes with median FPKM of 0 were removed, leaving 23,803 genes. We restricted analysis to protein coding genes, giving 15,304 unique Ensembl genes. Expression values were quantile-normalized to a standard normal distribution for each gene.

The replication set consisted of Illumina Human-6 v2 Expression BeadChip data from 730 HapMap lymphoblastoid cell lines from the EBI website. These datasets contain two replicates each for 730 unrelated individuals from 8 HapMap populations (YRI, CEU, CHB, JPT, GIH, MEX, MKK, LWK) and were generated as described by Stranger et al. [45]. Background corrected and summarized probeset intensities (by Illumina software) contained values for 7,655 probes. Additionally, probes containing common SNPs were removed [165]. Only probes with a one-to-one correspondence with Ensembl gene identifiers were retained. We removed probes with low concordance across replicates (Spearman correlation ≤ 0.5). In total we obtained 5,388 probes for downstream analysis.

Each probe was quantile-normalized to a standard normal distribution across all individuals separately for each replicate and then averaged across replicates. These values were quantile-normalized to a standard normal distribution for each probe.

3.6.4 eQTL association testing

Expression values were adjusted for individual sex, individual population membership, gene expression heterogeneity, and population structure (**Supplementary Note 3.7**). Adjusted expression values were used as input to the eSTR analysis. To restrict to STR loci with high quality calls, we filtered the call set to contain only loci where at least 50 of the 311 samples had a genotype call. To avoid outlier genotypes that could skew the association analysis, we removed any genotypes seen less than three times. If only a single genotype was seen more than three

times, the locus was discarded. To increase our power, we further restricted analysis to the most polymorphic loci with heterozygosity of at least 0.3. This left 80,980 STRs within 100kb of a gene expressed in our LCL dataset.

A linear model was used to test for association between normalized STR dosage and expression for each STR within 100kb of a gene. Dosage was defined as the sum of the deviations of the STR allele lengths from the hg19 reference. For example, if the hg19 reference for an STR is 20bp and the two alleles called are 22bp and 16bp, the dosage is equal to (22-20)+(16-20) = -2bp. STR genotypes were zscore-normalized to have mean 0 and variance 1. For genes with multiple transcripts, we defined the transcribed region as the maximal region spanned by the union of all transcripts. The linear model for each gene is given by:

$$\vec{y_g} = \alpha_g + \beta_{j,g}\vec{x}_j + \vec{\epsilon}_{j,g} \tag{3.1}$$

where $\vec{y_g} = (y_{g,1}, \ldots, y_{g,n})^T$ with $y_{g,i}$ the normalized covariate-corrected expression of gene g in individual i, n is the number of individuals, α_g is the mean expression level of homozygous reference individuals, $\beta_{j,g}$ is the effect of the allelic dosage of STR locus j on gene g, $\vec{x_j} = (x_{j,1}, \ldots, x_{j,n})^T$ with $x_{j,i}$ the normalized allelic dosage of STR locus j in the ith individual, and $\vec{\epsilon}_{j,g}$ is a random vector of length n whose entries are drawn from $N(0, \sigma^2_{\epsilon,j,g})$ where $\sigma^2_{\epsilon,j,g}$ is the unexplained variance after regressing locus j on gene g. The association was performed using the OLS function from the Python statsmodels package. For each comparison, we tested $H_0 : \beta_{j,g} = 0$ vs. $H_1 : \beta_{j,g} \neq 0$ using a standard t-test. We controlled for a gene-level false discovery rate (FDR) of 5% (see below).

3.6.5 Controlling for gene-level false discovery rate

We controlled for a gene-level FDR of 5%, assuming that most genes have at most a single causal eSTR. For each gene, we determined the STR association with the best p-value. This p-value was adjusted using a Bonferroni correction for the number of STRs tested per gene to give a p-value for observing a single eSTR association for each gene. Performing separate permutations for each gene was computationally infeasible, and was found to give similar results to a simple Bonferroni correction on a subset of genes. We then used this list of adjusted p-values as input to the qvalue R package to determine all genes with FDR at most 5%.

3.6.6 Partitioning heritability using linear mixed models

For each gene, we used a linear mixed model to partition heritability between the lead explanatory STR and other cis variants. We used a model of the form:

$$\vec{y}_g = \alpha_g + \beta_{j,g}\vec{x}_j + \vec{u}_g + \vec{\epsilon}_{j,g} \tag{3.2}$$

where $\vec{y}_g$, α_g, $\beta_{j,g}$, $\vec{x}_j$, and $\epsilon_{j,g}$ are as described above, $\vec{u}_g$ is a length n vector of random effects and $\vec{u}_g \sim MVN(0, \sigma^2_{u_g} K_g)$ with $\sigma^2_{u_g}$ the percent of phenotypic variance explained by cis bi-allelic variants for gene g, and K_g is a standardized $n \times n$identity by state (IBS) relatedness matrix constructed using all common bi-allelic variants (MAF$\geq$ 1%) reported by phase 1 of the 1000 Genomes Project within 100kb of gene g. This includes SNPs, indels, and several bi-allelic structural variants and is constructed as $K_g = \frac{1}{p}\sum_{i=0}^{p} \frac{1}{var(\vec{x}_i)}(\vec{x}_i - 1_n mean(\vec{x}_i))(\vec{x}_i - 1_n mean(\vec{x}_i))^T$ where p is the total number of variants considered, $\vec{x}_i$ is a length n vector of genotypes for variant i, and 1_n is a length n vector of ones. Note the mean diagonal element of K_g is equal to 1.

We used the GCTA program [166] to determine the restricted maximum likelihood estimates (REML) of $\beta_{j,g}$ and $\sigma^2_{u_g}$. To get unbiased values of $\sigma^2_{u_g}$, the --reml-no-constrain option was used.

We used the resulting estimates to determine the variance explained by the STR and the cis region. We can write the overall phenotypic variance-covariance matrix as:

$$var(\vec{y}_g) = \beta^2_{j,g} var(\vec{x}_j) + \sigma^2_{u_g} K_g + \sigma^2_{\epsilon_{j,g}} I_n \tag{3.3}$$

where $var(\vec{y}_g)$ is an $n \times n$ expression variance-covariance matrix with diagonal elements equal to 1, since expression values for each gene were normalized to have mean 0 and variance 1 and I_n is the $n \times n$ identity matrix.

This equation shows the relationship:

$$\sigma^2_p = h^2_{STR} + h^2_b + \sigma^2_\epsilon \tag{3.4}$$

where σ^2_p is the phenotypic variance, which is equal to 1, h^2_{STR} is the variance explained by the

STR, which is equal to $\beta_{j,g}^2 var(\vec{x}_j) = \beta_{j,g}^2$ since the STR genotypes were scaled to have mean 0 and variance 1, and h_b^2 is the variance explained by bi-allelic variants in the cis region. This is approximately equal to $\sigma_(u_g)^2$ since the local IBS matrix K_g has a mean diagonal value of 1.

We estimated the percent of phenotypic variance explained by STRs, $\beta_{j,g}^2$, using the unbiased estimator $\hat{h}_{STR}^2 = E[\beta_{j,g}^2] = \hat{\beta}_{j,g}^2 - SE^2$, where $\hat{\beta}_{j,g}$ is the estimate of $\beta_{j,g}$ returned by GCTA, and SE is the standard error on the estimate, using the fact that $\beta_{j,g} \sim N(\beta_{j,g}, SE)$. We estimated the percent of phenotypic variance explained by bi-allelic markers as $\hat{h}_b^2$. Note that for this analysis the STR was treated as a fixed effect. We also reran the analysis treating the STR as a random effect and found very little change in the results (**Supplementary Note 3.7**).

Results are reported for all eSTR-containing genes and for all genes with moderate total cis heritability, which we define as genes where $h_{STR}^2 + h_b^2 \geq 0.05$. We used this approach as to our knowledge there are no published results about the cis-heritability of expression of individual genes in LCLs from twin studies. We used 10,000 bootstrap samples of each distribution to generate 95% confidence intervals for the medians.

3.6.7 Comparing to the lead eSNP

We identified SNP eQTLs using SNPs with MAF $\geq$ 1% as reported by phase 1 of the 1000 Genomes Project. We used an identical pipeline to our eSTR analysis to identify SNP eQTLs after replacing the vector $\vec{x}_j$ with a vector of SNP genotypes (0, 1 or 2 reference alleles) that was z-normalized to have mean 0 and variance 1. To determine whether our eSTR signal was indeed independent of the lead SNP eQTL at each gene, we repeated association tests between STR dosages and expression levels while holding the genotype of the SNP with the most significant association to that gene constant. For this, we determined all samples at each gene that were either homozygous reference or homozygous non-reference for the lead SNP. For the SNP allele with more homozygous samples, we repeated the eSTR linear regression analysis and determined the sign and magnitude of the slope. We removed any genes for which there were less than 25 samples homozygous for the SNP genotype or for which there was no STR variation after holding the SNP constant, leaving 1,856 genes for analysis. We used a sign test to determine whether the direction of effects before and after conditioning on the lead SNP are more concordant than expected by chance.

We used model comparison to determine whether eSTRs can explain additional variation in gene expression beyond that explained by the lead eSNP for each gene. For each gene with a

significant eSTR and eSNP, we analyzed the ability of two models to explain gene expression:

$$\texttt{Model 1 (eSNP-only):} \quad \vec{y}_g = \alpha_g + \beta_{eSNP,g}\vec{x}_{eSNP,g} + \vec{\epsilon}_{j,g} \tag{3.5}$$

$$\texttt{Model 2 (joint eSNP+eSTR):} \quad \vec{y}_g = \alpha_g + \beta_{eSNP,g}\vec{x}_{eSNP,g} + \beta_{eSTR,g}\vec{x}_{eSTR,g} + \vec{\epsilon}_{j,g} \tag{3.6}$$

where α_g is the mean expression value for the reference haplotype, $\vec{y}_g$ is a vector of expression values for gene g, $\beta_{eSNP,g}$ is the effect of the eSNP on gene g, $\beta_{eSTR,g}$ is the effect of the eSTR on gene g, $\vec{x}_{eSNP,g}$ is a vector of genotypes for the lead eSNP for gene g, $\vec{x}_{eSTR,g}$ is a vector of genotypes for the best eSTR for gene g, and $\vec{\epsilon}_{j,g}$ gives the residual term. A major caveat is that the eSNP dataset has significantly more power to detect associations than the eSTR dataset due to the lower quality of the STR genotype panel (**Supplementary Note 3.7**), and this analysis is therefore likely to underestimate the true contribution of STRs to gene expression. We used ANOVA to test whether the joint model performs significantly better than the SNP-only method. We obtained the ANOVA p-value for each gene and used the qvalue package to determine the FDR.

3.6.8 Conservation analysis

Sequence conservation around STRs was determined using the PhyloP track available from the UCSC Genome Browser. To calculate the significance of the increase in conservation at eSTRs, we compared the mean PhyloP score for each eSTR to that for 1000 random sets of STRs with matched distributions of the distance to the nearest transcription start site. For each STR, we determined the mean PhyloP score for a given window size centered on the STR. The p-value given is the percentage of random sets whose mean PhyloP score was greater than the mean of the observed eSTR set.

3.6.9 Enrichment of STRs and eSTRs in predicted enhancers

H3K27ac peaks produced by the ENCODE Project [121] were used to determine predicted enhancers in GM12878. Peaks were downloaded from the UCSC Genome Browser and converted to hg19 coordinates using the liftOver tool. Any peak overlapping within 3kb of a transcription start site was removed to exclude promoter regions from the analysis.

3.6.10 Enrichment in histone modification peaks

Chromatin state and histone modification peak annotations generated by the ENCODE Consortium for GM12878 were downloaded from the UCSC Genome Browser. Because variants involved in regulating gene expression are more likely to fall near genes compared to randomly chosen variants, naive enrichment tests of eSTRs vs. randomly chosen control regions may return strong enrichments simply because of their proximity to genes. To account for this, we randomly shifted the location of eSTRs by a distance drawn from the distribution of distances between the best STR and lead SNP for each gene. We repeated this process 1,000 times. For each set of permuted eSTR locations, we generated null distributions by determining the percent of STRs overlapping each annotation. We used these null distributions to calculate empirical p-values for the enrichment of eSTRs in each annotation.

3.6.11 Effects of eSTRs on modulating regulatory elements

One potential mechanism by which eSTRs may act is by modulating epigenetic properties. The GERV (Generative Evaluation of Regulatory Variants)[152] model predicts ChIP-sequencing experiments directly from genomic sequences and optional covariates such as DNAse-seq data. We used the non-covariate version of this technique to assess the effect of STR variations on the occupancy of chromatin marks.

GERV builds on a kmer-based statistical model to predict the signal of ChIP-seq experiments from a DNA sequence context. Briefly, the model considers that each k-mer has a spatial effect on ChIP-seq read counts in a window of [-M, M-1] bp centered at the start of the k-mer. The read count at a given base is then modeled as the log-linear combination of the effects of all k-mers whose effect ranges cover that base, where k ranges from 1 to 8.

For each eSTR in our dataset, we generated sequences representing each observed allele. We filtered STRs with interruptions in the repeat motif, since the sequence for different allele lengths is ambiguous for these loci. For each mark, we used the model to predict the read count for each allele in a window of $\pm$M bp from the STR boundaries, where M was set to 1,000 for all marks except p300, for which M was set to 200. Previous findings of GERV showed that these values of M give the best correlation between predicted and real ChIP-seq signals using cross validation. For each alternate allele, we generated a score as the sum of differences in read counts from the reference allele at each position in this window. We regressed the number of

repeats for each allele on this score and took the absolute value of the slope for each locus. We repeated the analysis on a set of randomly chosen negative control loci. Control loci were chosen to match the distribution of repeat lengths and absolute signal for each mark in the reference genome. We used a Mann-Whitney rank test to compare the magnitudes of slopes between the eSTR and control sets for each mark.

3.6.12 Overlap of eSTR and GWAS genes

Aggregate results for seven common diseases (rheumatoid arthritis, Crohn's disease, type I diabetes, type 2 diabetes, blood pressure, bi-polar disorder, and coronary artery disease) were downloaded from the NHGRI GWAS catalog accessed on June 12, 2015. Relevant genes were taken from the columns "Reported Gene(s)" and "Mapped_gene". To generate a null distribution, we chose 1,000 sets of randomly selected genes matched to eSTR genes on expression in LCLs (difference in RPKM < 10) and on cis heritability (difference in variance explained by cis bi-allelic variants $< 5\%$). We compared the overlap of GWAS genes with eSTR genes vs. the 1,000 control sets to determine an empirical p-value.

3.6.13 eSTR associations with human traits

To generate STR genotypes for each of the individuals in the UK10K TwinsUK dataset, we ran lobSTR v2.0.3 on each BAM using the options fft-windowsize=16, fft-window-step=4 and bwaq=15. The resulting BAM files were analyzed using v2.0.3 of the lobSTR allelotyper using default options, resulting in STR genotypes for 1,685 individuals.

We then performed an association test between each STR and each phenotype. To control for population structure, we adjusted STR dosages and phenotypes for the top 10 ancestry principal components based on common SNPs (MAF$\geq$5%) after LD-pruning. Principal components were computed using EIGENSTRAT [167] v5.0.1. Phenotypes were further adjusted for the age at which the phenotype was measured. Association tests were performed between the adjusted dosages and the quantile-normalized adjusted phenotypes. We were able to analyze TwinsUK cohort for the following 38 phenotypes [in parentheses, the PMID reference given by TwinsUK to describe the phenotype measurement procedure]: Albumin (19209234), Alkaline phosphatase (19209234), Apolipoprotein A-I (15379757), Apolipoprotein B (15379757), Bicarbonate, Bilirubin (19209234), Body mass index, Creatinine (11017953), Diastolic blood pressure

(16249458), Heart Rate (19587794), FEV1 (17989158), FEV1/FVC ratio (17989158), FVC (17989158), Gamma-Glutamyl Transpeptidase (19209234), Glucose (19209234), High density lipoprotein (19016618), Standing height (17559308), Hemoglobin (19862010), Hip circumference (17228025), Homocysteine (18280483), C-reactive protein (21300955), Insulin (16402267), Mean corpuscular volume (19862010), Packed Cell Volume (10607722), Phosphate (12193151), Platelet count (19221038), Red blood cell count (19820697), Sodium (18179892), Systolic blood pressure (16249458), Total cholesterol (19820914), Triglycerides (15379757), Urea (18179892), Uric acid (19209234), Waist circumference (17228025), White blood cell count (19820697), Weight (17016694), and Waist to Hip ratio.

We then examined the association in the 666 eSTR loci that contained an eSTR that significantly improved the gene expression variance when combined with the lead eSNP (nominal ANOVA $p<0.05$). Out of these eSTRs, 499 were genotyped in >1,000 participants. For each phenotype, q values were calculated by adjusting the p-values using the Benjamini-Hochberg procedure. Only hits with a q-value < 0.1 were reported.

3.7 Supplementary Text

Please refer to the original publication

3.8 Supplementary Tables

Please refer to the original publication

3.9 Supplementary Figures

Please refer to the original publication

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Robust algorithms for genotyping, phasing and imputing short tandem repeats

Thomas Willems and Yaniv Erlich

Abstract: Recent advances in bioinformatics have enabled accurate methods to genotype, phase and impute single nucleotide polymorphisms. However, the development of similar approaches for other types of genetic variants has been far more limited. Short tandem repeats (STRs), highly mutable repetitive tracts of DNA, have proven particularly problematic to characterize. Despite increasing numbers of single gene and genome-wide studies linking STR variations to complex traits, genetic studies typically omit STRs due to high genotyping error rates. To overcome this limitation, we describe HipSTR, a novel haplotype-based method for genotyping and phasing STRs. Using gold standard datasets, we demonstrate that HipSTR outperforms all mainstream variant callers and is nearly an order of magnitude faster than the next best tool. We then use HipSTR to identify hundreds of replicable de novo mutations in a deeply sequenced cell line trio. Lastly, we illustrate the applicability of existing statistical phasing and imputation methods to STR loci.

4.1 Introduction

The advent of next-generation sequencing technologies has enabled the characterization of human genetic variation on an unprecedented scale. At the population-level, the 1000 Genomes Project [49] and UK10K Project [3] have uncovered the number and frequency of variants in the human genome. At the same time, family-based approaches have estimated the de novo mutation rates of single nucleotide polymorphisms, short indels and copy number variants [58, 59, 60, 61]. Motivated by these early insights, scientists are increasingly using genomics to

dissect the basis of complex traits and diseases. Through genome-wide association studies and large-scale trio-based studies, they have uncovered tens of thousands of variants with phenotypic links [4] and have shed light on the roles of common and rare variants in psychiatric disorders such as autism and schizophrenia [168, 169, 170, 171]. Recently, Genomics England announced plans to sequence 100,000 individuals as part of a national health care initiative, signaling the arrival of genomics in personalized medicine [172].

The impact of genomics in medicine is contingent upon its ability to identify causal genetic variants. While most genomic analyses focus on SNPs, short tandem repeats (STRs) represent a largely untapped type of genetic variation. Composed of a repeating 1-6 base pair motif, STRs are among the most polymorphic variants in the human genome and frequently mutate every 100-10,000 generations [56]. STR variations have been shown to have profound phenotypic consequences, as over 30 hereditary disorders such as Huntington's disease and Fragile X syndrome are caused by these markers [28]. Single-gene studies in humans have also demonstrated that STRs can modulate gene expression [41, 93, 94, 136] and regulate alternative splicing [109, 42, 34]. More recently, a genome-wide analysis estimated that STRs account for 10-15% of the cis-heritability of gene expression, underscoring their profound regulatory role [173].

Despite their putative role in phenotypic variation, STRs are typically omitted from large-scale genetic studies. Analyses of genotyping errors generated using high-coverage datasets have highlighted low complexity regions as being highly enriched in genotyping errors [174]. In addition, STRs are challenging to compare between tools because they typically contain multiple variant blocks, each of which has redundant representations. As a result, it has become common practice to mask STR regions.

Recent advances in genotyping algorithms have shown great promise in terms of reducing error rates. By using local assembly and haplotype alignment, tools such as GATK Haplotype Caller [53] and Platypus are more effective at identifying indels and are less prone to misaligned reads, resulting in higher sensitivity and specificity [54]. However, it remains unclear whether these and other mainstream tools are suitable for characterizing STRs. Advances in statistical phasing and imputation approaches have also led to increasingly refined haplotypes [175, 176, 177, 178]. However, the applicability of these approaches to STRs has received little attention. Given the widespread use of STRs in forensics, algorithms to accomplish these tasks would have profound implications in DNA identification. Furthermore, these algorithms could facilitate the incorporation of STRs into genome-wide association studies.

Here, we address each of these uncertainties using gold standard STR and high-throughout sequencing datasets. We begin by describing HipSTR, a novel STR-specific haplotype caller that accounts for prevalent sources of error in repetitive regions. By comparing HipSTR to GATK, Platypus, freebayes [179] and two other STR callers, we show that it performs optimally, offers exceptional computational tractability, and can reliably phase STR haplotypes onto SNP scaffolds. We then illustrate the added value of HipSTR by identifying hundreds of replicable de novo mutations in a well-studied trio. Finally, we demonstrate that Beagle [176], a widely used tool for phasing and imputation, can be successfully applied to STRs. In particular, STR genotypes imputed using Beagle recover a substantial fraction of the true underlying genetic variance and may be useful in association studies. Collectively, these analyses pave the way for reliably incorporating STRs into future genetic studies.

4.2 Results

4.2.1 Towards an improved STR variant caller

The structure and sequence of short tandem repeats pose unique challenges to genotyping methods. Due to their repetitive nature and wide range of allele lengths, STRs result in frequent alignment errors during read mapping. In addition, mechanisms similar to those that cause STR variation also introduce noise into PCR amplified reads around STRs [62, 180]. These stutter artifacts insert or delete copies of the repeat motif, resulting in sequences that differ in size from the original genotype and causing false positive heterozygous calls at homozygous sites.

To mitigate these and other STR-specific issues, we developed HipSTR, a haplotype-based STR caller **(Figure 4-1)**. HipSTR begins by learning the stutter model for each locus using an expectation maximization algorithm. After identifying an initial set of candidate STR alleles, it uses this stutter model and a specialized hidden Markov model (HMM) to realign reads to every candidate haplotype, mitigating the effects of alignment errors. The HMM accounts for two distinct sources of error: Illumina sequencing errors in regions flanking the STR and PCR stutter errors in the STR itself. HipSTR also uses information from phased SNP haplotypes to physically phase STRs onto SNP scaffolds. Using bases that overlap heterozygous SNPs, it computes the likelihood that each read originated from either SNP haplotype. These phasing likelihoods, in conjunction with the alignment likelihoods for each read-haplotype combination, are used to determine maximum likelihood phased genotypes.

HipSTR also leverages the unique mutational properties of STRs to improve genotyping accuracy. Because the sizes of both STR alleles and stutter artifacts occur in increments of the motif length, the HMM frequently aligns reads originating from unidentified candidate alleles as stutter artifacts. HipSTR therefore uses an iterative approach in which it progressively adds new candidate alleles based on frequently observed stutter artifacts. This enables it to detect large expansions and deletions without using local haplotype assembly approaches such as in Platypus and GATK-HC. For more details about each component of the HipSTR algorithm, please refer to the **Methods** section. HipSTR is open-source and freely available at https://github.com/tfwillems/HipSTR.

4.2.2 Benchmarking STR variant callers

In contrast to other types of insertions and deletions, ground truth datasets are readily available for STRs. To construct linkage maps of the human genome [9], highly polymorphic STRs were identified and then genotyped using capillary electrophoresis, providing accurate and challenging test cases. Here, we focus on a set of ∼600 STRs in the Marshfield panel of markers [20]. For 118 samples, both capillary genotypes [92] and high-throughout whole-genome sequencing data are available [181], where the latter was generated for 263 samples in the Simons Genome Diversity Project (SGDP) using 100 bp paired-end Illumina reads, over 30x coverage and a PCR-free library preparation protocol.

To benchmark variant callers, we sought to assess their ability to predict capillary genotypes from whole-genome sequencing data. In addition to HipSTR, we selected GATK HaplotypeCaller (GATK-HC), Platypus, freebayes, lobSTR [27] and RepeatSeq [52], as these tools are either widely used for general purpose or STR-specific variant calling. In conjunction with BWA-MEM alignments [182], we used each tool to jointly genotype the 263 SGDP samples in each STR region and compared the sum of each sample's indel sizes to the sum predicted by the capillary data **(Table 4.1)**. Under these default settings, freebayes and Platypus achieved relatively low accuracy. We therefore explored the effect of alternate command line options on accuracy **(Supplemental Table 4.2)** and reassessed each tool's performance using optimal settings, resulting in substantial gains for freebayes and Platypus and marginal gains for GATK-HC. Overall, HipSTR performed optimally and achieved over a 95% accuracy rate. GATK-HC was the second most accurate tool but required nearly 5 times as much computation time. In addition, the GATK-HC callset contained an average of 5.5 variant sites per STR region,

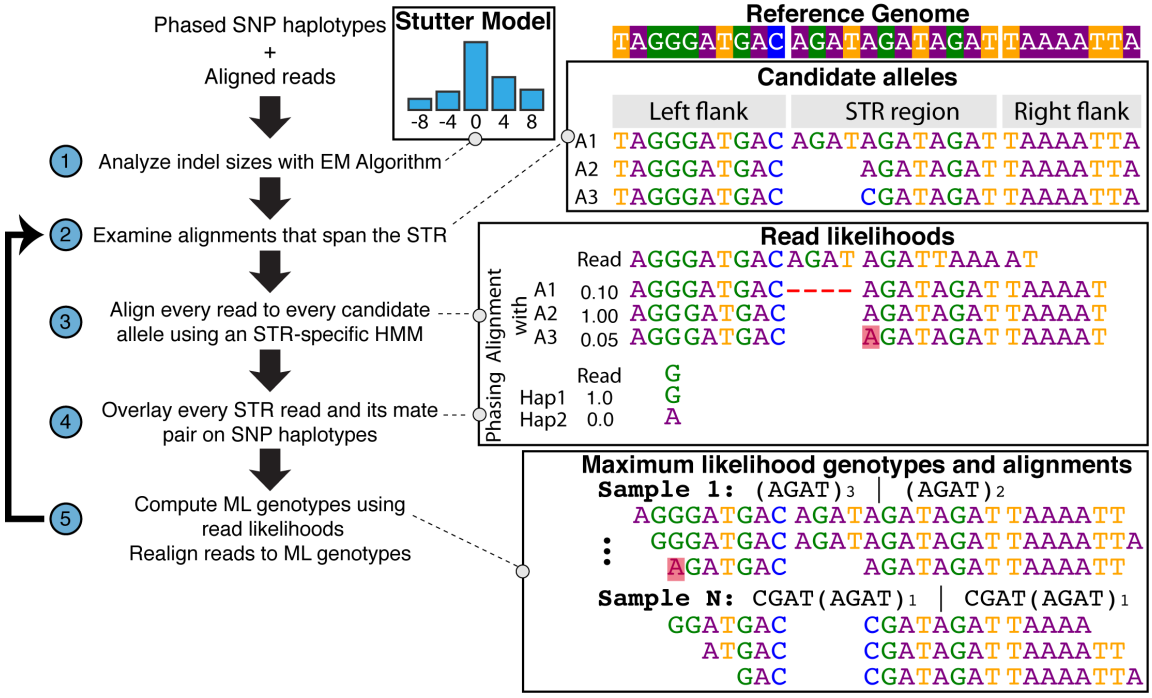

Figure 4-1: **Overview of HipSTR's algorithm**. In **step 1**, an expectation-maximization algorithm learns the PCR stutter model for the locus of interest in order to account for the frequency of these errors during genotyping. **Step 2** utilizes well-anchored alignments that span the STR to identify candidate alleles in the STR region and builds haplotypes using these alleles and the sequence upstream and downstream of the STR. In **step 3**, the PCR stutter model and an HMM are used to align every read to every candidate haplotype. **Step 4** analyzes reads that overlap heterozygous SNPs to determine the likelihood that each read came from either strand of the phased SNP haplotypes. Finally, **step 5** combines these two sets of likelihoods to determine each sample's maximum likelihood genotype. Every read is realigned relative to its sample's optimal genotype and if new reads span the STR, HipSTR returns to step 2 to identify new candidate alleles and repeat the process. This iterative procedure continues until no new candidate alleles are identified, at which point the maximum likelihood genotypes are output to a VCF file.

vastly complicating the interpretability of the calls. In contrast, HipSTR locally phased all variants within each STR region into two phased haplotypes. These locally phased haplotypes are particularly important for forensics applications and mutation rate studies as they require the individual lengths of STR alleles. Of note, lobSTR also obtained relatively high accuracy using very little computation time, but it offers no information about the sequence of the STR.

Table 4.1: Capillary-based benchmarking of STR genotyping tools

Method	Correct Calls[1]	Total Calls[1]	Accuracy (%)	Run Time (hrs)	Variants per STR[2]
HipSTR	58937	61942	95.2	1.3	1.0
GATK-HC*	59364	62966	94.3	5.8	5.5
GATK-HC	57947	62966	92.0	6.3	5.5
lobSTR	54627	61971	88.2	0.1	1.0
Platypus*	50196	62937	79.8	15.7	4.3
freebayes*	39197	57349	63.5	4.8	2.8
Platypus	39220	62347	62.9	1.9	2.9
RepeatSeq	30289	52393	57.8	0.3	1.0
freebayes	9011	61741	15.7	3.6	3.0

*Tool was run with settings optimized for accurate STR calling
[1]The reported number of calls refer to the 118 samples with capillary genotypes
[2]Average number of VCF records per STR region

Next, we sought to assess how variant filtration impacts the relative performance of these tools. Based on best practice guidelines and manual selection, we chose features for each tool that might be indicative of genotype quality **(Supplemental Table 4.3)**. We then used these features and the true capillary genotypes to train boosted regression tree classifiers to distinguish between correct and incorrect calls. Five-fold cross validation revealed that these classifiers successfully ranked the quality of each tool's genotypes, as precision generally increased with classifier confidence **(Figure 4-2)**. Except for at very low recall levels, the HipSTR classifier produced the most precise call sets and precision only noticeably declined when including the 5% least confident calls. GATK-HC's classifier generally produced the second most precise call sets, largely mirroring the performance trends we observed in the unfiltered scenario. Collectively, these comparisons suggest that HipSTR is the best performing method, both overall and on filtered subsets of the data, but that GATK-HC also provides robust STR calls.

Encouraged by HipSTR's ability to accurately determine the length of STRs, we sought to assess whether it can also reliably determine their sequence. Using 100 bp reads from the Illumina Platinum Genomes Project and 250 bp reads from an additional study **(Methods)**, we generated BWA-MEM alignments and used HipSTR to genotype each Marshfield STR in a CEPH trio (NA12891, NA12892 and NA12878). Out of 564 markers with no missing genotypes and at least one length polymorphism, 562 had genotypes consistent with Mendelian inheritance. We

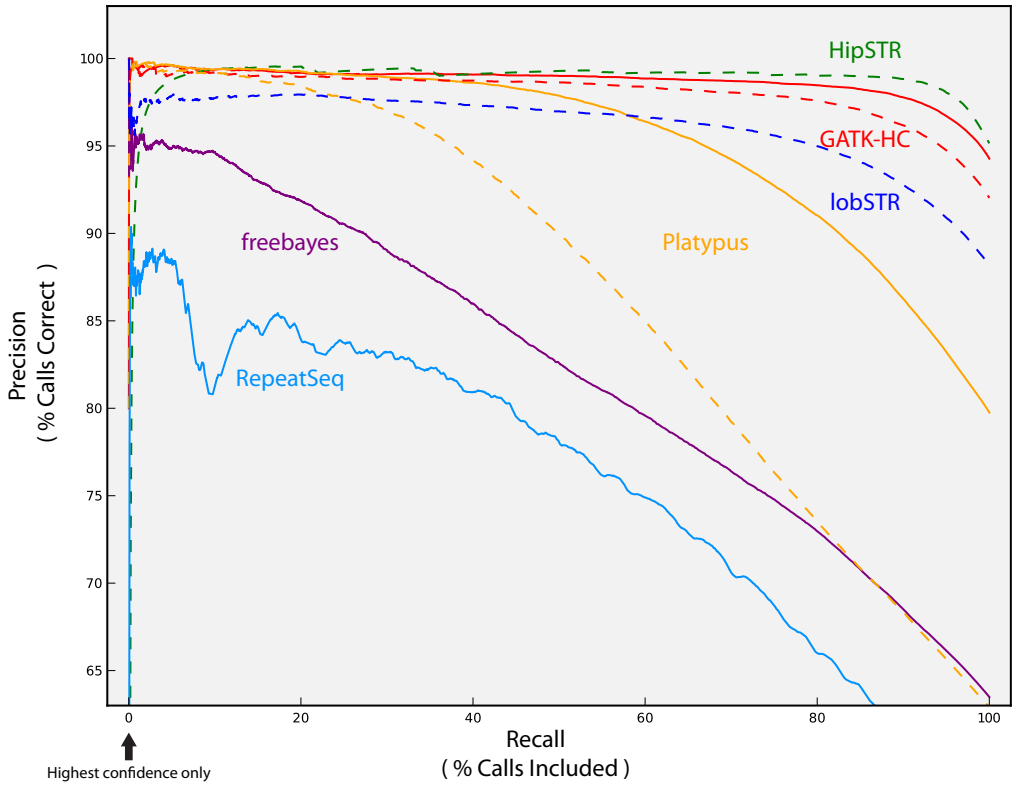

Figure 4-2: **Precision-recall curves for STR genotype classifiers**. Each variant caller's classifier can distinguish correct from incorrect calls. Dashed lines denote tools run using default settings, while solid lines denote tools run using settings optimized for STR genotyping. The curve for freebayes run in default mode is not shown due to poor performance.

manually inspected the two markers where HipSTR genotypes were incompatible with Mendelian inheritance. In both instances, one of which is outlined in **Figure 4-3**, the alignments strongly supported a de novo mutation. For 47 out of the 568 markers, HipSTR identified two or more alleles with the same length but different sequences. **Figure 4-4** depicts one such instance for a complex marker comprised of adjacent GATA and CATA repeats. The HipSTR genotype for NA12878, which is well supported by the alignments, consists of two alleles with the same length but different numbers of GATA and CATA repeats. These analyses suggest that HipSTR is able to determine the sequence of STRs in an accurate manner.

Lastly, we sought to validate HipSTR's ability to physically phase STRs onto SNP scaffolds. Using Illumina Platinum Genomes data for the CEPH trio and GATK-HC, we generated SNP

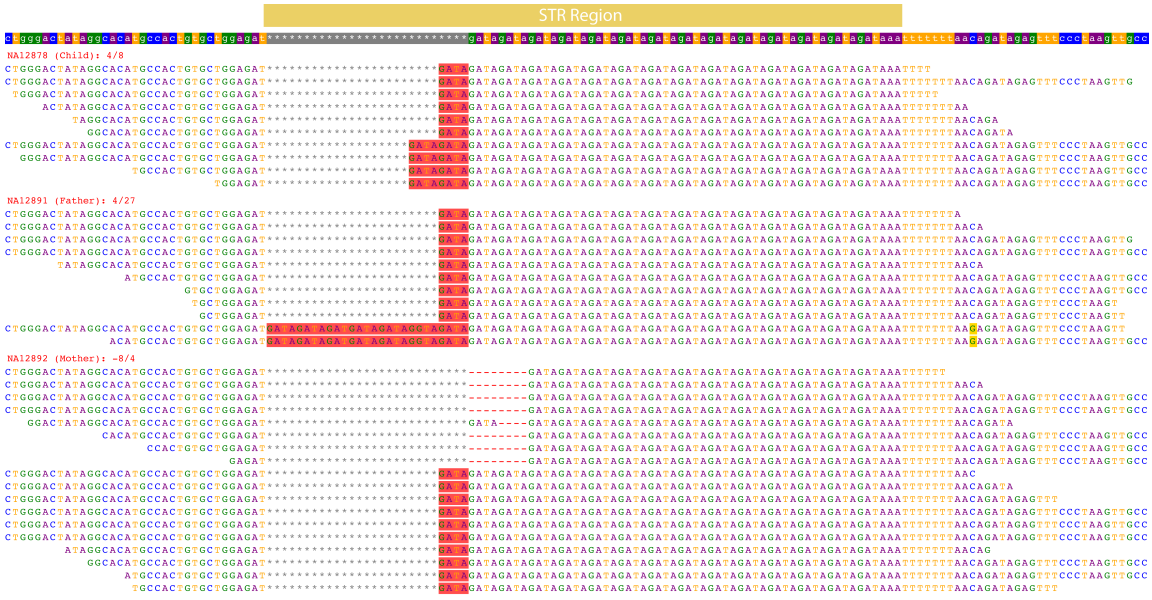

Figure 4-3: **Candidate de novo mutation at a Marshfield locus**. The child (NA12878) has an 8 base pair insertion that is not observed in either its father's (NA12891) or mother's (NA12892) genotypes. The top row denotes the reference genome sequence, with the STR region highlighted by the gold box. Subsequent rows illustrate the maximum likelihood alignment for each read relative to the reference genome. Bases with a red background or a red dash denote an insertion or deletion relative to the reference genome, respectively. To facilitate visualization, only a third of reads that span the STR region are shown.

calls and phased these calls into scaffolds using Beagle's pedigree option [176] and the 1000 Genomes Phase 3 reference panel [49]. Rerunning HipSTR with these SNP scaffolds as an additional source of input resulted in the same level of Mendelian consistency as described above but also phased ∼50% of the heterozygous STR calls onto their SNP scaffolds. For 178 markers where HipSTR phased the child's genotype, we could determine the transmitted paternal and maternal alleles from the unphased STR genotypes. In all of these instances, HipSTR correctly phased the paternal STR allele onto the paternal SNP scaffold, demonstrating the accuracy of its physical phasing approach.

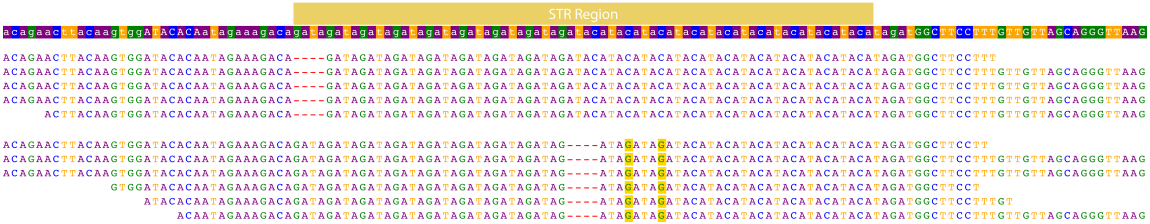

Figure 4-4: **HipSTR can reliably detect STR sequence differences**. At this marker, the HipSTR genotype for NA12878 consists of two alleles, both of which have four base pair deletions but different underlying sequences: $(GATA)_8(CATA)_9$ and $(GATA)_{11}(CATA)_6$. The top row denotes the reference genome, where the $(GATA)_9(CATA)_9$ STR sequence is highlighted in yellow. Each subsequent row illustrates the maximum likelihood alignment of the reads supporting the two alleles relative to the reference genome. Bases with yellow boxes or red dashes indicate mismatches and deletions relative to the reference genome, respectively. To facilitate visualization, only a third of reads that span the STR region are shown.

4.2.3 Quantifying STR de novo mutations

Encouraged by HipSTR's extreme precision and the candidate mutations it identified in the Marshfield markers, we used it to identify de novo STR mutations genome-wide. We again focused on the CEPH trio described above as these individuals have been deeply sequenced and provide the best datasets to identify STR variants. To generate a set of putative mutations, we used BWA-MEM to align datasets with 50-200x sequencing coverage from the Illumina Platinum Genomes Project. We then ran HipSTR using these alignments and a reference panel of STR regions we constructed in our previous studies [138], resulting in genotypes for ∼1.6 million STRs, many of which are homopolymer repeats. After stringent quality filtering **(Methods)**, we identified 423 loci where the genotypes were inconsistent with Mendelian inheritance and the child had a novel allele.

Next, we sought to validate these mutations using orthogonal datasets. We generated BWA-MEM alignments using Illumina datasets from the Genome in a Bottle Consortium [183] and 1000 Genomes Project [49] and used HipSTR to recall each site. For each marker, we compared individuals' allele lengths in the discovery and validation call sets. The lengths matched for all individuals at 358 of the sites, resulting in an ∼85% replication rate. The resulting set of high quality de novos predominantly occur at homopolymer repeats (293) but also occur at 34, 10, 17, 3 and 1 sites with 2-6 base pair repeat units, respectively.

Lastly, we sought to determine an upper bound for the number of these de novo mutations that

occur in the germline. A prior study of de novo SNP mutations in the CEPH trio identified ∼1000 mutations, only 49 of which occurred in the germline [184]. As a result, we postulated that many of the high confidence STR mutations may have arisen either somatically or within the cell line. We used HipSTR and sequencing data with 50x coverage from the Illumina Platinum Genomes to genotype the 11 children of NA12878 at each of the 358 replicated sites. After discarding loci where NA12878's de novo allele was absent from all its children, only 75 of the sites remained. This preliminary analysis suggests that most STR de novos for this trio do not occur within the germline, consistent with the prior SNP findings. In future work, we aim to use more detailed approaches to further exclude false positives from this set of 75 germline mutation candidates.

4.2.4 Phasing and imputing STRs

Statistical genetic approaches for phasing and imputation have been extensively used to construct phased haplotypes [2, 49, 3] and boost power in genome-wide association studies [185]. However, most of these approaches are limited to biallelic variants. Beagle [176] is one widely used tool that can accommodate multiallelic variants, motivating us to explore its applicability to STRs. To generate datasets for these analyses, we again focused on the Marshfield markers described earlier. We constructed phased SNP haplotypes by applying Beagle to preexisting unphased SNP calls for the SGDP samples [186], using the 1000 Genomes as a reference panel. We then used the SGDP HipSTR calls and capillary data to obtain high accuracy STR calls for 400 Marshfield markers across 263 individuals **(Methods)**.

We first sought to phase the STR calls onto the SNP haplotypes. For each marker, we generated in silico heterozygous STR calls with known phase by randomly pairing samples with homozygous STRs. We then used Beagle to jointly phase the in silico and remaining true heterozygotes onto the SNP scaffolds. Overall, the inferred and true phases matched for 85% of the in silico calls. To gauge whether we could improve this accuracy by filtering samples with uncertain phasing, we modified the Beagle algorithm to report the phasing confidence for each genotype and reran the analysis **(Methods)**. Requiring a minimum confidence of 90% retained over 74% of the phased calls and improved the accuracy to 94%. By examining the observed phasing accuracy for various confidence bins, we found that most reported confidences vastly overestimated the true accuracy **(Figure 4-7)**. However, we also discovered a large step-like increase in accuracy at the 90% confidence cutoff, suggesting that this is an appropriate threshold for filtering.

Encouraged by these results, we used Beagle to phase the original genotypes for each of the 400 markers. Overall, Beagle was able to obtain high confidence ($>$ 90%) phasings for 70% of heterozygous calls. Roughly 31% of the STR calls were homozygous, leaving 21% of STR calls with unknown phase after filtering. The fraction of calls Beagle could confidently phase varied widely by marker, ranging from 8-100% **(Figure 4-5)**, but was largely unaffected by the fraction of homozygous calls at each locus ($R^2 = 0.004$).

Next, we assessed the ability of Beagle to impute STRs. For each marker, we randomly selected 25 samples to impute and constructed a reference panel of phased STR-SNP haplotypes with the remaining samples. We then used the reference panel and the SNP haplotypes of the omitted samples to impute the sum of their STR allele sizes. Repeating this procedure ten times for each marker resulted in an overall accuracy of 43.4% when comparing the sum of allele sizes in the imputed and true genotypes **(Figure 4-5)**. In contrast, a naive strategy that imputed genotypes using the most commonly observed allele resulted in an accuracy of 23%, suggesting that Beagle effectively leverages the linkage disequilibrium around each STR to substantially improve accuracy. Despite the low imputation accuracy, imputed genotypes recovered more than 50% of the variance for more than half of the markers. Across loci, imputation accuracy and R^2 were both highly positively correlated with the ability to confidently phase heterozygous calls ($R^2 = 0.60$ and $R^2 = 0.66$, respectively).

Given the widespread use of STRs in forensics, we were particularly interested in the imputability of these markers. Of the 13 STRs used by the FBI for DNA identification, 3 were included in our panel of markers: D13S317, D8S1179 and TPOX. The imputation accuracy for D13S317 and D8S1179 were 30% and 33%, respectively, but the imputed genotypes recovered 60% of the genotyping variance for both markers **(Figure 4-6)**. In contrast, imputed genotypes for TPOX had a far higher accuracy of 66%. In addition, when we restricted comparisons to the 25% of imputed calls with posteriors greater than 90%, the TPOX genotypes were 100% accurate. The stark disparity in imputation performance between these three loci likely stems from the difference in their mutation rates. NIST has estimated per-generation mutation rates of $1.4 * 10^{-3}$ for both D8S1179 and D13S317 and $1 * 10^{-4}$ for TPOX. As all other CODIS markers have mutation rates comparable to D13S317 (apart from TH01), we believe that accurate imputation of most CODIS markers will remain challenging and error-prone.

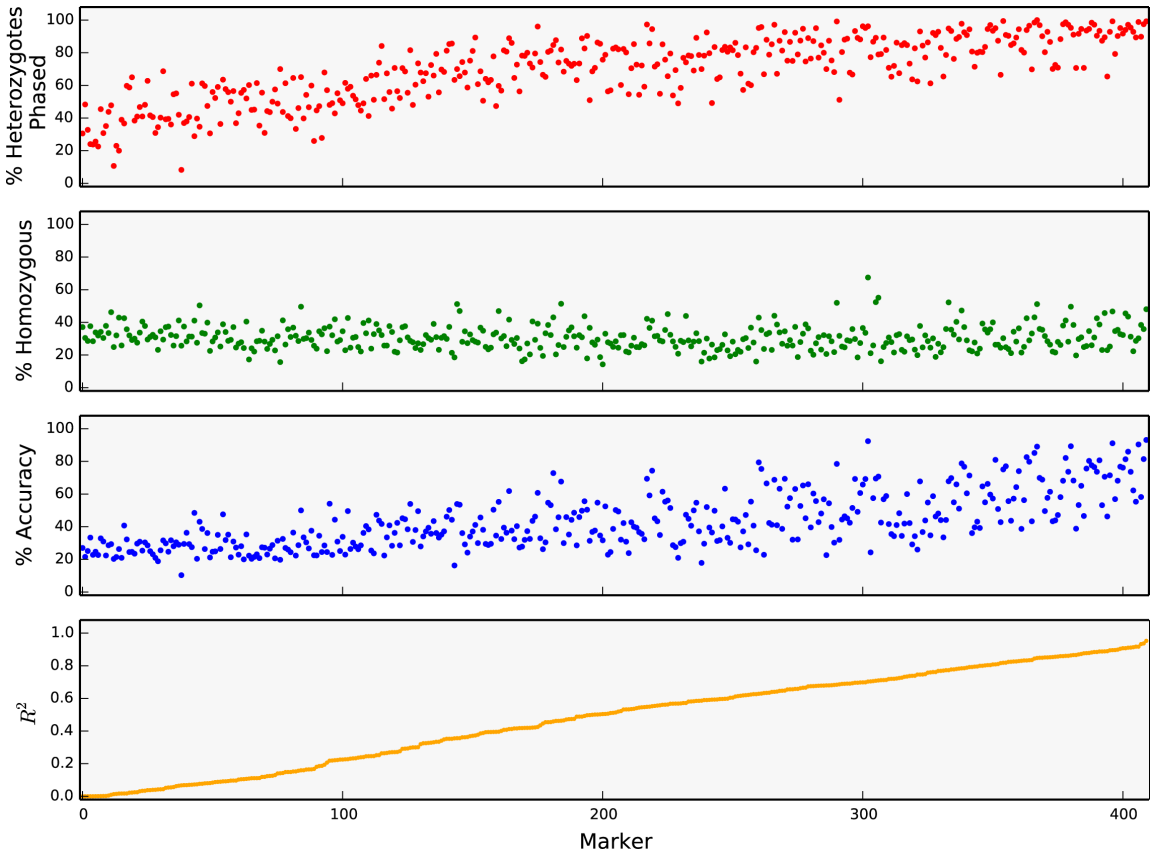

Figure 4-5: **Feasibility of phasing and imputing Marshfield STRs.** For each of the 400 Marshfield markers (x-axis), the fraction of heterozygous genotypes that can be phased with high confidence varies widely by marker. Similarly, the imputation accuracy (blue) and the R^2 (yellow) of imputed STR dosages with true STR dosages also vary widely by marker.

4.3 Discussion

Despite the abundance of STRs in the human genome and their extreme levels of variability, large-scale genomic studies frequently omit them from their analyses. This omission largely stems from the fact that STRs have traditionally been regarded as non-functional genomic elements. However, single-gene and genome-wide analyses continue to uncover their roles in gene expression regulation and phenotypic variation, largely dispelling this notion [42, 34, 173]. Difficulties associated with characterizing STRs from high-throughout sequencing data have also been a major cause of this omission. As a result, the focus of this chapter was to develop and

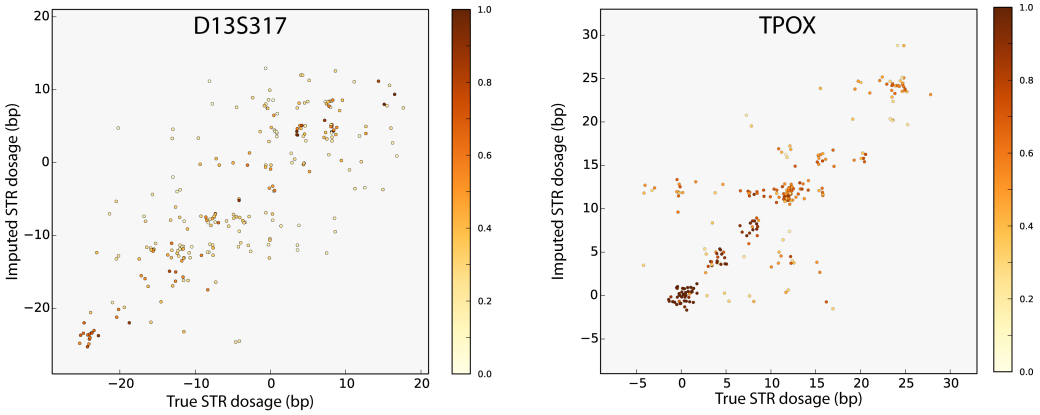

Figure 4-6: **Imputation of markers used in DNA identification.** The scatter plot depicts the ability of Beagle to impute genotypes for **(a)** D13S317 and **(b)** TPOX, 2 of 13 STRs used by the FBI for DNA identification. Every point depicts the sum of the STR allele sizes in the actual genotype (x-axis) versus the imputed genotype (y-axis) colored by the imputed genotype's posterior probability. Jitters in the x and y values are used to facilitate visualization.

assess the accuracy of approaches for incorporating STRs into future genomic analyses.

Our comparison of variant callers resulted in several rules of thumb for future studies interested in genotyping STRs. In general, most of the variant callers performed quite poorly in STR regions. Tweaking tool settings alleviated some of these performance issues, but Platypus, freebayes and RepeatSeq still resulted in call sets with high genotyping error rates. Future studies that integrate STR call sets from these tools should be wary of their limitations. On the other hand, GATK Haplotype Caller performed exceptionally well in STR regions. As this tool is frequently used in many large-scale analyses [187], the resulting STR calls may be extremely useful for understanding STR variation and mutability. One major caveat to this is that GATK frequently produces multiple variants per STR region, complicating downstream analysis and interpretability. We therefore believe that in addition to its improved accuracy and speed, HipSTR's ability to generate locally phased STR haplotypes makes it ideal for STR-related analyses.

Our study also provides novel suggestions for phasing and imputing STRs using existing approaches. We found that a minor modification to Beagle makes it well-suited for phasing STR genotypes. These phased haplotypes may be useful for obtaining more detailed insights into the regulatory roles of STRs, such as their role in allele-specific expression [44]. The results we obtained while imputing the Marshfield markers also suggest that imputed STR genotypes may

be useful for downstream analyses. While the raw genotype accuracy was relatively low for most markers, imputed dosages recovered a substantial amount of the genetic variance. As a result, STR imputation may provide an effective means of incorporating STRs into GWAS.

Despite the novelty of our study, it has several major limitations. While the most prevalent STRs in the human genome are homopolymer and dinucleotide repeats [138], the majority of the markers considered in this study had 3-4 bp repeat units. As a result, the observed levels of variant caller accuracy may not reflect performance genome-wide. We do note, however, that the Marshfield panel is comprised of highly polymorphic and long STRs, creating added difficulties absent at most STRs in the genome. We were also limited by the small number of samples in the SGDP dataset during our assessment of phasing and imputation. Future studies may benefit from assessing whether large reference panels of STRs substantially improve the accuracy and R^2 of imputed genotypes, as has been observed for SNP genotypes [188].

Lastly, our results raise some very exciting avenues for future STR research. In addition to demonstrating its high precision, HipSTR's ability to identify several hundred replicable de novo mutations highlights STRs as a rich potential source of de novo variation. Prior analyses for the CEPH trio have suggested that most de novo SNP mutations are cell-line specific [184] and our preliminary analyses are consistent with this conclusion. Nonetheless, characterizing STR mutations in trio-based cohorts will be invaluable to understanding the true genetic load of STRs and their putative contributions to complex traits.

4.4 Methods

4.4.1 Modeling PCR stutter

PCR stutter artifacts add or remove copies of an STR's motif to sequencing reads, resulting in observed STRs that differ in size from the true underlying genotype [62, 180]. To mitigate these effects, HipSTR constructs a stutter model θ_x for each STR locus x. The model contains the probability that stutter adds (u) or removes (d) repeats from the true allele in an observed read, and a geometric distribution with parameter ρ_s that controls the size of the stutter-induced changes. In our framework, the probability of observing a stutter artifact δ repeat units in size is:

$$P(stutter = \delta|\theta_x) \begin{cases} 1-u-d & \delta = 0 \\ u\rho_s(1-\rho_s)^{\delta-1} & \delta > 0 \\ d\rho_s(1-\rho_s)^{-\delta-1} & \delta < 0 \end{cases}$$

To estimate each locus' stutter model parameters, we extract the size of the STR observed in each read for all individuals in the population. We then use an Expectation-Maximization approach [189] to learn the parameters. The E-step computes each sample's genotype posteriors:

$$P(g_i = (j,k)|R,\theta_x^t) \propto f_j f_k \prod_{m=1}^{n_{reads,i}} \sum_{a \in j,k} \begin{cases} 1-u-d & r_{m,i} = r_a \\ u\rho_s(1-\rho_s)^{r_{m,i}-r_a-1} & r_{m,i} > r_a \\ d\rho_s(1-\rho_s)^{r_a-r_{m,i}-1} & r_{m,i} < r_a \end{cases}$$

Here, R denotes the set of all reads, g_i denotes the phased genotype for the i^{th} individual, $n_{reads,i}$ denotes the number of reads for the i^{th} sample, $r_{m,i}$ denotes the number of repeats in the m^{th} read for the i^{th} individual, r_a denotes the number of repeats in the a^{th} allele and f_j denotes the frequency of the j^{th} allele. For each possible phased genotype, the E-step also computes the conditional probability that a read originated from either allele:

$$P(a_{m,i} = j|g_i = (j,k),\theta^t) \propto \begin{cases} 1-u-d & r_{m,i} = r_j \\ u\rho_s(1-\rho_s)^{r_{m,i}-r_j-1} & r_{m,i} > r_j \\ d\rho_s(1-\rho_s)^{r_j-r_{m,i}-1} & r_{m,i} < r_j \end{cases}$$

The M-step updates the stutter model parameters and allele frequencies using these probabilities:

$$u^{t+1} = \frac{1}{Q}\sum_{i=1}^{N}\sum_{j=1}^{A}\sum_{k=1}^{A} P(g_i = (j,k)|R,\theta^t) \sum_{m=1}^{n_{reads,i}} \sum_{a \in j,k} P(a_{m,i} = a|g_i = (j,k),\theta^t) I(r_{m,i} > r_a)$$

$$d^{t+1} = \frac{1}{Q}\sum_{i=1}^{N}\sum_{j=1}^{A}\sum_{k=1}^{A} P(g_i = (j,k)|R,\theta^t) \sum_{m=1}^{n_{reads,i}} \sum_{a \in j,k} P(a_{m,i} = a|g_i = (j,k),\theta^t) I(r_{m,i} < r_a)$$

$$\rho_s^{t+1} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{A}\sum_{k=1}^{A} P(g_i = (j,k)|R,\theta^t) \sum_{m=1}^{n_{reads,i}} \sum_{a \in j,k} P(a_{m,i} = a|g_i = (j,k),\theta^t) I(r_{m,i} \neq r_a)}{\sum_{i=1}^{N}\sum_{j=1}^{A}\sum_{k=1}^{A} P(g_i = (j,k)|R,\theta^t) \sum_{m=1}^{n_{reads,i}} \sum_{a \in j,k} P(a_{m,i} = a|g_i = (j,k),\theta^t) \, |r_{m,i} - r_a|}$$

$$f_j^{t+1} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{A} P(g_i = (j,k)|R,\theta^t) + P(g_i = (k,j)|R,\theta^t)$$

Intuitively, the update rules for the stutter probabilities u and d compute the fraction of times a read's STR allele is either larger or smaller than its underlying allele. The update rule for the step size parameter ρ_s is more involved, but it first restricts the computation to reads with non-zero stutter. It then computes the inverse of the mean weighted step size, consistent with a maximum likelihood estimator for a geometric distribution.

4.4.2 Generating candidate alleles

To identify an initial set of STR alleles, HipSTR selects reads that span the STR region and are well-anchored. In particular, it requires that both ends of a read match the reference genome for at least 10 base pairs and that no longer end matches are present 15 base pairs upstream or downstream from the read. Using this subset of reads, it includes a sequence as a candidate allele if is present in 2 or more and least 20% of a sample's reads.

HipSTR also uses an iterative approach to identify new candidate alleles. At the end of every round of its pipeline, it computes the ML genotype for each sample and realigns every read relative to the most probable allele in its sample's genotype. Each of these retraced alignments generates a sequence in the STR region. If the same sequence is observed in a sample in 2 or more alignments with stutter artifacts, HipSTR identifies the sequence as a new candidate allele. This iterative approach enables it to identify larger and larger indels each iteration.

4.4.3 Computing genotype likelihoods

HipSTR's genotype likelihood model integrates information about every read's phasing likelihood and alignment likelihood. For the m^{th} read for sample i, $P(p_{m,i}|h=1)$ and $P(p_{m,i}|h=2)$ denote the phasing likelihoods of the read originating from the first and second SNP haplotypes, while $P(s_{m,i}|a=j)$ denotes the alignment likelihood of the read to the j^{th} allele. We use a uniform prior for each unphased genotype, such that heterozygous phased genotypes have half the prior probability of their homozygous counterparts. The likelihoods for the phased genotypes are:

$$P(g_i=(j,j)|R_i) \propto 2 \prod_{m=1}^{n_{reads,i}} \left[P(p_{m,i}|h=1)+P(p_{m,i}|h=2)\right] P(s_{m,i}|a=j)$$

$$P(g_i=(j,k)|R_i) \propto \prod_{m=1}^{n_{reads,i}} \left[P(p_{m,i}|h=1)P(s_{m,i}|a=j)+P(p_{m,i}|h=2)P(s_{m,i}|a=k)\right]$$

4.4.4 Read phasing likelihoods

To compute the phasing likelihoods for each read, HipSTR examines bases in the read or its mate pair that are aligned to heterozygous SNPs. If the base (b_i) truly came from the haplotype under consideration, the likelihood of it matching the SNP base (h_j) is given by the base quality q_{b_i}, while the likelihood of it not matching is one third of the residual probability. We express this as:

$$Q(b_i,h_j) = \begin{cases} q_{b_i} & b_i = h_j \\ \frac{1-q_{b_i}}{3} & b_i \neq h_j \end{cases} \tag{4.1}$$

We compute each read's total phasing likelihood by multiplying $Q(b_i,h_j)$ for every base b_i aligned to a heterozygous haplotype SNP h_j in the read or its mate pair. In practice, SNP calls in and around STR regions are likely to be error-prone. We therefore exclude SNPs that are within 15 base pairs of the STR region when computing the phasing likelihoods.

4.4.5 Aligning reads to flanking sequences

HipSTR assumes that each haplotype is comprised of two distinct types of regions: flanking sequences and STR sequences. Within flanking sequences, the model aims to account for

Illumina sequencing errors using a previously described hidden Markov model [190]. Each read base is aligned with a haplotype base or an inserted sequence, while every spanned haplotype base is aligned with a read base or a deletion. To efficiently perform these computations, we use three matrices to recursively compute the maximum log-likelihood of aligning read bases $b_1 \dots b_i$ with haplotype bases $h_1 \dots h_j$. The matrices M, I and D are used to track the likelihoods that read base b_i is aligned to haplotype base h_j, an insertion or a deletion, respectively.

In conjunction with values for $t_{X \to Y}$, the log-probability of transitioning from hidden state X to hidden state Y, we use the following recursions to fill in each matrix column-by-column:

$$M(i,j) = Q(b_i, h_j) + \max \begin{cases} M(i-1, j-1) + t_{M \to M} \\ D(i-1, j-1) + t_{M \to D} \\ I(i-1, j-1) + t_{M \to I} \end{cases}$$

$$I(i,j) = Q(b_i, b_i) + \max \begin{cases} I(i-1, j) + t_{I \to I} \\ M(i-1, j) + t_{I \to M} \end{cases}$$

$$D(i,j) = \max \begin{cases} M(i, j-1) + t_{D \to M} \\ D(i, j-1) + t_{D \to D} \end{cases}$$

4.4.6 Aligning reads to STR sequences

In STR regions, HipSTR utilizes an entirely different alignment model. It assumes that at most one indel occurs within the region and requires that its magnitude D be a multiple of the motif length, capturing the unique nature of PCR stutter artifacts. If no indel has occurred, the likelihood of the observed sequence is governed by the agreement between each base in the read and its corresponding haplotype base. The probability of aligning base b_i and its preceding bases to an STR sequence $h_1 \dots h_L$ of length L is:

$$P(b_{\max(1,i-L+1)} \dots b_i | D, h_1 \dots h_L) = \prod_{k=0}^{\min(L-1,i-1)} Q(b_{i-k}, h_{L-k})$$

If a deletion occurs, we assume that it can arise anywhere within the STR region. We iterate over these configurations, each of which has a likelihood given by the agreement between the

sequenced bases and their corresponding haplotype bases:

$$P(b_{\max(1,i-L+D+1)} \dots b_i | D, h_1 \dots h_L) =$$
$$\frac{1}{L-D+1} \sum_{d=0}^{L-D} \prod_{k=0}^{\min(d-1,i-1)} Q(b_{i-k}, h_{L-k}) \prod_{k=d}^{\min(L-D-1,i-1)} Q(b_{i-k}, h_{L-D-k})$$

Finally, if an insertion occurs, we assume that it can precede any base in the STR region. As PCR stutter insertions typically contain sequences that copy the local repeat structure, we assume that inserted sequences are periodic copies of the STR sequence directly preceding the insertion. We therefore measure the likelihood of inserted bases according to their agreement with this sequence. For an STR with a repeat motif of length M, iterating over each possible insertion position results in the likelihood:

$$P(b_{\max(1,i-L-D+1)} \dots b_i | D, h_1 \dots h_L) =$$
$$\frac{1}{L+1} \sum_{d=0}^{L} \prod_{k=0}^{\min(d-1,i-1)} Q(b_{i-k}, h_{L-k})$$
$$\prod_{k=d}^{\min(d+D-1,i-1)} \begin{cases} Q(b_{i-k}, h_{L-d-((k-d) \bmod M)}), & L-d-((k-d) \bmod M) \geq 1 \\ Q(b_{i-k}, h_{M+L-d-((k-d) \bmod M)}), & \text{otherwise} \end{cases}$$
$$\prod_{k=d+D}^{\min(L+D-1,i-1)} Q(b_{i-k}, h_{L+D-k})$$

4.4.7 Trio genotypes for the Marshfield markers

We ran HipSTR using data from both the Illumina Platinum Genomes Project and an additional study of the effects of PCR amplification on sequencing errors. Data from accession numbers SRX264762, SRX264792 and ERX168849 were used to genotype NA12891. Similarly, data from SRX264763, SRX264794 and ERX168850 were used to genotype NA12892, while accessions SRX264790, SRX264793 and ERX168836 provided data for NA12878. All three samples were genotyped using HipSTRv0.2 and the options *--def-stutter-model --use-all-reads --read-qual-trim #*.

4.4.8 Filtering de novo calls

We applied a series of stringent filters to the CEPH trio call set to reduce the likelihood that genotyping errors introduce false positive de novos. In particular, we required that all three individuals in the trio have a minimum genotype posterior (Q) of 0.9, no more than 10% of reads containing either a stutter artifact or a flanking sequence indel (DSTUTTER/DP and DFLANKINDEL/DP) and at least 10 reads spanning the STR region. Lastly, we required that the ratio of spanning reads supporting the alleles be at least 20%.

4.4.9 Generating the STR call set for imputation and phasing assessment

To generate a robust set of calls for imputation and phasing assessment, we used the HipSTR call set for the SGDP dataset and applied a series of additional filters. In particular, we removed an individual call if the posterior probability (Q) was less than 90%, more than 15% of reads had a stutter artifact or indel in the flanking sequence (DSTUTTER/DP and DFLANKINDEL/DP) or the call had no spanning reads. We then compared the remaining calls to the capillary data and removed genotypes if they were inconsistent with the capillary lengths. Finally, to avoid issues caused by high levels of missingness, we discarded loci if fewer than 85% of samples had genotypes after filtering, resulting in 410 loci for downstream analyses.

4.4.10 Modifying Beagle to emit phasing confidence

Internally, v4.0 of Beagle keeps track of unphased genotypes. To modify the program to emit phasing confidences, we altered these data structures such that they use phased genotypes. The phasing confidences are then readily generated based on the observed frequencies of the two phased genotypes. This modified version of Beagle is available at https://github.com/tfwillems/PhasedBEAGLE.

4.5 Supplemental Figures

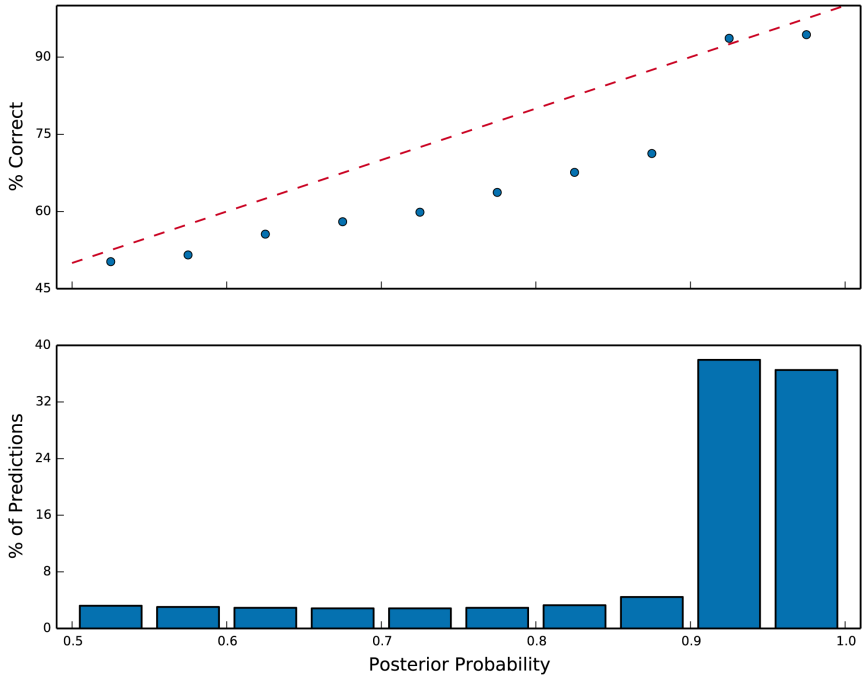

Figure 4-7: **Phasing accuracy versus reported phasing confidence.** In general, the observed phasing accuracy is far below the reported confidence level. However, a large increase in accuracy is observed for confidences above 90%, suggesting that this is an appropriate threshold for filtering the phased STR genotypes.

4.6 Supplemental Tables

Table 4.2: Optimal settings for STR genotyping

Method	Version	Explored Options	Optimal Options
freebayes	1.0.2-6 (g3ce827d)	genotype-qualities no-partial-observations report-genotype-likelihood-max	genotype-qualities no-partial-observations
GATK HC	3.5-0 (g36282e4)	max_alternate_alleles pcr_indel_model kmerSize	max_alternate_alleles 25
HipSTR	0.2	None	None
lobSTR	4.0.0	filter-mapq0 filter-clipped max-repeats-in-ends min-read-end-match	None
Platypus	0.8.1	maxVariants minVarFreq	maxVariants 25 minVarFreq 0.0
RepeatSeq	0.8.2	None	None

Table 4.3: Features used to build STR genotype classifiers

Method	Feature Name	Source[1]	Description
freebayes	GQ	F	Genotype quality
	DP	F	Genotype read depth
	GLDIFF	F	Difference between genotype likelihoods of the called and next best genotypes
	READRATIO[2]	F	Minimum ratio of read depths for called alleles
	UNCALLEDFRAC[2]	F	Fraction of reads observed for an uncalled allele
GATK HC	QD	I	Ratio of variant quality to read depth
	FS	I	P-value for strand bias
	ReadPosRankSum	I	Z-score for test of alt vs. ref read position bias
	GQ	F	Genotype quality
	DP	F	Genotype read depth
HipSTR	Q	F	Genotyper posterior probability
	SPANDP	F	Minimum number of reads spanning either allele
	DSTUTTER/DP	F	Fraction of reads with a stutter artifact
	DFLANKINDEL/DP	F	Fraction of reads with an indel flanking the STR
lobSTR	SB	F	P-value for strand bias
	Q	F	Genotype posterior probability
	DP	F	Genotype read depth
	DISTENDS	F	Average difference between distance of STR to read ends
Platypus	QD	I	Ratio of variant quality to read depth
	SbPval	I	P-value for strand bias
	GQ	F	Genotype quality
	GOD	F	Genotype goodness of fit value
RepeatSeq	GQ	F	Genotype quality
	DP	F	Genotype read depth

[1] Describes whether the field is calculated from FORMAT fields (F) or INFO fields (I) in the variant caller's VCF

[2] Computed using the RO and DPR FORMAT fields

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Population-scale sequencing data enables precise estimates of Y-STR mutation rates

---

This chapter is awaiting publication as:

> **Willems TF**, Gymrek M, Poznik GD, Tyler-Smith C, The 1000 Genomes Project Chromosome Y Group, Erlich Y. Population-Scale Sequencing Data Enables Precise Estimates of Y-STR Mutation Rates. *Am J Hum Genet*. (In Press).

Abstract: Short Tandem Repeats (STRs) are mutation-prone loci that span nearly 1% of the human genome. Previous studies have estimated the mutation rates of highly polymorphic STRs using capillary electrophoresis and pedigree-based designs. While this work has provided insights into the mutational dynamics of highly mutable STRs, the mutation rates of most others remain unknown. Here, we harnessed whole-genome sequencing data to estimate the mutation rates of Y-chromosome STRs (Y-STRs) with 2-6 base pair repeat units that are accessible to Illumina sequencing. We genotyped 4,500 Y-STRs using data from the 1000 Genomes Project and the Simons Genome Diversity Project. Next, we developed MUTEA, an algorithm that infers STR mutation rates from population-scale data using a high-resolution SNP-based phylogeny. After extensive intrinsic and extrinsic validations, we harnessed MUTEA to derive mutation rate estimates for 702 polymorphic STRs by tracing each locus over 222,000 meioses, resulting in the largest collection of Y-STR mutation rates to date. Using our estimates, we identified determinants of STR mutation rates and built a model to predict rates for STRs across the genome. These predictions indicate that the load of de novo STR mutations is at least 75 mutations per generation, rivaling the load of all other known variant types. Finally, we identified Y-STRs with potential applications in forensics and genetic genealogy, assessed the ability to differentiate between the Y-chromosomes of father-son pairs, and imputed Y-STR genotypes.

5.1 Introduction

Mutations provide the fuel for evolutionary processes. The rates at which new mutations arise play a central role in a range of genetic applications, including dating phylogenetic events [191], informing disease studies [192], and evaluating forensic evidence [69]. The advent of high-throughput sequencing has enabled genome-wide measurements of the number of de novo mutations using a broad range of strategies. A host of studies have evaluated the mutation rates of nearly every type of genetic variation, ranging from SNPs [184, 193, 58, 59] and short indels [194] to large structural variations [60]. These sequencing studies have concluded that approximately 50-100 de novo mutations arise each generation, most of which are point mutations. However, these studies have largely overlooked the contribution of short tandem repeats (STRs).

STRs are one of the most abundant types of repeats in the human genome. They consist of a repeating 2-6 base pair (bp) motif and span a median of 25bp. Approximately 700,000 STR loci exist in the human genome that in aggregate occupy $\sim$1% of its total length. STR variations have been implicated in more than 30 hereditary disorders [28], and emerging lines of evidence have highlighted their involvement in complex traits in both humans [41, 93, 94] and model organisms [38, 40, 129]. The repetitive nature of STRs causes error-prone DNA-polymerase replication events that can insert or delete copies of the repeat motif in subsequent generations, leading to markedly elevated mutation rates [23, 62].

Previous studies estimated the rates and patterns of de novo STR mutations using capillary electrophoresis genotyping of specialized sets of markers, such as the Marshfield panel, the CODIS markers, or specific Y-chromosome STRs (Y-STRs). These studies have estimated that the average STR mutation rate per locus is 10^{-3} to 10^{-4} mutations per generation (mpg) [23, 22, 195, 56, 196]. However, most STRs characterized in these studies were chosen for their relatively high levels of diversity in the population. As such, it is not clear whether their mutation rates and patterns reflect most STRs in the genome. Furthermore, as most previously studied STRs have tri- and tetranucleotide motifs, the field lacks robust mutation rate estimates for other motif lengths, specifically dinucleotides, the most prevalent type of STR. Finally, capillary electrophoresis has relatively low throughput, and most STRs were never genotyped in these studies, leaving the specific mutation rates of most STRs unknown.

The rapid advancement of next-generation sequencing technologies has provided the opportunity

to genotype STRs beyond those on existing panels and to do so on a larger scale. Coupled with vast improvements in the depth, read length, and quality of whole-genome sequencing (WGS) datasets, algorithmic progress in STR genotyping tools has made it possible to robustly call these markers from high-throughput data [27, 52, 197]. In our previous study, we found that 90% of the STRs in the genome are accessible to Illumina technology, and we showed that hemizygous STRs can be called with very high accuracy [138].

Here, we leveraged population-scale high-throughput sequencing data to systematically estimate the mutation rates and analyze the mutational dynamics of STRs across the Y-chromosome. To gain power, we used two independent datasets, the 1000 Genomes Project [49] and the Simons Genome Diversity Project (SGDP) [181]. The Y-chromosomes in these datasets confer rich genealogical information, enabling the analysis of complex STR mutation models without the need for familial information. To leverage this genealogical information, we developed an algorithm, Measuring Mutation Rates using Trees and Error Awareness (MUTEA), which infers the mutational dynamics along the Y-chromosome branches. After validating MUTEA via intrinsic and extrinsic tests, we scanned 4,500 Y-STRs and used the algorithm to infer the mutation rates of 702 polymorphic Y-STRs. To the best of our knowledge, this is the largest collection of Y-STR mutation rates to date. We show the value of this large collection of mutation rates by uncovering the sequence determinants of mutability, predicting the genetic load of de novo STR mutations across the genome, and exploring a series of forensic applications.

5.2 Materials and methods

5.2.1 Sequencing datasets

We analyzed 179 male samples in the SGDP cohort from widely dispersed populations across Africa, Asia and the Americas. The SGDP samples were sequenced to over 30x coverage using a PCR-free library preparation protocol and 100bp paired-end Illumina reads. As our previous results demonstrate that this protocol substantially reduces the rate of PCR stutter at STR loci [48], the SGDP cohort provides a high-quality dataset for calling Y-STRs. We also analyzed 1,244 unrelated male samples from phase 3 of the 1000 Genomes Project. These samples are from 26 globally diverse populations and were sequenced to an average autosomal coverage of 7x using 75-100 bp paired-end Illumina reads.

5.2.2 Y-SNP phylogeny

To construct the SGDP Y-chromosome haplotype tree, we downloaded VCF files containing the Y-SNP calls generated by the SGDP analysis group. As many of these SNPs lie in pseudoautosomal regions or regions with low mappability, we applied a series of filters to reduce the frequency of genotyping errors. We first removed loci where more than 10% of individuals were heterozygous using VCFtools [198]. For the remaining SNPs, we removed individual SNP calls that were heterozygous, had fewer than 7 supporting reads, or had more than 10% of reads supporting an uncalled allele. Lastly, we discarded SNP loci if fewer than 150 samples met these criteria or if more than 10% of reads had zero mapping quality. Overall, we obtained nearly 39,000 high-quality polymorphic SNPs.

We then used the high-quality SNPs to build the Y-chromosome phylogenetic tree using RAxML [199] and the options *-m ASC_GTRGAMMA -f d -asc-corr lewis*. The SGDP samples included 3 representatives of haplogroup A1b1 and no members of the more basal clades (A00, A0, and A1a), so we used Dendroscope [200] to root the phylogeny along the branch marked by the M42 and M94 mutations, markers associated with the split between A1b1 and megahaplogroup BT. For the 1000 Genomes phase 3 dataset, we used a RAxML-generated phylogeny that was built by the 1000Y analysis group [201].

Although the maximum-likelihood phylogeny generated for each dataset has numerical branch lengths, these lengths are not scaled in units of generations as required by our method. We therefore tested two scaling approaches. First, we selected the factor that most closely equated the total number of generations in each phylogeny to the corresponding value based on published Y-SNP mutation rates. To do so, we used a recently published Y-SNP mutation rate of 310^{-8} mutations per base per generation [202, 203] and the numbers of called SNPs and called sites in each SNP dataset. As an alternative method, we scaled the trees using mutation rate estimates for 15 loci in the Y-chromosome Haplotype Reference Database (YHRD), a large compendium of individual Y-STR mutational studies (individually cited therein) [204]. We chose to calibrate using these loci because their mutation rate estimates are each based on more than 7,000 father-son pairs per locus and should therefore be relatively precise. For the 1000 Genomes data, we used the available PowerPlex capillary data for each locus, assumed error-free genotypes, scaled the phylogeny using a range of factors, and estimated the set of mutation rates for each scaling factor using MUTEA (see below). The choice of scaling factor had essentially no effect on the correlation with the YHRD estimates, resulting in an R^2 of 0.89 across all tested factors

(Supplemental Figure 5-6). However, the total squared error between the estimates was minimized for a factor of ∼2,800, which we therefore selected as the optimal scaling. For the SGDP data, we performed an analogous analysis using HipSTR genotypes (see below) for 9 of these 15 loci, again resulting in a uniform R^2 of 0.91 and an optimal scaling factor of ∼3,200 **(Supplemental Figure 5-6)**.

The resulting scaling factors were remarkably concordant between the methods, with the factors determined by the Y-SNP method ∼25% greater. However, to maximize the concordance with pedigree estimates, we used the second method. After scaling the branches, we found that the approximate total lengths of the SGDP and 1000 Genomes phylogenies are 60,000 and 160,000 meioses, respectively.

5.2.3 Defining and identifying Y-STRs

To identify Y-STRs, we used a quantitative procedure developed in our previous work [138]. Briefly, this procedure uses Tandem Repeats Finder (TRF) to score each genomic sequence according to its purity, length, and nucleotide composition [89]. It then uses extensive simulations of random nucleotide sequences to determine a scoring threshold that distinguishes random DNA from DNA that is truly repetitive, selecting regions with scores above this threshold as STRs. Our previous results suggest that this approach has less than a 1.4% probability of omitting a polymorphic STR and has a false positive rate of approximately 1%.

We applied this procedure to the Y-chromosome sequence of the hg19 reference genome. As TRF occasionally identifies regions that overlap, we ensured that every locus has a unique STR annotation using the following steps: (1) We merged two STR regions if the higher scoring one contained 85% of the bases in the union of the regions (2) Overlapping entries that failed this criterion but which had the same period were also merged. For example, adjacent $[GATA]_{10}$ and $[TACA]_8$ entries were merged into one STR (3) Since we intended to use sequencing alignments relative to either hg19 or GRCh38 coordinates, we removed hg19 STR regions that failed to liftOver [205] to the GRCh38 assembly or were lifted from the Y-chromosome to the X-chromosome.

We also added coordinates for Y-STR loci whose mutation rates have been characterized in prior studies [56, 57]. For these markers, we used the published set of primer sequences and the isPCR tool [205] to map the primers to hg19 coordinates. We then ran TRF on each region and pinpointed the coordinates using the published repeat structure. Lastly, we applied

TRF to additional regions previously published as part of comprehensive Y-STR maps to obtain coordinates for labeled markers whose mutation rates have not been characterized [206]. In total, we added 261 annotated Y-STRs, ∼190 of which have mutation rate estimates from prior studies.

5.2.4 Y-STR call set and its accuracy

We downloaded BWA-MEM [182] alignments for the SGDP samples from the project website and extracted and merged the Y-chromosome alignments into a single BAM file using SAMtools [182]. STR genotypes were then generated using HipSTR, an improved version of lobSTR, an STR caller for Illumina data we developed in our previous studies [27].

HipSTR provides additional capabilities over lobSTR by using a specialized hidden Markov model (HMM) to account for PCR stutter artifacts. Briefly, to genotype an STR, HipSTR creates a list of candidate alleles from the alignments observed in the population. For each sample, it then realigns every read to each putative allele using the HMM, selects the allele with the highest total likelihood as the genotype, and returns each read's alignment relative to this genotype. This haplotype-based approach produces highly accurate STR genotypes and eliminates many read misalignments that occur if reads are aligned individually or are only aligned to the reference genome. We used HipSTR to genotype each STR region in the Y-STR reference described above using the merged BAMs and the following options: *--min-reads 25 --haploid-chrs chrY --hide-allreads*. Similarly, we downloaded BWA-MEM alignments from the 1000 Genomes phase 3 data release. As these alignments were relative to the GRCh38 assembly, we ran HipSTR using the corresponding GRCh38 STR regions and the options *--min-reads 100 --haploid-chrs chrY --hide-allreads*.

We employed several strategies to enhance the quality of the SGDP STR call set: (1) To avoid errors introduced by neighboring repeats, we omitted genotyped loci that overlapped one another or multiple STR regions (2) We discarded loci if more than 5% of samples' genotypes had a non-integer number of repeats, such as a three base pair expansion in an STR with a tetranucleotide motif. These types of events occur quite rarely and usually reflect genotyping errors rather than genuine STR polymorphisms [27] (3) We removed Y-STRs sites that were called in at least 2 SGDP females, as they are likely to have high X-chromosome or autosome homology (4) We omitted sites if more than 15% of reads had a stutter artifact or more than 7.5% of reads had in indel in the sequence flanking the STR. These HipSTR-reported statistics typically indicate that

the locus is not well captured by HipSTR's genotyping model and may arise if duplicated sites are mapping to the same reference genome location (5) For the remaining loci, we discarded unreliable calls on a per-sample basis if more than 10% of an individual's reads had an indel in the flank sequence (6) Finally, we removed loci in which fewer than 100 samples had genotype posteriors greater than 66%, as these loci had too few samples for accurate inference.

To filter the 1000 Genomes call set, we first removed loci that did not pass the SGDP dataset filters. We then applied a set of filters identical to those described above except that we only removed loci with more than 15 genotyped females and did not apply a stutter frequency cutoff. These alterations account for the 1000 Genomes dataset's larger sample size and use of PCR amplification during library preparation.

Importantly, we found that both the SGDP and 1000 Genomes HipSTR call sets had high quality. We compared our STR genotypes to capillary electrophoresis datasets available for the same samples. For the SGDP, we observed a 99.7% concordance rate when comparing the HipSTR and capillary results for 3,300 calls at 48 Y-STRs [207]. For the 1000 Genomes, a comparison of 4,050 calls at 15 loci in the PowerPlex Y23 panel resulted in a 97.5% concordance rate [208].

5.2.5 Measuring mutation rates using trees and error awareness (MUTEA): theory

Previously developed methods estimate STR mutation rates from population data by comparing the mean squared difference in allele lengths between samples to the time to the most recent common ancestor (TMRCA) [13, 14]. However, these methods generally assume simple mutation models, can be sensitive to haplogroup size fluctuations [209] and require exact error-free genotypes. We therefore sought to develop an algorithm that can address these issues by leveraging detailed Y-SNP phylogenies.

Figure 5-1 outlines the steps underlying MUTEA. Under a naive setting without genotyping error, MUTEA uses Felsenstein's pruning algorithm [210] and numerical optimization to evaluate and improve the likelihood of a mutation model until convergence. However, due to the error-prone and low-coverage nature of WGS-based STR call sets, using these genotypes would result in vastly inflated mutation rate estimates. To avoid these biases, MUTEA learns a locus-specific error model and uses this error model to compute genotype posteriors. It then uses these posteriors rather than fixed genotypes during the mutation model optimization process to obtain robust estimates. In addition, MUTEA uses a flexible computational framework for STR

mutations that includes length constraints and allows for multi-step mutations. We describe each step below.

5.2.6 Mutation model likelihood

We used Felsenstein's pruning algorithm to evaluate the likelihood of an STR mutation model. Let M denote the STR mutation model, D denote the dataset containing STR genotype likelihoods, and T denote the Y-chromosome phylogeny rooted at node R. The likelihood of the data is:

$$P(D|M,T) = \sum_r P(R=r, D|M,T) = \sum_r P(R=r|M,T)P(D|R=r,M,T)$$

Let D_{N_i} denote the genotype likelihoods of all nodes that are in the subtree rooted at node N_i. If node N_i has genotype g, the conditional probability of the data in its subtree is given by:

$$\begin{aligned} P(D_{N_i}|N_i=g,M,T) &= \prod_{C_j \in child(N_i)} \sum_{b \in alleles} P(C_j=b, D_{C_j}|N_i=g,M,T) \\ &= \prod_{C_j \in child(N_i)} \sum_{b \in alleles} P(C_j=b|N_i=g,M,T)P(D_{C_j}|C_j=b,M,T) \end{aligned}$$

While descending the phylogeny, this recursive relation applies until a node with no children is encountered. These leaf nodes represent sequenced individuals and the conditional probability of the data is given by the individuals' genotype likelihoods. Therefore, the likelihood of a mutation model can be calculated using a post-order tree traversal. First, the algorithm computes the genotype likelihoods at each leaf node. It then progresses to each internal node and calculates the conditional probability of the data for each potential genotype after computing its descendants' probabilities. Finally, upon reaching the root node, the total data likelihood is computed using the root node's conditional probabilities and a uniform prior for the root node's genotype.

In practice, we compute the total log-likelihood to avoid numerical underflow issues. Because normalizing the genotype likelihoods of each sample does not affect the relative model likelihoods, we calculated genotype posteriors using a uniform prior and used them throughout our analysis.

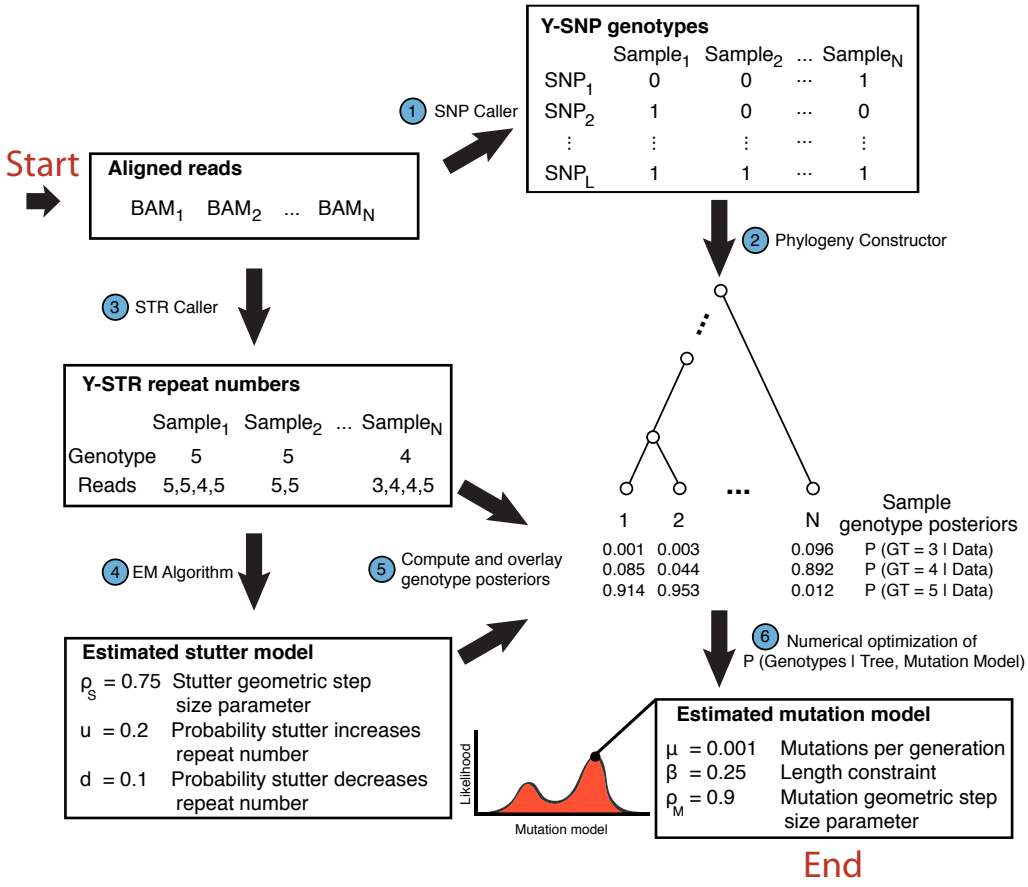

Figure 5-1: **Y-STR mutation rate estimation method.** Schematic of our procedure to estimate Y-STR mutation rates. The method first genotypes Y-SNPs (step 1) and uses these calls to build a single Y-SNP phylogeny (step 2). This phylogeny provides the evolutionary context required to infer Y-STR mutational dynamics, with samples in the cohort occupying the leaves of the tree and all other nodes representing unobserved ancestors. Steps 3-6 are then run on each Y-STR individually. After using an STR genotyping tool to determine each sample's maximum-likelihood genotype and the number of repeats in each read (step 3), an EM-algorithm analyzes all of these repeat counts to learn a stutter model (step 4). In combination with the read-level repeat counts, this model is used to compute each sample's genotype posteriors (step 5). After randomly initializing a mutation model, Felsenstein's pruning algorithm and numerical optimization are used to repeatedly evaluate and improve the likelihood of the model until convergence. The mutation rate in the resulting model provides the maximum-likelihood estimate.

5.2.7 STR mutation model

To model STR mutations, we used a generalized stepwise mutation model with a length constraint. Each mutation model M is characterized by three parameters: a per-generation mutation rate μ, a geometric step size distribution with parameter ρ_M and β, a spring-like length constraint that causes alleles to mutate back towards the central allele. In this framework, the central allele is assigned a value of zero, and nonzero allele values indicate the number of repeats from this reference point. Given a starting allele a_t observed at time t, the probability of observing a particular allele k the following generation is:

$$p(a_{t+1} = k|a_t) = \begin{cases} 1-\mu & k = a_t \\ \mu f_i \rho_M (1-\rho_M)^{k-a_t-1} & k > a_t \\ \mu f_d \rho_M (1-\rho_M)^{a_t-k-1} & k < a_t \end{cases}$$

where the fraction of mutations increasing and decreasing the size of the STR are $f_i = \frac{1-\beta\rho_M a_t}{2}$ and $f_d = 1-f_i$; f_i values greater than one or less than zero were clipped and set to one and zero, respectively. These two model features act as spring-like length constraints that attract alleles back towards the central allele. To avoid biologically implausible models, we constrained β to have non-negative values, where $\beta = 0$ reduces to a traditional generalized stepwise mutation model and increasingly positive values of β model STRs with stronger tendencies to mutate back towards the central allele. Values of ρ_M close to one primarily restrict models to single-step mutations, while smaller values of this parameter enable frequent multistep mutations.

5.2.8 Computing STR genotype likelihoods

To calculate the likelihood of the data D observed in the leaf nodes, we needed to account for STR genotyping errors. These errors are mainly caused by PCR stutter artifacts that insert or delete STR repeat units in the observed sequencing reads. We therefore developed a method to learn each STR's distinctive stutter noise profile.

Let θ_x denote the stutter model for STR locus x. θ_x is parameterized by the frequency of each STR allele (F_i), the probability that stutter adds (u) or removes (d) repeats from the true allele in an observed read, and a geometric distribution with parameter ρ_s that controls the size of the stutter-induced changes. Given a stutter model and a set of observed reads (R), the posterior

probability of each individual's haploid genotype is:

$$P(g_i = j|R, \theta_x) \propto F_j \prod_{k=1}^{n_{reads,i}} \begin{cases} 1-u-d & r_{k,i} = s_j \\ u\rho_s(1-\rho_s)^{r_{k,i}-s_j-1} & r_{k,i} > s_j \\ d\rho_s(1-\rho_s)^{s_j-r_{k,i}-1} & r_{k,i} < s_j \end{cases}$$

where g_i denotes the genotype of the i^{th} individual, $n_{reads,i}$ denotes the number of reads for the i^{th} individual, $r_{k,i}$ denotes the number of repeats observed in the k^{th} read for the i^{th} individual, and s_j denotes the number of repeats in the j^{th} allele. Analogous to the step size parameter in the mutation model, small values of ρ_s allow for frequent multistep stutter artifacts while values near one restrict artifacts to single step changes.

We implemented an expectation-maximization (EM) framework to learn these model parameters [189]. The E-step computes the genotype posteriors for every individual given the observed reads and the current stutter model parameters. The M-step then uses these posterior probabilities to update the stutter model parameters as follows:

$$u^{t+1} = \frac{1}{Q} \sum_{i=1}^{N} \sum_{j=1}^{A} P(g_i = j|R, \theta^t) \sum_{k=1}^{n_{reads,i}} I(r_{k,i} > s_j)$$

$$d^{t+1} = \frac{1}{Q} \sum_{i=1}^{N} \sum_{j=1}^{A} P(g_i = j|R, \theta^t) \sum_{k=1}^{n_{reads,i}} I(r_{k,i} < s_j)$$

$$\rho_s^{t+1} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{A} P(g_i = j|R, \theta^t) \sum_{k=1}^{n_{reads,i}} I(r_{k,i} \neq s_j)}{\sum_{i=1}^{N} \sum_{j=1}^{A} P(g_i = j|R, \theta^t) \sum_{k=1}^{n_{reads,i}} |r_{k,i} - s_j|}$$

$$F_j^{t+1} = \frac{1}{N} \sum_{i=1}^{N} P(g_i = j|R, \theta^t)$$

Here, N denotes the number of samples, A denotes the number of putative alleles, Q denotes the number of sequencing reads and I is the indicator function. As ρ_s is the parameter of a geometric step size distribution, the M-step updates its value using the inverse of the mean weighted step size for reads with non-zero stutter.

Locally misaligned reads can also introduce genotyping errors if they cause a miscalculation in a read's repeat length. However, these errors introduce artifacts that are relatively similar to those caused by PCR stutter. As a result, the EM procedure learns stutter models that correct for the

combined frequencies of PCR stutter and misalignment, resulting in robust genotype posteriors for downstream analyses.

5.2.9 MUTEA computation

Given genotype likelihoods for an STR of interest, we used a maximum-likelihood approach to estimate the underlying mutation model. Our approach first estimates the central allele of the mutation model by computing the median observed STR length and then normalizes all genotypes relative to this reference point. Next, it randomly selects mutation model parameters μ, β, and ρ_M subject to the constraint that they lie within the ranges of 10^{-5} to 0.05, 0 to 0.75 and 0.5 to 1.0, respectively. Using these bounds, the Nelder-Mead optimization algorithm [211], and the outlined method for computing each model's likelihood, we iteratively update the mutation model parameters until the likelihood converges. After repeating this procedure using three different random initializations to increase the probability of discovering a global optimum, our algorithm selects the optimized set of parameters with the greatest total likelihood.

For each STR in the SGDP and 1000 Genomes call sets that passed the requisite quality control filters, we first used the EM algorithm to learn a PCR stutter model. To run this algorithm, we obtained the size of the STR observed in each read from the MALLREADS VCF field. HipSTR uses this field to report the maximum-likelihood STR size observed in each read that spans its sample's most probable haplotype. In conjunction with a uniform prior, the learned stutter model was then used to compute the genotype posteriors for each sample with a HipSTR quality score greater than 0.66. Samples with quality scores below this threshold were omitted because the genotype uncertainty can result in erroneous reported read sizes. Finally, together with the optimization procedure and the appropriate scaled Y-SNP phylogeny, we used these genotype posteriors to obtain a point estimate of the STR's mutation rate.

5.3 Results

5.3.1 Verifying MUTEA using simulations

We validated MUTEA's inferences by running the algorithm on simulated data from a wide range of Y-STR mutation models **(Supplemental Text 5.5.1)**. We tested mutation rates (μ) from 10^{-5} to 10^{-2} mpg, a range that encompasses most known polymorphic Y-STRs. We also

varied the distribution of step-sizes for each STR mutation from a single step ($\rho_M = 1$) to a wide range of mutation steps ($\rho_M = 0.75$) and added various spring-like length constraints that ranged from no constraint ($\beta = 0$) to a strong attractor towards the central allele ($\beta = 0.5$).

MUTEA obtained unbiased estimates of the simulated mutation rate for nearly all scenarios **(Supplemental Figure 5-7)**. We only observed a slight upward bias in the estimates for the slowest simulated mutation rate ($\mu = 10^{-5}$) due to the lower bound imposed during numerical optimization. In contrast, mutation rates estimated using simpler mutation models limited to single-step mutations or no length constraints were far more biased in these scenarios **(Supplemental Figure 5-8)**. MUTEA's inferences were also robust to the presence of simulated PCR stutter noise. After forward simulating STRs, we simulated reads for each genotype and distorted their repeat numbers using various PCR stutter models **(Supplemental Text 5.5.2)**. We then input these repeat counts into MUTEA instead of the STR genotypes. Although MUTEA was completely blind to the selected stutter parameters, it reported unbiased estimates of the Y-STR mutation rates, step sizes, and stutter models for nearly all scenarios **(Figure 5-2, Supplemental Figures 5-9 to 5-11)**, with just a slight bias for the lowest simulated mutation rate, as was the case for the exact genotypes scenario described above. As a negative control, we again ran MUTEA on the stutter-affected reads but without employing the EM stutter correction method. With this procedure, posteriors based on the fraction of reads supporting each genotype resulted in marked biases, particularly for low mutation rates, demonstrating the importance of correctly accounting for stutter artifacts in these settings **(Figure 5-2, Supplemental Figures 5-10 to 5-11)**.

5.3.2 MUTEA estimates are internally and externally consistent

Encouraged by the robustness of our approach, we turned to analyze real Y-STR data from the SGDP and the 1000 Genomes Y-STR call sets. In total, we examined ~4,500 STR loci, 702 of which displayed length polymorphisms in both datasets, with the rest nearly fixed. We ran MUTEA on each of these polymorphic STRs to estimate its mutation rate (μ), expected step size (ρ_M), and stutter parameters (u, d, ρ_s) in both datasets **(Supplemental Table 5.4)**.

The MUTEA mutation rate estimates were largely consistent between the datasets **(Figure 5-3)**. We obtained an R^2 of 0.92 when comparing the log mutation rate estimates from the 1000 Genomes and SGDP datasets for the 702 polymorphic markers. Importantly, this high concordance was achieved despite substantial differences between the analyzed populations,

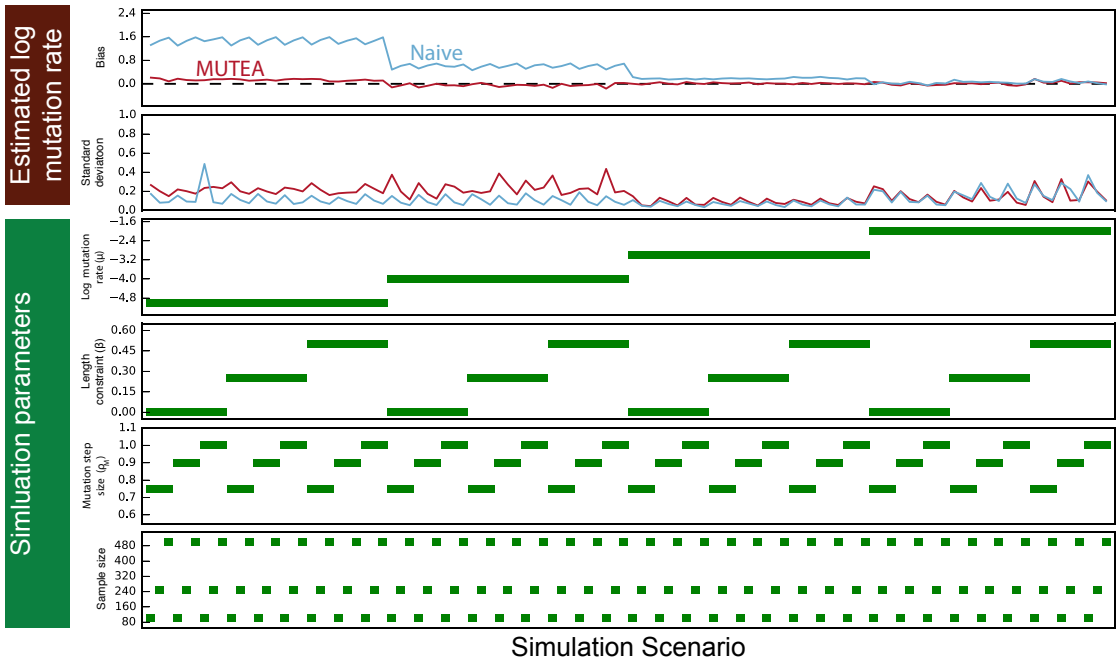

Figure 5-2: **Validating MUTEA using simulations.** STR sequencing reads with PCR stutter noise were simulated for a variety of sample sizes and mutation models (simulation parameters panels). Applying MUTEA (red line) to these reads led to relatively unbiased mutation rate estimates (upper panel) with small standard deviations (second panel). As a negative control, we also applied a naive approach to correct for stutter noise (blue line). This approach computed genotype posteriors using the fraction of supporting reads, resulting in markedly biased mutation rate estimates.

sample sizes, and sequencing data quality. The 1000 Genomes data should have higher rates of stutter than the SGDP data due to the PCR amplification used in the sequencing library preparation. Consistent with this expectation, MUTEA learned higher stutter probabilities in the 1000 Genomes data, as compared to the SGDP data, for most loci **(Supplemental Figure 5-12, left panels)**. Nonetheless, the mutation rate estimates were highly concordant. In addition, we found that despite differences in the overall probability of stutter, the downward and upward stutter rates were highly correlated between the two datasets ($R^2 = 0.88$ and $R^2 = 0.68$ on the log scale, respectively), reflecting the algorithm's ability to capture each locus' distinctive error profile **(Supplemental Figure 5-12, right panels)**.

Genotyping technology played only a small role in explaining the estimate concordance between

the two datasets. We re-ran MUTEA on the 1000 Genomes Y-tree using capillary genotypes for 15 Y-STR loci that were available for the same samples **(Figure 5-3)**. Comparing the resulting log mutation rate estimates to those obtained using sequencing-generated genotypes, we obtained an R^2 of 0.98. These comparisons demonstrate that our method obtains robust locus-specific mutation rate estimates while accounting for varying degrees of PCR stutter artifacts and alignment and genotyping errors. Furthermore, the inter-dataset concordance suggests that there are either very few errors in the phylogenies or that these errors have little impact on the resulting mutation rate estimates.

We also validated our mutation rate estimates by comparing them to results from previous studies that used pedigree-based designs and capillary electrophoresis for genotyping. In these studies, Burgarella et al. [57] and Ballantyne et al. [56] estimated Y-STR mutation rates for specialized panels of Y-STRs by examining approximately 500 and 2,000 father-son duos per Y-STR, respectively. We observed only a moderate replicability between the reported mutation rates from these two prior studies (R^2 of 0.34, **Figure 5-3**). This low correlation presumably stems from the very small number of transmissions used by Burgarella et al. On the other hand, we observed an R^2 of $\sim$0.65 when we compared either the SGDP or the 1000 Genomes estimates to those from Ballantyne et al., despite considerably different methodological approaches **(Figure 5-3)**. One limitation of this comparison is that Ballantyne et al. could not report precise mutation rates for slowly mutating Y-STRs due to the number of meioses events examined in their study. As a result, their estimates were effectively restricted to a lower bound of $\mu = 10^{-3.5}$ mpg **(Figure 5-3, inset)**. In contrast, our deep phylogeny enabled us to accurately estimate much lower rates, highlighting the advantage of analyzing population data, rather than father-son pairs, for slowly mutating STRs. Comparing our estimates to those from Burgarella et al. resulted in an R^2 of $\sim$0.3, but restricting this evaluation to the subset of loci they characterized using more than 5000 father-son duos resulted in a substantially higher R^2 of 0.87 **(Supplemental Figure 5-13)**. These results demonstrate that our estimates are concordant with prior father-son based results, provided that the latter were generated using sufficiently many pairs.

5.3.3 Characteristics and determinants of Y-STR mutations

Next, we analyzed the STR mutation patterns. To obtain a single mutation rate estimate for each Y-STR, we averaged the estimates from the SGDP and 1000 Genomes datasets. We

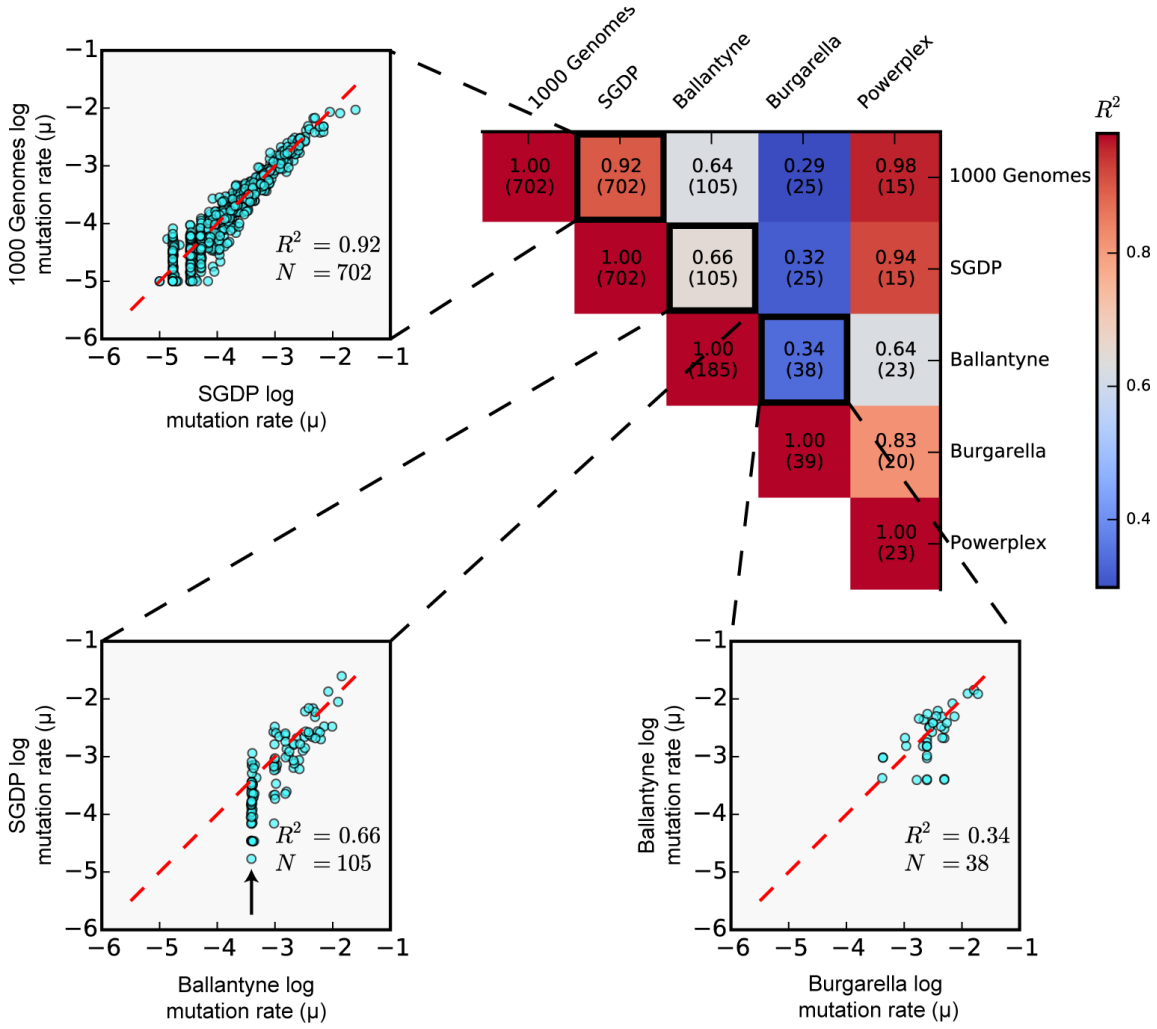

Figure 5-3: **Concordance of mutation rate estimates across datasets.** The heat map in the upper right corner presents the correlation of log mutation rates obtained from two father-son capillary-based studies ("Ballantyne" and "Burgarella") with those obtained in this study using the 1000 Genomes WGS data ("1000 Genomes"), the Simons Genome WGS data ("SGDP") and the capillary data available for samples in the 1000 Genomes ("Powerplex"). Each cell indicates the number of markers involved in the comparison and the resulting R^2. Representative scatterplots for three of these comparisons depict the pair of estimates for each marker (cyan) and the x = y line (red). The black arrow in the SGDP vs. Ballantyne comparison shows the effective lower limit of the Ballantyne et al. mutation rate estimates.

found that the distribution of Y-STR mutation rates has a substantial right tail, with most STRs mutating at very slow rates and only a few loci mutating at high rates **(Figure 5-4)**. On average, a polymorphic Y-STR mutates at a rate of $3.8 * 10^{-4}$ mpg and has a median mutation rate of $8.7 * 10^{-5}$ mpg. The average Y-STR mutation rate is an order of magnitude lower than previous estimates from panel-based studies. This difference cannot be explained by our phylogenetic measurement procedure since inspection of the same markers yielded relatively concordant numbers. Instead, it likely stems from the ascertainment strategy of STR panels, which select highly diverse loci that do not reflect the mutation rates of most STRs. One caveat in this analysis is that very long Y-STR markers were not accessible to Illumina reads. These loci might affect the calculated average mutation rate and, to a smaller extent, the median mutation rate. Consistent with these explanations, our mutation rate estimates for previously characterized loci were upwardly enriched relative to our estimates for all markers **(Figure 5-4)**.

Leveraging our Y-STR mutation rate catalog, we searched for loci with relatively high mutation rates. These loci help to distinguish Y-chromosomes of highly related individuals and can help to precisely date patrilineal relatedness among individuals, which is important for forensics and genetic genealogy. Most of the markers with the greatest estimated mutation rates have been characterized in prior studies **(Table 5.1)**, but we identified six loci whose mutation rates were estimated to be greater than $\sim 2 * 10^{-3}$ mpg and are yet to be reported **(Tables 5.2-5.3)**. Two of these markers, DYS548 and DYS467, have been used in previous genealogical panels but to the best of our knowledge, their mutation rates were never reported. In addition, we identified more than 65 loci with dinucleotide motifs and mutation rates greater than $\sim 3.33 * 10^{-4}$ mpg **(Table 5.3, Supplemental Table 5.4)**.

We observed wide variability in the mutation rates and patterns between motif length classes. STRs with tetranucleotide motifs had the greatest median mutation rate ($\mu = 1.76 * 10^{-4}$ mpg), followed by those with trinucleotide ($\mu = 1.22 * 10^{-4}$ mpg), pentanucleotide ($\mu = 1.19 * 10^{-4}$ mpg), dinucleotide ($\mu = 7.7 * 10^{-5}$ mpg), and hexanucleotide motifs ($\mu = 3.28 * 10^{-5}$ mpg) **(Figure 5-4)**. However, within each motif class, mutation rates varied by two or more orders of magnitude, indicating that other factors contribute to STR variability and highlighting that aggregate mutation rate statistics depend on the set of loci under consideration. We also found marked differences in the mutation patterns between motif classes. Loci with dinucleotide motifs and mutation rates greater than 10^{-4} mpg had a median step size parameter of $\rho_M = 0.8$, implying that many of the de novo mutations are expected to be greater than one repeat unit. On the other hand, the median step size parameter for longer motif classes within this mutation

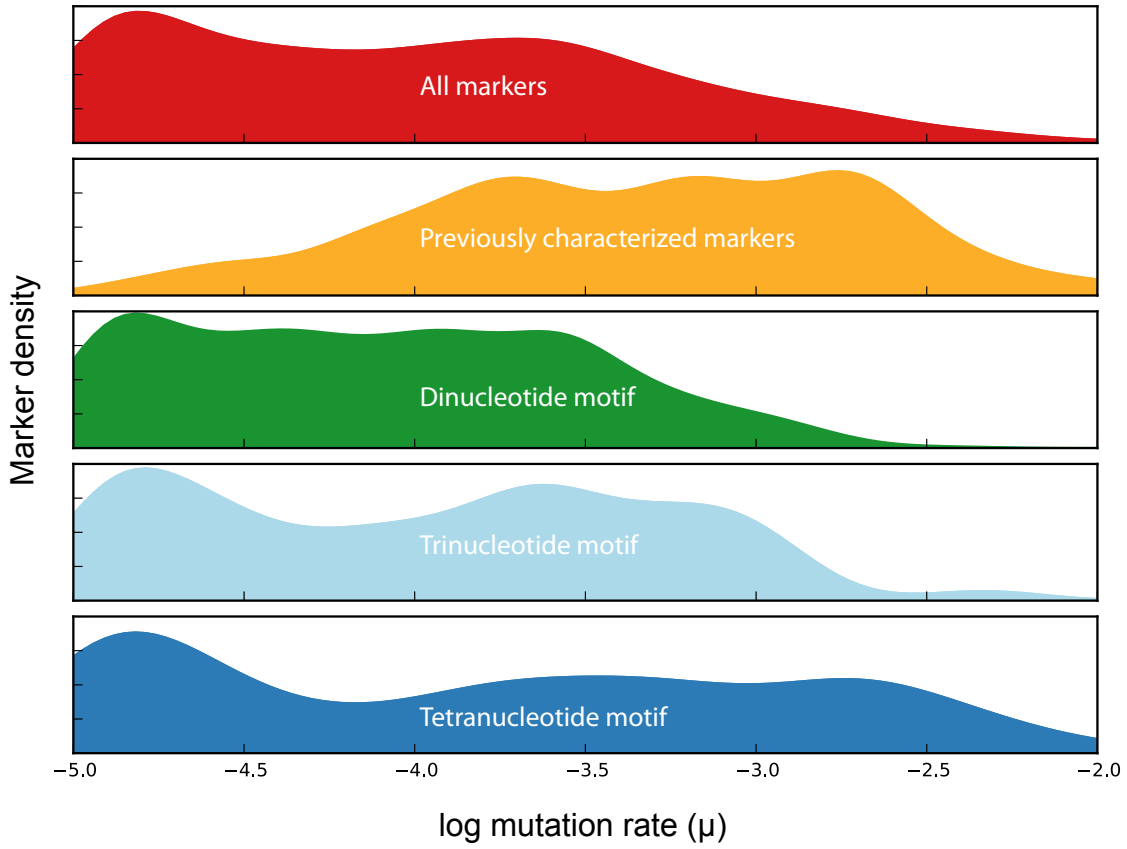

Figure 5-4: **Distribution of Y-STR mutation rates.** In red, we show the distribution of mutation rates across all STRs in this study. The set of loci with previously characterized mutation rates (orange) is substantially enriched for more mutable loci. When stratified by motif length, loci with tetranucleotide motifs (dark blue) are the most mutable, followed by loci with trinucleotide (light blue) and dinucleotide (green) motifs.

rate range was closer to one, implying that nearly all de novo events involve single step mutations.

Next, we harnessed the large number of Y-STR mutation rate estimates to identify the sequence determinants of mutation rates. For STRs without repeat structure interruptions, the length of the major allele explains a substantial fraction of the variance in log mutation rates for loci with di-, tri-, and tetranucleotide motifs ($R^2 = 0.83$, $R^2 = 0.67$, and $R^2 = 0.82$, respectively; pentanucleotide motifs were not assessed due to a small number of data points). However, when analyzing all STRs, including those with interruptions, the length of the major allele is

a poor predictor that explains only a modest amount of the variance ($R^2 = 0.16$, $R^2 = 0.25$, and $R^2 = 0.42$) **(Figure 5-5, left panels)**. To construct an improved model, we analyzed the relationship between the log mutation rate and the length of the longest uninterrupted repeat tract, regardless of the number of interruptions **(Figure 5-5, right panels)**. This model explained more than 75% of the variance in mutability for each of the three motif length classes. To assess the impact of the repeat motif on the mutation rate, we stratified loci with dinucleotide motifs by repeat sequence (AC, AG, or AT) and once again regressed the log mutation rate on the length of either the major allele or longest uninterrupted tract **(Supplemental Figure 5-14)**. Major allele length was again a relatively poor predictor of the log mutation rate, but uninterrupted tract length explained more than 80% of the variance for each motif. Although these motif-specific models improved the R^2, the increase was quite limited, suggesting that conditioned on the uninterrupted tract length, the repeat motif itself plays a minor role in the mutation rate. Taken together, our results show that a simple model of motif size and longest uninterrupted tract length largely explains STR mutation rates.

5.3.4 Predicting genome-wide STR mutation rates

We estimated the number of de novo mutations across the entire genome using the determinants found above. For each repeat motif length, we trained a non-linear mutation rate predictor using the uninterrupted tract lengths and mutation rates of the polymorphic Y-STRs. To account for the fixed STRs in our dataset and to better fit the model at shorter tract lengths, we assigned each fixed locus a mutation rate of 10^{-5} mpg, the lower mutation rate boundary used by MUTEA **(Supplemental Figure 5-15)**, and we jointly trained the predictors across all STRs. To validate these predictors, we used them to estimate the mutation rates of paternally transmitted autosomal CODIS markers, which the National Institute of Standards and Technology (NIST) has previously estimated using conventional means. Our predictors explained about 75% of the variance in the log mutation rates for these markers. In addition, the median mutation rate reported by NIST ($\mu = 1.3 * 10^{-3}$ mpg) closely matched the result reported by our predictors ($\mu = 1.0 * 10^{-3}$ mpg), suggesting that they generate reliable predictions.

Next, we ran our predictors on each STR in the human genome with 2-4 bp motifs, resulting in mutation rate estimates for each of the ~590,000 loci **(Supplemental Table 5.5)**. Since our model was trained using Y-STR mutation rates, these estimates refer only to the paternally inherited half of the genome. We discarded estimated rates below 1.2510^{-5} mpg, as these are

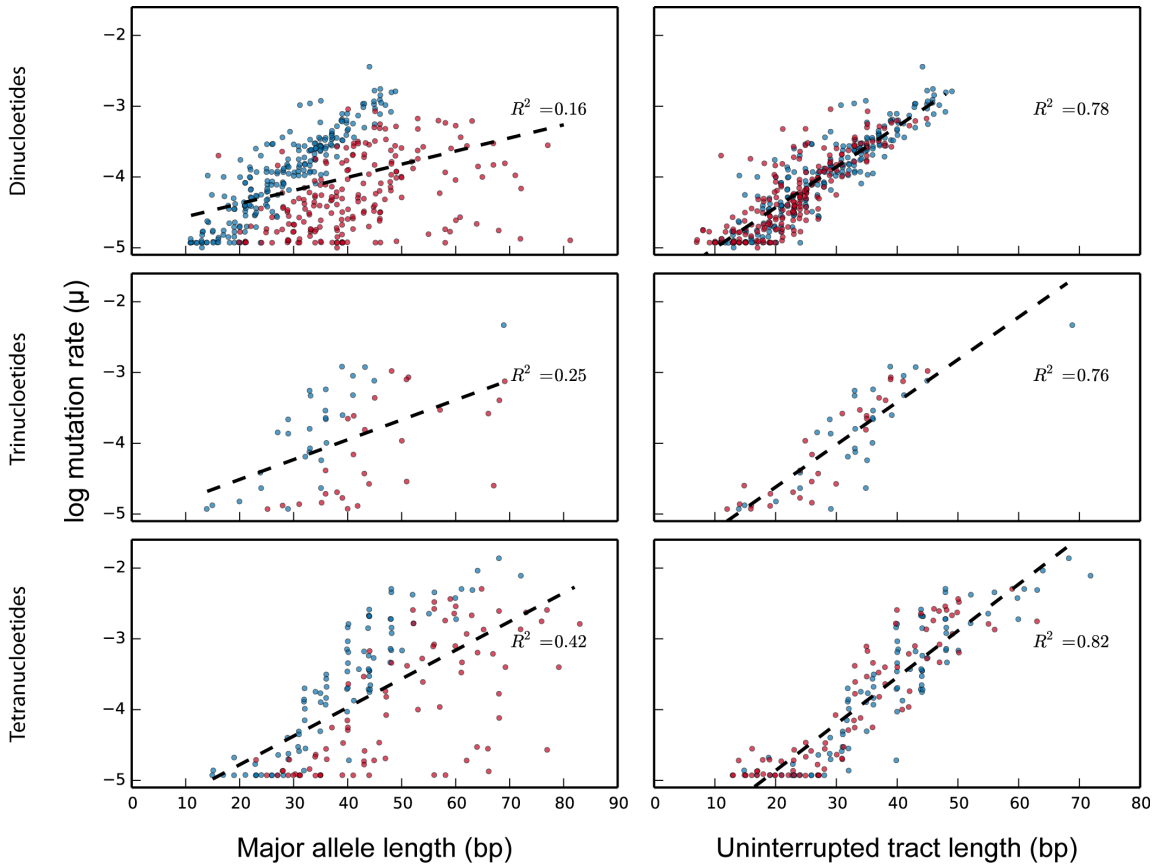

Figure 5-5: **Sequence determinants of Y-STR mutability.** Each panel presents the estimated log mutation rates (y-axis) of STRs versus either the major allele length (left panels, x-axis) or the longest uninterrupted tract length (right panels, x-axis) for various repeat motif sizes (rows). The black lines represent the mutation rate predicted by a simple linear model. For a given allele length (left panels), Y-STRs with no interruptions to the repeat structure (blue) are generally more mutable than those with one or more interruptions (red). Whereas major allele length alone is poorly correlated with mutation rate (left panels), the longest uninterrupted tract length (right panels) is strongly correlated regardless of the number of interruptions.

too close to the MUTEA lower boundary and may therefore be upwardly biased. After filtering, our model predicts that there are ∼70,000 STRs with mutation rates greater than 10^{-4} mpg, ∼44,000 loci with mutation rates greater than 1 in 3000 mpg and that an STR should mutate at an average rate of $4.4 * 10^{-4}$ mpg. Stratifying our results by motif length, we predict 29, 3, and 33 de novo STR mutations for loci with di-, tri- and tetranucleotide motifs on the paternally

inherited set of chromosomes.

Overall, we predict that 76-85 de novo STR mutations occur each generation for the full set of chromosomes. To account for the maternal chromosomes, we extrapolated our paternal results using prior estimates of the male to female STR mutation rate ratio (3.3:1 to 5.5: [22, 212]). We posit that our estimates for STR de novo mutational load are likely to be conservative. First, we omitted loci with 5-6 bp motifs for which we did not have sufficient data to build a mutation rate model. Second, for autosomal STRs whose uninterrupted tract lengths exceeded the maximal length observed in our study, we estimated their mutation rates using the maximal Y-STR length. Given the strong positive correlation between tract length and mutation rate observed in our study, these loci are probably far more mutable. Despite our conservative approach, the estimated number of genome-wide de novo STR mutations rivals that of any known class of genetic variation, including SNPs ($\sim$70 events per generation), indels (1-3 events), and SV and interspersed repeats ($<$1 event per generation) [58, 59, 60, 61]. As such, our results highlight the putative contribution of STRs to de novo genetic variation.

5.3.5 Y-STRs in forensics and genetic genealogy

We assessed the applicability of our Y-STR results to the genetic genealogy and forensic DNA communities. First, we considered whether it would be possible to distinguish between closely patrilineally related individuals from high-throughput sequencing data. Based on the entire Y-STR set reported by our study, we expect roughly one de novo mutation to occur every four generations. In addition, from WGS data, one also expects to identify approximately one de novo SNP every 2.85 generations [203], resulting in a 60% theoretical probability of differentiating between a father and son's Y-chromosome haplotype using high-throughput sequencing. Previous studies have suggested that capillary genotyping of 13 rapidly mutating Y-STRs can discriminate between father-son pairs in 20-27% of the cases [56, 213]. However, these particular markers are largely inaccessible to whole-genome sequencing data due to their long length and highly repetitive flanking regions that preclude unique mapping. With increased interest in high-throughput sequencing among genetic genealogy services (e.g. FullGenomes and Big Y by FamilyTreeDNA) and the forensics community, our results suggest that WGS can achieve better patrilineal discrimination compared to common panel-based methods. Of course, the main caveat is that WGS technology is at least an order of magnitude more expensive than a panel-based approach. However, if the current trajectory of sequencing cost decline continues,

shotgun sequencing to discriminate between closely patrilineally related individuals might soon become economically viable.

We also assessed the accuracy of imputing Y-STR profiles from Y-SNP data. This capability may be useful in forensic cases involving a highly degraded male sample for which complete Y-STR profiles would be difficult to obtain. In such cases, since there are many more SNPs than STRs on the Y-chromosome, it might be possible to salvage some of those markers with a high-throughput method and impute Y-STRs profiles for compatibility with common forensic or genealogical databases.

For imputation, we created a framework called MUTEA-IMPUTE. Briefly, after building a SNP phylogeny relating all samples and learning a mutation model as outlined in **Figure 5-1**, MUTEA-IMPUTE passes two sets of messages along the phylogeny to compute the exact marginal posteriors for each node, resulting in imputation probabilities for samples without observed Y-STR genotypes **(Supplemental Text 5.5.4)**. We assessed the accuracy of our algorithm by imputing the 1000 Genomes individuals for the PowerPlex Y23 panel, a set of markers regularly used in forensic cases involving sex crimes. Over 100 iterations, we randomly constructed reference panels of 500 samples and used MUTEA-IMPUTE to calculate the maximum a posteriori genotypes for a distinct set of 70 samples.

Despite the small size of the reference panel, we were able to correctly impute an average of 66% of the genotypes without any quality filtration **(Supplemental Table 5.6)**. Importantly, the resulting imputed probabilities roughly matched the average accuracy, indicating that the posteriors computed using this technique are well calibrated **(Supplemental Figure 5-16)**. Discarding imputed genotypes with posteriors below 70% resulted in an overall accuracy of 88% and retained about 40% of the calls. On a marker-by-marker basis, accuracy was generally inversely proportional to the estimated mutation rates, with the most slowly mutating markers having accuracies on the order of 95%. This trend stems from the fact that as the mutation rate increases, shorter branch lengths are required to obtain an estimate with similar confidence. We envision that a larger panel will substantially increase the ability to correctly impute Y-STRs and might facilitate work with highly degraded samples, a common issue in forensics casework.

5.4 Discussion

Advances in sequencing technology have fundamentally altered Y-STR analyses. The initial scarcity of SNP genotypes led to the development of methods to infer coalescent models from Y-STR genotypes alone. Methods designed to also learn STR mutational dynamics either marginalized over these coalescent models [214] or aimed to simultaneously infer the coalescent and mutational models [215, 216]. With the advent of population-scale WGS datasets, many of these STR-centric approaches have instead used SNPs, resulting in substantially more detailed phylogenies. For the Y-chromosome, these detailed phylogenies now provide the evolutionary context required to interpret Y-STR mutations, obviating the need for computationally expensive tree enumeration or marginalization approaches. However, the errors prevalent in WGS-based Y-STR genotypes require methods capable of accounting for genotype uncertainty, precluding the application of many traditional microsatellite distance measures designed for capillary data [13, 14].

In this study, we developed MUTEA, a method that leverages population-scale sequencing data to estimate Y-STR mutation rates. One inherent advantage of our approach is its ability to model and learn many of the salient features of microsatellite mutations. By incorporating a geometric step-size distribution, we allow both single-step mutations that predominate at tetranucleotide loci [22, 217] as well as multistep mutations that frequently occur at dinucleotide loci [22, 218]. In addition, the model's length constraint parameter captures the intra-locus phenomenon of shorter STR alleles preferentially expanding and longer alleles preferentially contracting [218, 100]. As these parameters are learned from observed STR genotypes, our method avoids many biases that stem from imposing single-step mutations or assuming parameters a priori.

In addition to its mutational model flexibility, our approach has both high throughput and a high dynamic range. With whole-genome sequencing data, we were able to assess every Y-STR that is accessible to Illumina sequencing, dramatically increasing the catalog of polymorphic loci with estimated mutation rates. In addition, by leveraging deep Y-chromosome phylogenies, we were able to obtain mutation rate estimates for very slowly mutating loci. Our estimates were highly replicable and consistent, as demonstrated by the strong concordance between the estimates from the two whole-genome sequencing datasets.

Our approach has several inherent limitations. Because Illumina datasets are currently comprised of 75-100 base pair reads, we were unable to genotype and characterize the mutation rates of

both long Y-STRs and Y-STRs that reside in heterochromatic regions. Due to the strong relationship between tract length and mutation rate, we anticipate that more rapidly mutating loci reside on the Y-chromosome. In addition, we were unable to characterize the mutation rates of homopolymers due to a rapid degradation of base quality scores with increasing allele length. As a result, future studies may benefit from reapplying our analyses as sequencing technologies, particularly those enabling longer reads, continue to mature. Another limitation is that our mutation model does not capture the full complexity of STR mutational dynamics, as it ignores intra-locus mutation rate variation [99]. Incorporating these and other mutational characteristics may be of interest to future studies.

One longstanding question regarding Y-STR mutation rates has been the apparent discrepancy between evolutionary and pedigree-based mutation rates. Several studies have suggested that evolutionary rates are 3-4 times lower, resulting in substantial inconsistencies in Y-STR-based lineage dating and large discrepancies from Y-SNP-based TMRCA estimates [195, 209, 219]. Because our study harnessed evolutionary data, we sought to avoid any potential issues by scaling each phylogeny such that our estimates best matched those from pedigree-based studies. Nonetheless, our investigations into an alternative scaling based on a SNP molecular clock resulted in similar scaling factors that only differed by $\sim$25%. Coupled with the strong concordance we observed with pedigree-based estimates, our study provides little evidence for a substantial difference between mutation rates estimated from these two types of data. Future work may benefit from assessing whether these previously reported discrepancies were due to the simplified Y-STR mutation models used in the approaches to obtain evolutionary-based Y-STR mutation rates.

Our large corpus of mutation rate estimates has enabled us to dissect the sequence factors governing STR mutability. We determined that the longest uninterrupted tract length is a strong predictor of the log mutation rate. This observation matches the exponential relationship between mutation rate and tract length previously reported in several pedigree-based studies [56, 212, 217, 100]. We also found that the total length of the major allele was a poor predictor. Coupled with the fact that Y-STRs without interruptions were much more mutable than interrupted ones with the same major allele length, our study provides strong evidence that interruptions to the repeat structure decrease mutation rates. This finding supports what has long been posited in STR evolutionary models [220, 221] and has been shown in a handful of small-scale experimental studies of STR mutability [222, 223]. However, it contradicts the recent findings of Ballantyne et al. in which no effect was observed [56].

Another open question is why STRs with dinucleotide motifs have lower mutation rates, given their higher levels of polymorphisms in the population. A previous large-scale panel-based study reported that loci with dinucleotide motifs have lower mutation rates than loci with tetranucleotide motifs [22]. Our survey confirmed this observation without ascertainment of STRs directly based on their polymorphism rates. However, genome-wide analyses of STRs have shown that dinucleotides have more diverse allelic spectra than tetranucleotides [27, 138]. These results are unlikely to be due to genotyping errors as a study of an individual sequenced to a depth of 120x also showed that dinucleotide repeats are more polymorphic than other types of STRs [27]. One potential explanation is that STRs with dinucleotide motifs have larger step sizes but lower mutation rates. However, we cannot exclude other explanations such as a difference in length constraint.

Our large compendium of mutation rate estimates has also enabled predictions about genome-wide STR variation. Prior studies have estimated a rate of approximately 75 de novo mutations per generation [184, 194] but have largely ignored STRs, despite their elevated mutation rates. Based on our projections for paternally inherited chromosomes, the number of de novo STR mutations is likely to rival the combined contribution of all other types of genetic variants. As several lines of evidence have highlighted the involvement of STR variations in complex traits [41, 93, 94, 173], it will be important to assess the biological impact of these de novo STR variations on human phenotypes.

5.5 Supplemental Text

5.5.1 Simulating exact STR genotypes

Values of μ, β, and ρ_M ranging from 10^{-5} to 10^{-2}, 0 to 0.5, and 0.75 to 1.0, respectively, were used to simulate genotypes under a wide range of mutation models. Using either the 1000 Genomes phylogeny or the SGDP phylogeny, each simulation was performed as follows:

1. Randomly assign the root node an STR allele between -4 and 4 and mark it as active
2. Remove an active node and mark it as inactive. For each of this node's children:
   (a) Calculate the child's allele probabilities using the branch length, the true mutation model and the parent node's genotype

(b) Randomly select an STR allele based on these probabilities

(c) Mark the descendant node as active

3. While active nodes remain, go to step 2
4. Report the exact STR alleles for a random subset of the samples (leaf nodes) based on the required sample size

5.5.2 Simulating STR sizes in reads with PCR stutter

We first used the procedure above to simulate STR genotypes down the phylogeny. We then used the true genotype for a particular sample g_i and a given stutter model to simulate the STR sizes observed in each read as follows:

1. Sample the number of observed reads $n_{reads,i}$ for each sample with genotype g_i from the read count distribution
2. For each read from 1 through $n_{reads,i}$, sample a number $c \sim U(0,1)$
   (a) If $c < d$, randomly sample an artifact size a_j from a geometric distribution with parameter ρ_s. Report the read's STR size as $g_i - a_j$
   (b) If $d \leq c < 1 - u$, report the read's STR size as g_i
   (c) Otherwise, randomly sample an artifact size a_j from a geometric distribution with parameter ρ_s. Report the read's STR size as $g_i + a_j$

To assess whether estimates would be accurate for even the most sparsely sequenced loci, we used read count distributions obtained from both Y-STR call sets corresponding to loci in the 10^{th} coverage percentile. For **Figure 2**, we used a stutter model with $d = 0.15$, $u = 0.01$ and $\rho_s = 0.8$, and we used 1, 2 and 3 reads for 65%, 25% and 10% of samples, respectively.

5.5.3 Confidence interval estimation

We used a delete-d jackknife approach to estimate mutation rate confidence intervals [224]. For each Y-STR, we sampled without replacement half of the STR genotypes a total of 100 times and estimated the log mutation rate using each of these subsets. Given these subsample estimates and the log estimate obtained using all samples, the standard error (SE) and confidence interval

(CI) for the log mutation rate were calculated according to:

$$SE = \sqrt{\frac{1}{100}\sum_{i=1}^{100}(\log \mu_i - \frac{1}{100}\sum_{j=1}^{100}\log \mu_j)^2}, \qquad CI = \log \mu_{tot} \pm 1.96 * SE$$

where μ_{tot} is the estimate based on the full dataset.

5.5.4 Y-STR imputation

We extended MUTEA to impute missing STR genotypes. Using the approach outlined in **Figure 1**, we first construct a phylogeny relating all samples and learn a mutation model. We then use this learned mutation model to pass two sets of messages along the tree and compute exact posteriors for each node, resulting in imputation probabilities for samples with missing genotypes. For node N_i with parent P_i, sibling S_i and children C_{1i} and C_{2i}, its conditional genotype probability given the observed data D is:

$$\begin{aligned} P(N_i|D) &= P(N_i|D_{C_{1i}}, D_{C_{2i}}, D_{-N_i}) = P(N_i, D_{C_{i1}}, D_{C_{2i}}|D_{-N_i})/P(D_{C_{i1}}, D_{C_{2i}}|D_{-N_i}) \\ &= P(N_i|D_{-N_i})P(D_{C_{1i}}, D_{C_{2i}}|N_i, D_{-N_i})/P(D_{C_{i1}}, D_{C_{2i}}|D_{-N_i}) \\ &\propto P(N_i|D_{-N_i})P(D_{C_{i1}}|N_i)P(D_{C_{2i}}|N_i) \end{aligned}$$

Here, D_{N_i} and D_{-N_i} denote the genotype likelihoods in and not in node N_i's subtree, respectively. We note that each of these terms is conditioned on the STR mutational model M and the Y-chromosome phylogeny T, but we omit these terms here and below for brevity.

The second and third terms in the node posterior expression are computed using a bottom-up traversal of the tree from the leaves to the root node. Each node in the tree combines information from its two children using the recurrence

$$\begin{aligned} P(D_{C_{1i}}|N_i) &= \sum_{a \in alleles} P(D_{C_{1i}}, C_{1i} = a|N_i) = \sum_a P(D_{GC_{1i}}, D_{GC_{2i}}, C_{1i} = a|N_i) \\ &= \sum_a P(C_{1i} = a|N_i)P(D_{GC_{i1}}|C_{1i} = a)P(D_{GC_{2i}}|C_{1i} = a) \end{aligned}$$

Here, GC_{1i} and GC_{2i} denote the two children of node C_{1i}. This recurrence applies to all nodes except the leaves, where genotype posteriors or a uniform prior are used for samples with and

without genotype information, respectively.

Similarly, the first term in the node posterior expression is computed using a top-down traversal of the tree from the root to the leaves. After assigning the root node a uniform prior probability, each node combines information from its parent and sibling:

$$\begin{aligned} P(N_i|D_{-N_i}) &= \sum_{a \in alleles} P(N_i, P_i = a|D_{S_i}, D_{-P_i}) = \frac{\sum_a P(N_i, P_i = a, D_{S_i}|D_{-P_i})}{P(D_{S_i}|D_{-P_i})} \\ &= \frac{\sum_a P(P_i = a|D_{-P_i}) P(D_{S_i}|P_i = a, D_{-P_i}) P(N_i|P_i = a, D_{S_i}, D_{-P_i})}{P(D_{S_i}|D_{-P_i})} \\ &\propto \sum_a P(P_i = a|D_{-P_i}) P(D_{S_i}|P_i = a) P(N_i|P_i = a) \end{aligned}$$

5.6 Acknowledgments

M.G. was supported by the National Defense Science and Engineering Graduate Fellowship. G.D.P. was supported by the National Science Foundation Graduate Research Fellowship under grant DGE-1147470. C.T.-S. was supported by The Wellcome Trust grant 098051. Y.E. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. This study was supported by NIJ grant 2014-DN-BX-K089 (Y.E. and T.W.). Y.E. is a SAB member of Identity Genomics, BigDataBio and Solve Inc. G.D.P is an employee of 23andMe. None of these entities played a role in the design, execution, interpretation, or presentation of this study.

5.7 Tables

Table 5.1: The most mutable Y-STRs with previously characterized mutation rates.

Chrom	Hg19 start	Hg19 end	Motif	Mutation rate (mpg)	Homogeneous tract length (bp)	Name
Y	7,053,359	7,053,426	AAAG	$1.37 * 10^{-2}$	68	DYS576
Y	7,867,880	7,867,943	AAAG	$9.20 * 10^{-3}$	64	DYS458
Y	6,861,231	6,861,298	AAAG	$7.80 * 10^{-3}$	72	DYS570
Y	14,515,312	14,515,363	AGAT	$5.08 * 10^{-3}$	48	DYS439
Y	8,426,378	8,426,443	AAG	$4.67 * 10^{-3}$	69	DYS481
Y	21,520,224	21,520,275	AGAT	$4.50 * 10^{-3}$	48	DYS549
Y	18,718,889	18,718,940	AGAT	$4.20 * 10^{-3}$	52	Y-GATA-A10
Y	4,270,960	4,271,019	AGAT	$3.77 * 10^{-3}$	60	DYS456
Y	19,372,273	19,372,328	AGAT	$2.88 * 10^{-3}$	48	DYS543
Y	14,761,101	14,761,160	AGAT	$2.65 * 10^{-3}$	46	DYS442

Table 5.2: The most mutable Y-STRs with tetranucleotide motifs and previously uncharacterized mutation rates.

Chrom	Hg19 start	Hg19 end	Motif	Mutation rate (mpg)	Homogeneous tract length (bp)	Name
Y	14,612,456	14,612,520	AGAT	$5.07 * 10^{-3}$	59	DYS467
Y	5,409,729	5,409,801	AAAG	$5.06 * 10^{-3}$	61	N/A
Y	19,500,594	19,500,656	AAAG	$4.89 * 10^{-3}$	63	N/A
Y	14,200,743	14,200,802	AGAT	$4.54 * 10^{-3}$	56	N/A
Y	21,665,702	21,665,764	AAAT	$3.66 * 10^{-3}$	50	DYS548

Table 5.3: The most mutable Y-STRs with dinucleotide motifs and previously uncharacterized mutation rates.

Chrom	Hg19 start	Hg19 end	Motif	Mutation rate (mpg)	Homogeneous tract length (bp)	Name
Y	2,807,025	2,807,064	AT	$3.62 * 10^{-3}$	44	N/A
Y	2,708,412	2,708,457	AG	$1.75 * 10^{-3}$	46	N/A
Y	3,832,234	3,832,278	AC	$1.66 * 10^{-3}$	45	N/A
Y	6,398,638	6,398,684	AC	$1.62 * 10^{-3}$	49	N/A
Y	17,109,092	17,109,141	AC	$1.57 * 10^{-3}$	48	N/A

5.8 Supplemental Figures

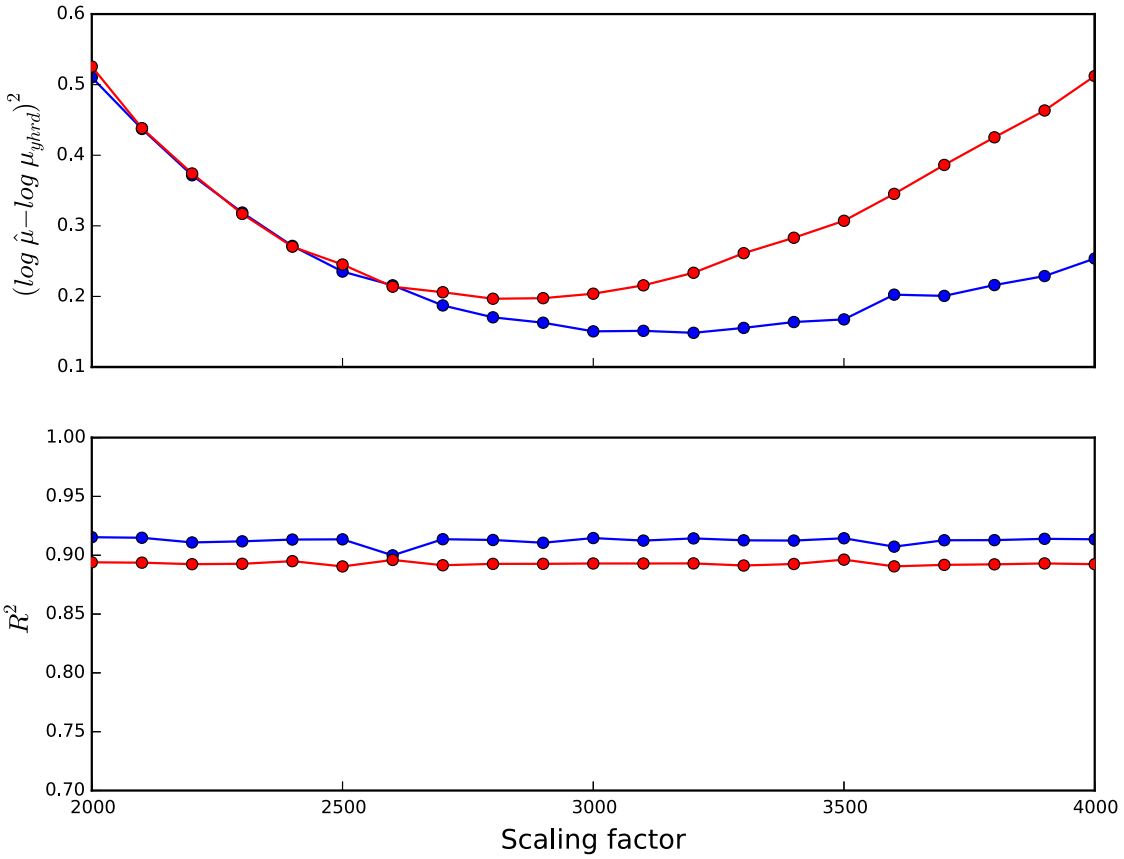

Figure 5-6: **Scaling the Y-SNP phylogenies.** Comparison of mutation rates for loci in the Y-Chromosome Haplotype Reference Database to estimates for the same loci obtained using data from the Simons Genome Project (blue) and the 1000 Genomes Project (red), over a range of scaling factors. Although the scaling factor had little effect on the R^2, it substantially impacted the total squared error in the log estimates. For each data set, we selected the optimal scaling factor as the value that minimized this squared error.

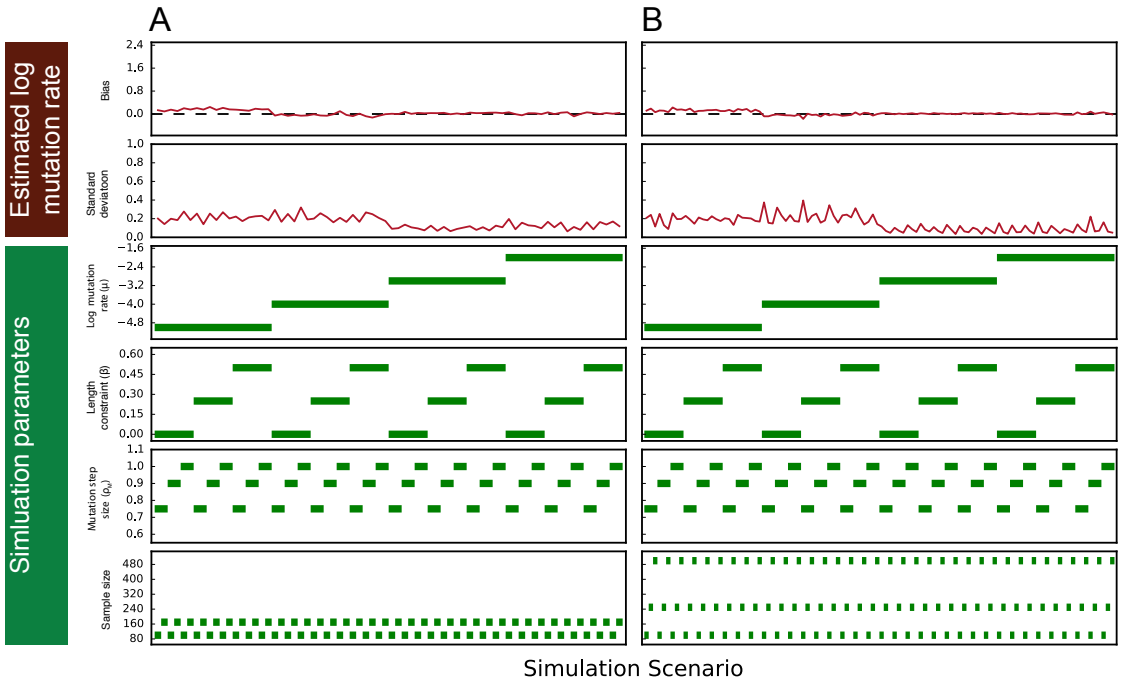

Figure 5-7: **MUTEA obtains accurate mutation rate estimates from exact genotypes.** STR genotypes were simulated for a variety of sample sizes and mutation models (four lower rows) for both the Simons Genomes phylogeny **(A)** and the 1000 Genomes phylogeny **(B)**. After assigning each sample's genotype a posterior probability of one and iterating 25 times for each simulation scenario, mutation rate estimates were unbiased (upper row) and have reasonably low standard deviations (second row).

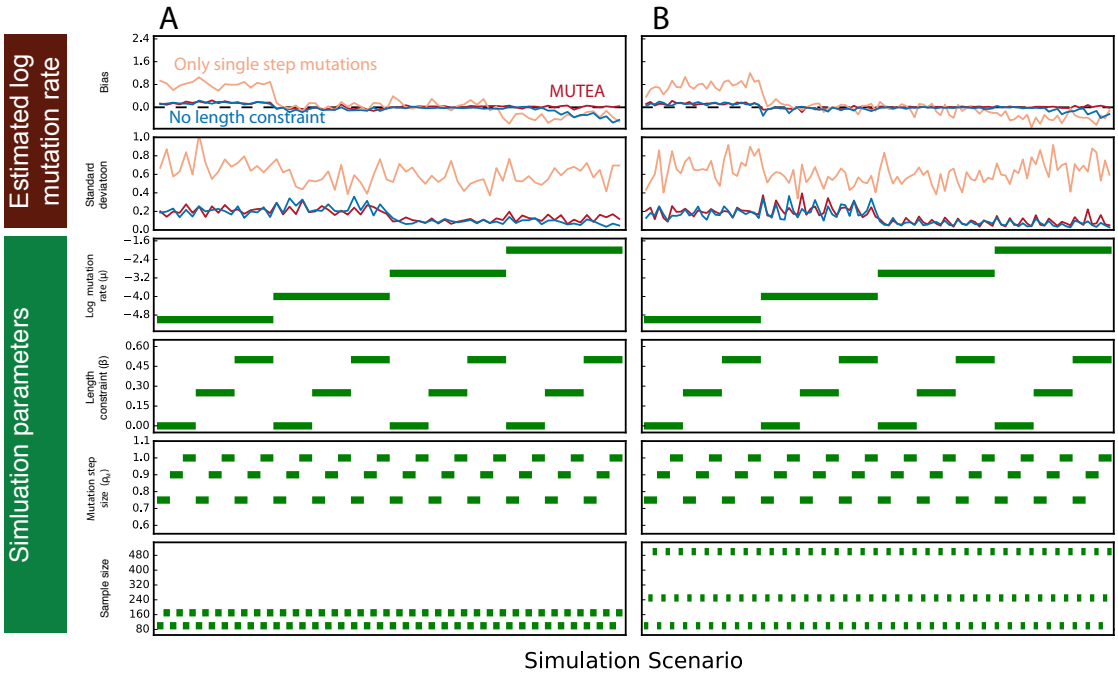

Figure 5-8: **Simplifying mutation models results in biased mutation rate estimates.** STR genotypes were simulated for a variety of sample sizes and mutation models (four lower rows) for both the Simons Genomes phylogeny **(A)** and the 1000 Genomes phylogeny **(B)**. After assigning each sample's genotype a posterior probability of one and iterating 25 times for each simulation scenario, mutation rate estimates were biased (upper row) when the estimated model is restricted to singe-step mutations (orange) or to no length constraint (blue), but not when the estimated model is unrestricted, as in MUTEA (red).

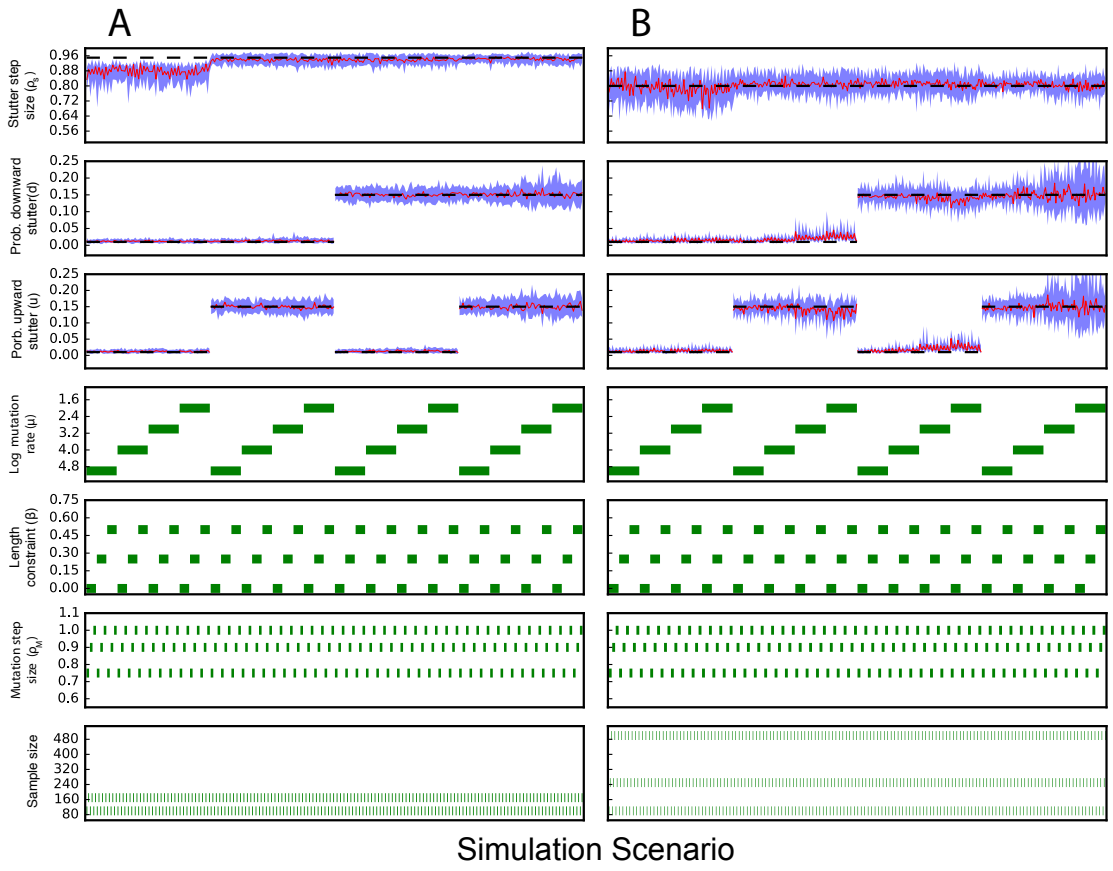

Figure 5-9: **MUTEA accurately recovers the underlying stutter model.** STR genotypes were simulated for a variety of sample sizes and mutation models (four lower rows). We then simulated observed reads for each set of genotypes using various PCR stutter models (dashed black lines in three upper rows) and input these simulated reads to MUTEA. Across 25 iterations of each scenario, the median inferred stutter parameters (red lines) were relatively unbiased. Blue lines indicate the lower and upper quartiles of the estimates for each scenario. **A.** 1, 2, 3, 4, 5 or 6 reads were generated for 19%, 27%, 21%, 15%, 8% and 10% of Simons Genome Diversity project samples, respectively. **B.** 1, 2 or 3 reads were generated for 65%, 25% and 10% of 1000 Genomes Project samples, respectively.

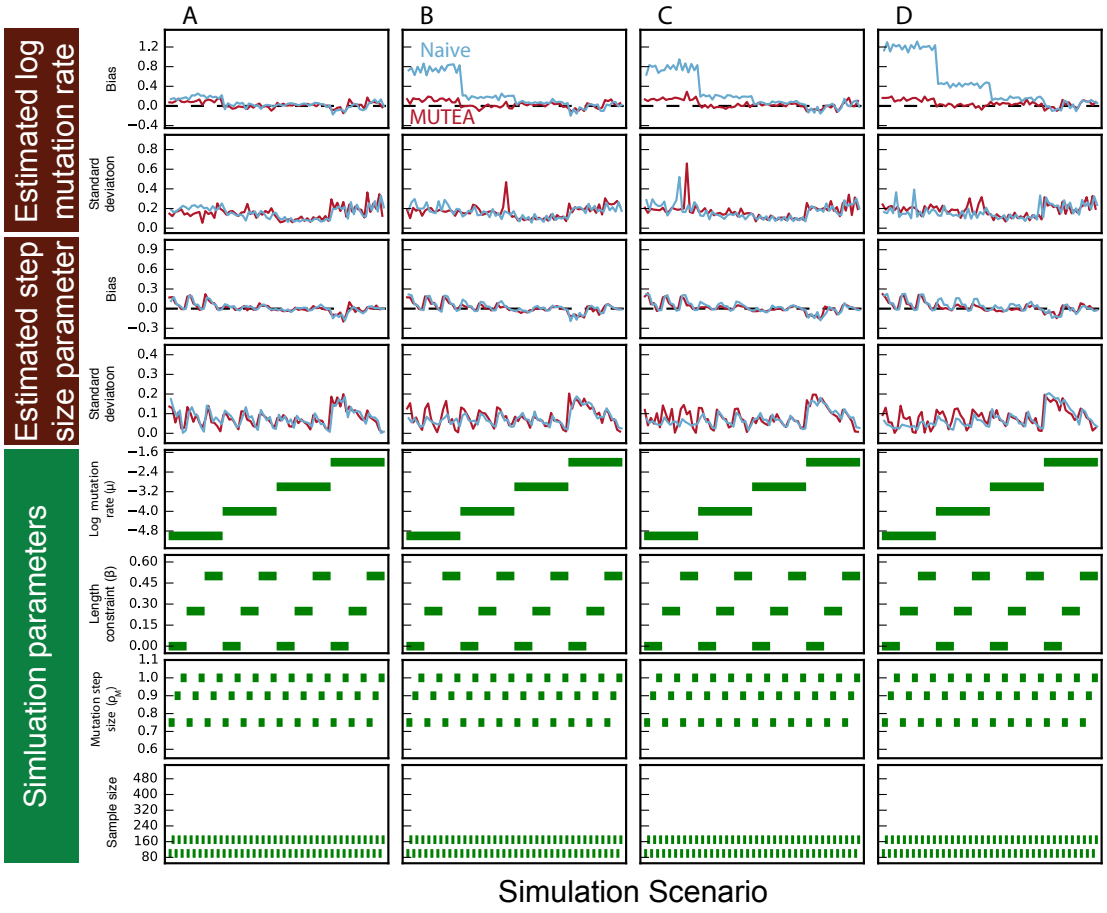

Figure 5-10: **MUTEA infers unbiased mutation rates and step size parameters from stutter-affected reads using the Simons Genome Diversity Project phylogeny.** STR genotypes were simulated for a variety of sample sizes and mutation models (four lower rows) using the Simons Genome phylogeny. Reads for each set of genotypes were then simulated using various PCR stutter models and input to MUTEA. Across 25 iterations for each scenario, MUTEA inferred unbiased estimates for the log mutation rate (red lines, upper row) and the step size parameter (third row). In contrast, a naive method that computes genotype posteriors based on the fraction of supporting reads resulted in biased mutation rate estimates (blue lines). For each simulation, 1, 2, 3, 4, 5, or 6 reads were simulated for 19%, 27%, 21%, 15%, 8%, and 10% of samples using a stutter model with $\rho_s = 0.95$ and **(A)** $d = 0.01$ and $u = 0.01$, **(B)** $d = 0.15$ and $u = 0.01$, **(C)** $d = 0.01$ and $u = 0.15$, or **(D)** $d = 0.15$ and $u = 0.15$.

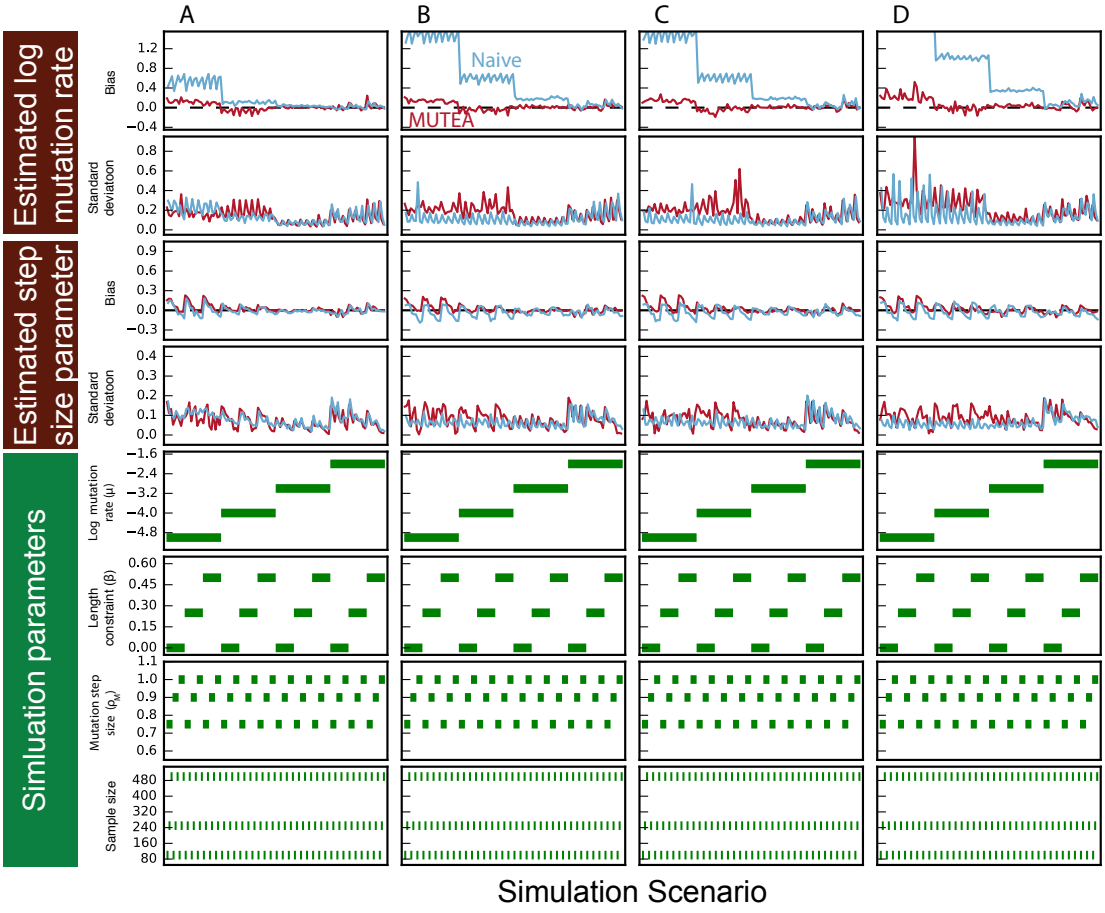

Figure 5-11: **MUTEA infers unbiased mutation rates and step size parameters from stutter-affected reads using the 1000 Genomes Project phylogeny.** STR genotypes were simulated for a variety of sample sizes and mutation models (four lower rows) using the 1000 Genome phylogeny. Reads for each set of genotypes were then simulated using various PCR stutter models and input to MUTEA. Across 25 iterations for each scenario, MUTEA inferred unbiased estimates for the log mutation rate (red lines, upper row) and the step size parameter (third row). In contrast, a naive method that computes genotype posteriors based on the fraction of supporting reads resulted in biased mutation rate estimates (blue lines). For each simulation, 1, 2, or 3 reads were generated for 65%, 25%, and 10% of samples using a stutter model with $\rho_s = 0.8$ and **(A)** $d = 0.01$ and $u = 0.01$, **(B)** $d = 0.15$ and $u = 0.01$, **(C)** $d = 0.01$ and $u = 0.15$, or **(D)** $d = 0.15$ and $u = 0.15$.

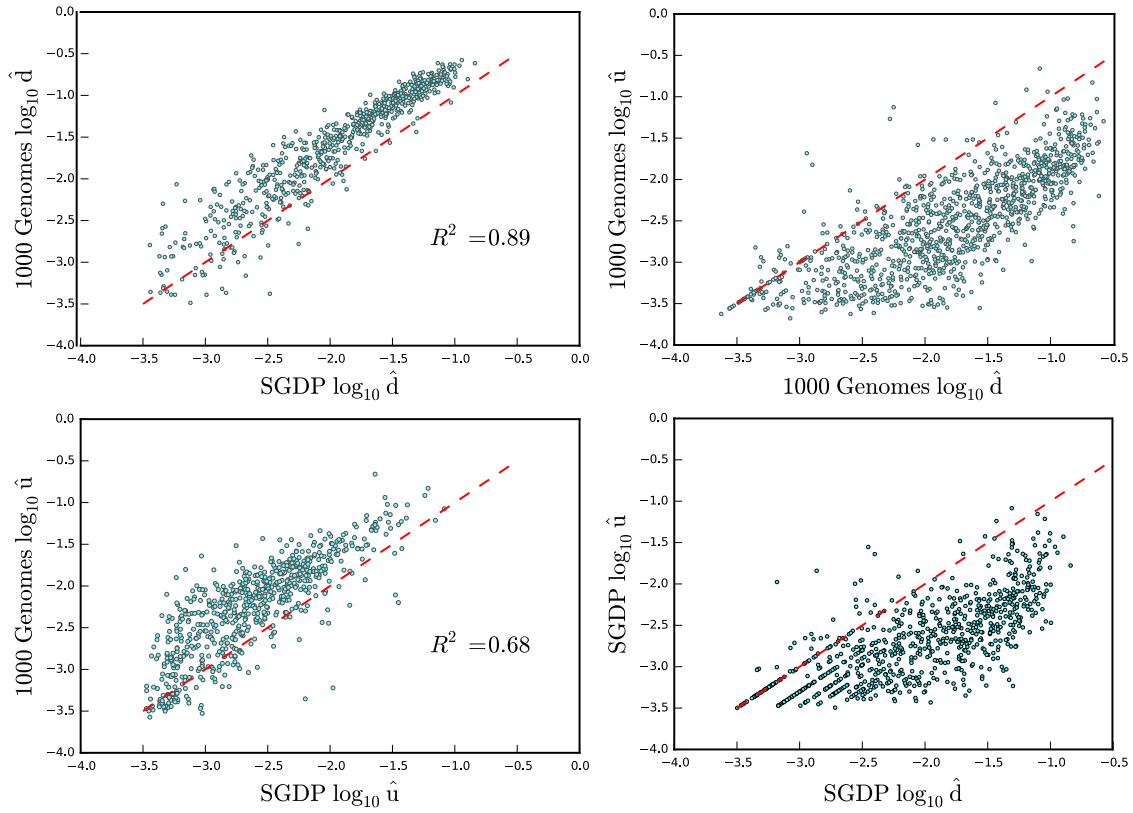

Figure 5-12: **Relationship between stutter probabilities within and across datasets.** For a given Y-STR locus, the probabilities of stutter increasing (u) or decreasing (d) the size of the STR in each read were highly correlated between input datasets (first column). However, the 1000 Genomes stutter rates largely fell above the diagonal (red line), indicating higher rates of stutter in this dataset. Within each dataset, nearly all loci had a higher rate of downward stutter than of upward stutter (second column).

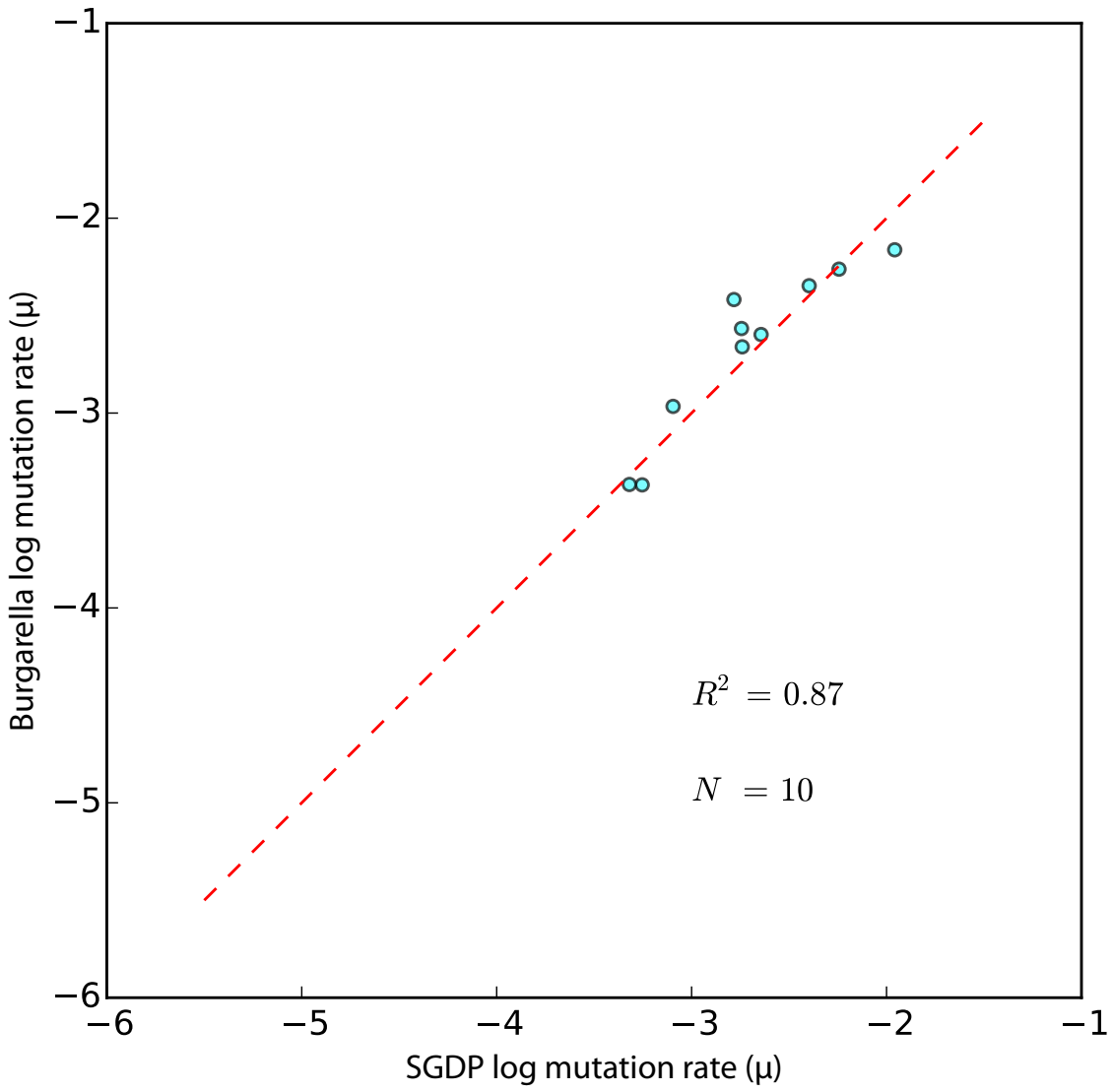

Figure 5-13: **SGDP estimates replicate those Burgarella estimates that were based on large numbers of father-son pairs.** The ten mutation rate estimates generated by Burgarella et al. using more than 5,000 father-son pairs are highly concordant with estimates we obtained using the SGDP data.

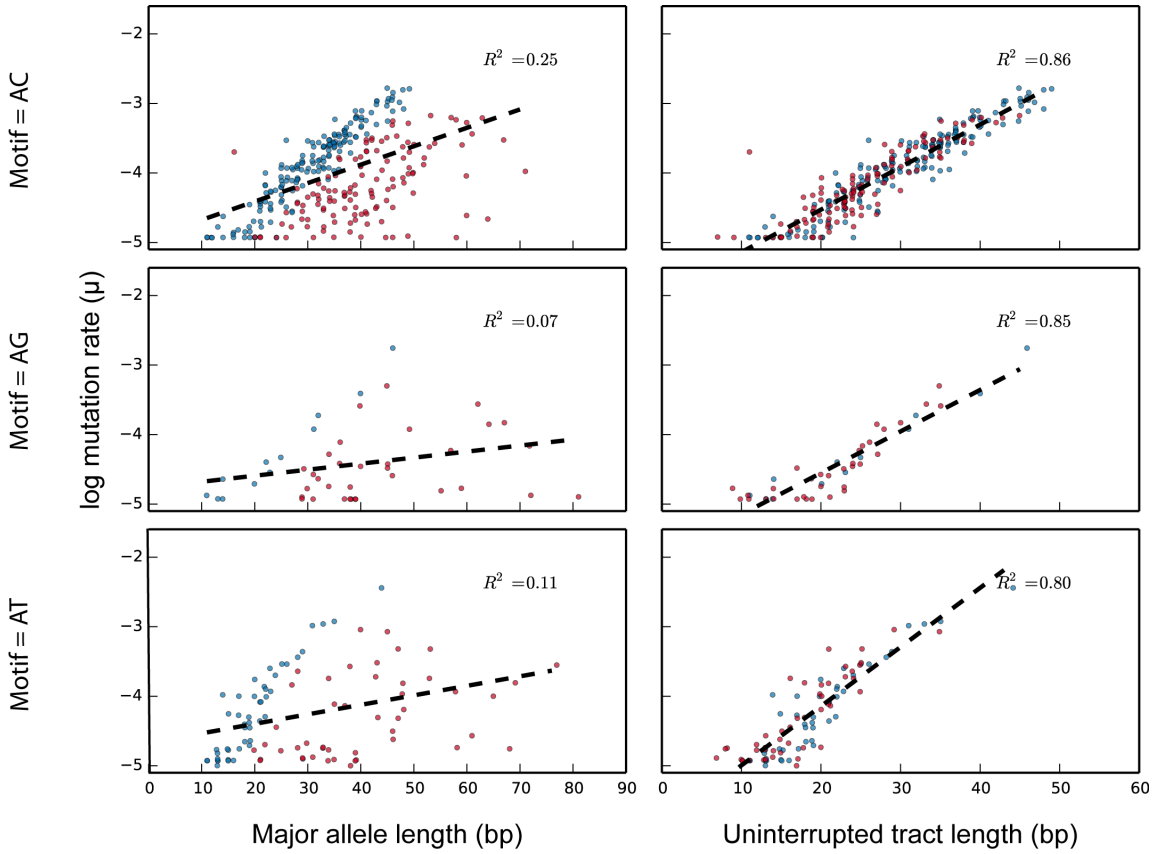

Figure 5-14: **Sequence determinants of Y-STR mutability for loci with dinucleotide repeat units.** Stratified by repeat motif (rows), loci with no interruptions to the repeat structure (blue) are generally more mutable than those with one or more interruptions (red). Whereas major allele length is a poor predictor of mutability (first column), the length of the longest uninterrupted tract is a very strong predictor of the log mutation rate for each motif length (second column).

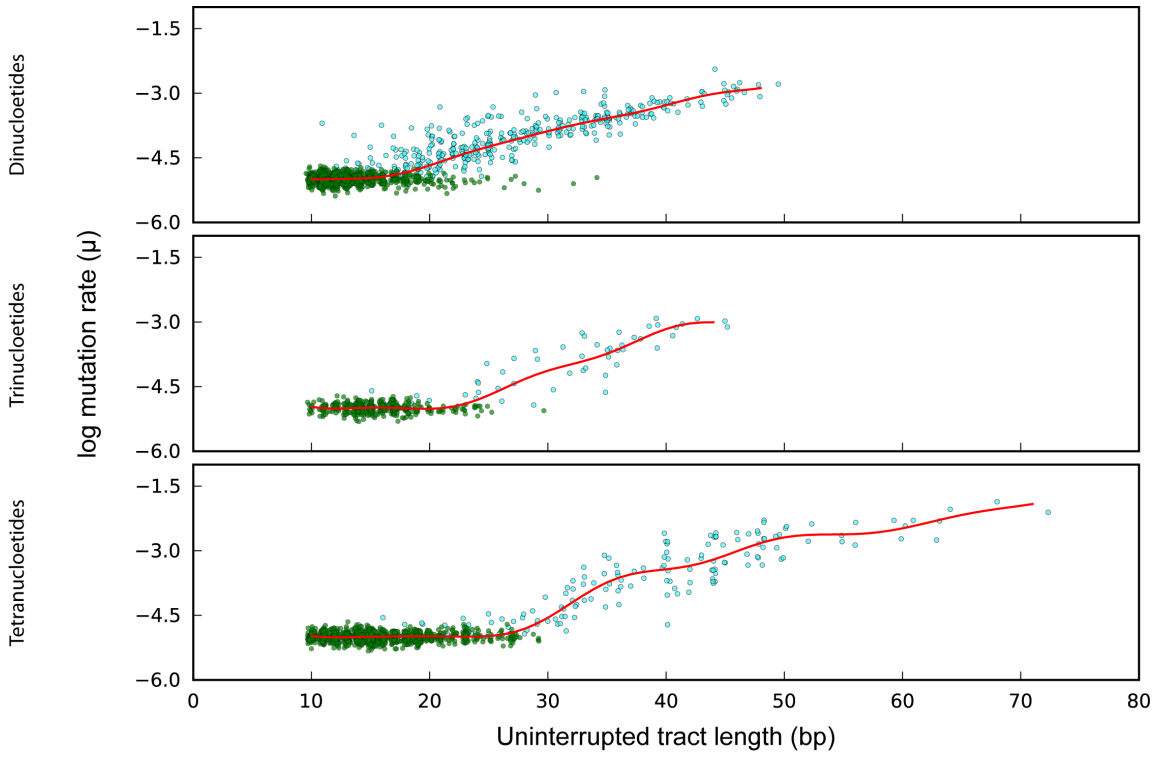

Figure 5-15: **Sequence-based predictors of Y-STR mutation rates.** For Y-STRs with di-, tri- and tetranucleotide motifs (rows), the mutation rates for polymorphic Y-STRs (cyan) and fixed Y-STRs (green) were used to fit predictive models of mutation rate (red). In general, the models predict a monotonic increase in log mutation rate with increasing uninterrupted tract lengths. Fixed Y-STRs were assigned a flat rate of 10^{-5} mpg and are displayed using jittered y-values to facilitate visualization.

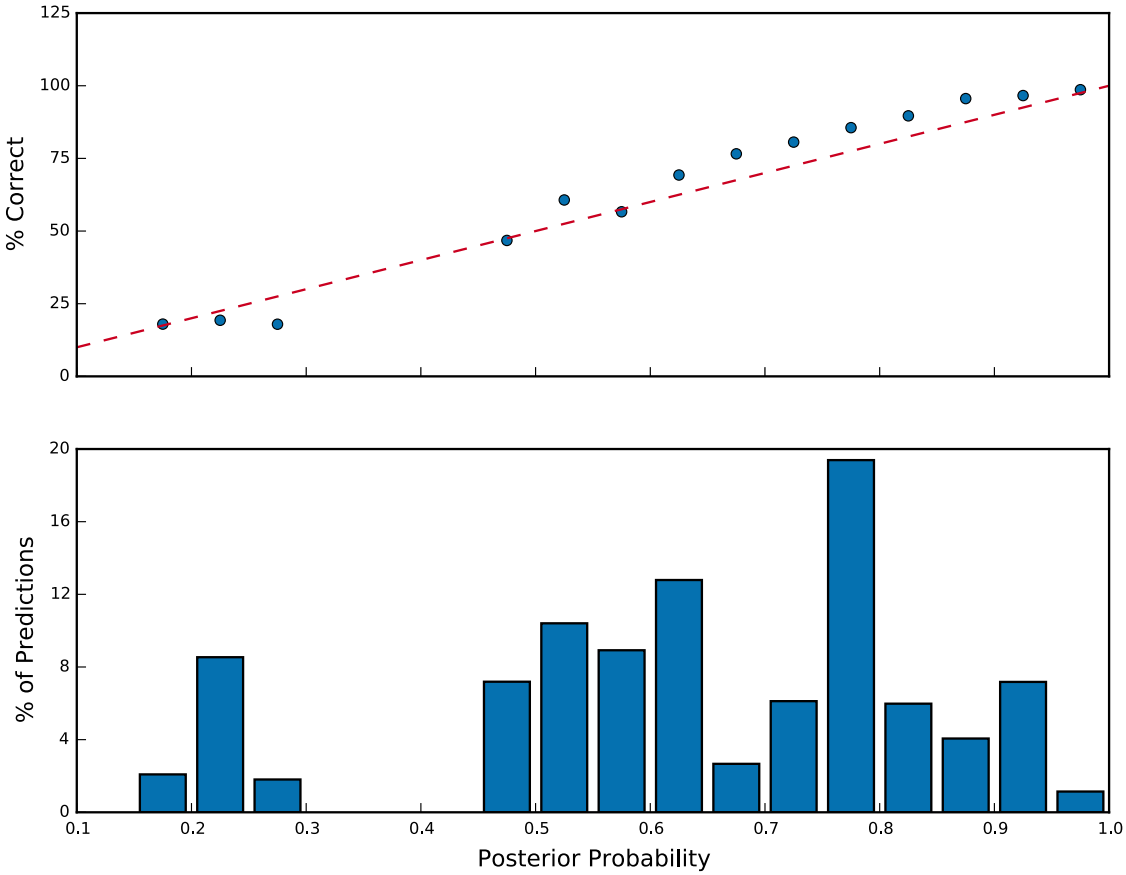

Figure 5-16: **MUTEA-IMPUTE produces well-calibrated imputation probabilities.** Y-STR genotypes for 1000 Genomes samples were imputed for loci in the PowerPlex Y23 panel across 1000 iterations, each using a reference panel of 500 samples and an imputation set of 70 samples. The accuracy for each posterior probability bin (upper panel) largely followed the diagonal (red line), demonstrating that the imputation probabilities reflect the true probability of correct imputation.

5.9 Supplemental Tables

Table 5.4: Estimated Y-STR mutation rates and stutter parameters

Please see the original publication

Table 5.5: Estimated mutation rates for each STR genome-wide

Please see the original publication

Table 5.6: Imputation accuracy for loci in the PowerPlex Y23 Panel.

		Posterior > 0%		Posterior > 70%	
Marker	μ (mpg)	% Calls	% Correct	%Calls	% Correct
DYS392	0.0006	100	93.1	96.4	94.5
DYS438	0.0006	100	93.1	95.4	95.1
DYS437	0.0007	100	92.7	95.0	94.3
DYS393	0.0014	100	82.5	76.9	87.8
DYS448	0.0017	100	81.9	80.2	89.4
DYS533	0.0017	100	78.1	74.2	85.5
DYS643	0.0019	100	80.0	78.1	86.3
DYS391	0.0022	100	73.5	54.8	80.1
Y-GATA-H4	0.0023	100	72.2	45.3	87.3
DYS390	0.0025	100	76.4	51.4	84.2
DYS385a	0.0025	100	74.3	64.5	87.9
DYS389I	0.0027	100	72.2	28.7	86.5
DYS19	0.0028	100	70.3	35.0	88.5
DYS635	0.0032	100	68.6	54.9	81.7
DYS456	0.0037	100	60.5	20.5	91.6
DYS549	0.0043	100	55.8	4.9	85.8
DYS439	0.0050	100	50.0	3.0	83.5
DYS481	0.0051	100	56.8	24.0	83.0
DYS385b	0.0052	100	53.7	17.2	87.2
DYS389II	0.0056	100	49.1	6.6	86.8
DYS458	0.0079	100	38.3	0.7	34.8
DYS570	0.0095	100	41.3	0.8	94.3
DYS576	0.0096	100	33.6	0.5	63.9
All		100	67.3	38.8	88.5

Chapter 6

The future of STRs

6.1 STRs: then and now

Throughout this thesis, I've described our efforts to characterize and address unknown aspects of STR variation and mutability. **Chapter 2** primarily sought to address two main questions: how many polymorphic STRs exist in the human genome and what drives their levels of variation? Based on low coverage data, we uncovered over 200,000 such STRs and identified their period, length and purity as the major drivers of variation. In **Chapter 3**, motivated by known examples of STR involvement in heritable diseases and single-gene studies, we used a genome-wide approach to assess their contribution to gene regulation. Our results suggest that STRs play a major role in shaping gene expression, providing an avenue by which they can contribute to complex traits. Lastly, in **Chapter 5**, we sought to extend the understanding of Y-STR mutation rates to the entire chromosome by leveraging population-scale data, resulting in mutation rates for hundreds of novel markers. The large number of mutation rate estimates enabled us to infer levels of STR de novo variation genome-wide and suggested that period, length and purity are also what determine STR mutation rates.

Collectively, these analyses highlight STRs as extremely polymorphic and mutable genetic variants with putative regulatory roles. Although our analyses have resulted in novel insights, they have also been hindered by the impact of STR genotyping errors. In **Chapter 4**, we described HipSTR, a novel state-of-the-art tool for genotyping STRs that dramatically increases genotyping accuracy. Coupled with rapidly progressing genomic datasets, we believe that future applications of HipSTR and other bioinformatics approaches may result in substantial insights. We highlight some of these potential applications in the sections that follow.

6.2 Refining the landscape of STR variation

In **Chapter 2**, we used low-coverage datasets from the 1000 Genomes Project to assess STR variation across a population of individuals. However, the quality of the sequencing data and lobSTR genotyping errors severely limited our ability to genotype longer STRs and may have introduced spurious alleles. Therefore, applying HipSTR to a population-scale high-coverage dataset will likely result in novel and more precise insights into the landscape of human STR variation. This analysis would also be beneficial for characterizing homopolymer repeats and the sequence variation within STRs. Furthermore, a dataset comprised of long reads would enable insights into the distribution of alleles at loci implicated in Mendelian diseases such as Fragile X syndrome and Huntington's disease [28].

6.3 STR imputation

Future studies may also benefit from developing and enhancing approaches for STR imputation. In **Chapter 5**, we outlined MUTEA-IMPUTE, which to our knowledge is the first method for Y-STR imputation. Using a reference panel of only 500 individuals, we were able to impute some of the most mutable Y-STRs with $\sim$90% accuracy. As the accuracy of our approach is largely governed by the time to the most recent common ancestor between an imputation sample and a reference sample, we believe that reassessing MUTEA-IMPUTE with a large reference panel of $\sim$10000 individuals will lead to substantial improvements in accuracy. Demonstrating this improvement would have profound implications for forensic and genealogical applications in which Y-STRs are heavily utilized [56, 17].

Developing approaches for autosomal STR imputation is also likely to be of interest but presents a much greater challenge. In **Chapter 4**, we explored the feasibility of imputing markers in the Marshfield panel using Beagle [176]. Though running Beagle was tractable in our study due to the small sample size and limited numbers of markers, it would not be practical to impute STRs genome-wide or in thousands of samples using this tool, necessitating the development of new methods. One promising avenue of research involves using the positional Burrows-Wheeler transform [225] to reconstruct an STR's phylogeny. While work in this area of research is still preliminary, these phylogenies would enable the application of MUTEA-IMPUTE to autosomal markers, resulting in substantially improved computational tractability and potentially more

accurate results.

6.4 STRs and complex traits

Together with preexisting knowledge about STRs involved in heritable diseases and single-gene studies, the findings in **Chapter 3** underscore the importance of assessing STR contributions to more complex traits. One potential avenue towards accomplishing this aim is to incorporate STRs into genome-wide association studies. As most large GWAS currently use SNP-chips to genotype individuals, STRs would need to be imputed before testing for association. While we have explored some promising preliminary avenues towards imputing STRs, further work will be required to render it computationally tractable. Alternatively, as whole-genome sequencing cohorts continue to grow in sample size and sequencing quality, it may be feasible to genotype STRs in these cohorts using tools like HipSTR. Genomics England's recently announced goal of sequencing 100,000 whole-genomes [172] suggests that such an approach may be feasible in the next 5 years. Coupled with the fact that WGS STR genotypes are far more accurate than their imputed counterparts, we believe that this provides the promising approach for identifying STRs involved in complex traits.

6.5 The promise of long reads

Throughout this thesis, we performed STR analyses using sequencing data obtained from the Illumina platform. This technology has largely dominated the sequencing market in the last 5 years as it has produced increasingly large volumes of reads with extremely low error rates. However, as explained in the **Introduction**, Illumina technology has primarily been limited to short 100 bp sequencing reads that have hindered our ability to characterize the variation and mutability of longer STRs.

Fortunately, major breakthroughs in sequencing technology may soon overcome these limitations. One particularly exciting development is the advent of single molecule real time sequencing (SMRT) and nanopore sequencing [226, 227]. In contrast to Illumina, in which clusters of amplified DNA are generated for each DNA segment, these technologies directly sequence the input DNA segments. The SMRT platform developed by Pacific Biosciences uses highly sensitive nanoscale detectors called zero-mode waveguides to monitor individual polymerase molecules as

they sequence DNA molecules [226, 50]. Using lasers and fluorescently tagged nucleotides, this technology generates real-time fluorescence profiles that are then used to infer the underlying bases. Nanopore sequencers also use a real-time approach to characterize DNA sequences, but they monitor the change in a nanopore's current as the DNA moves through it. As each DNA base and its surrounding bases induce different changes in current, the resulting current "squiggles" can also be used to infer the sequence of the DNA [227]. These technologies are particularly exciting because they can generate sequencing reads that are thousands of bases long.

Empowered by these technological advances, future studies will be able to characterize STRs that have been largely inaccessible to Illumina technology. Longer reads will be particularly invaluable in understanding the STRs involved in human diseases, as these typically involve large expansions. In addition, given the strong increase in STR mutation rates with allele length, longer reads may help uncover STRs that mutate every 100 generations, enabling the differentiation between father-son Y-STR haplotypes in sex crimes [213]. We anticipate that future studies with access to these datasets will be able to use MUTEA, our mutation rate estimation algorithm, to help identify these rapidly mutating markers.

6.6 De novo variation

Multiple aspects of the work outlined in this thesis suggest that STRs are a rich source of de novo variation. In **Chapter 4**, we used data from a deeply sequenced trio to identify hundreds of de novo mutations. While most of these mutations likely occurred within the cell line, preliminary analyses suggest that as many as 70 of them occurred within the germline. Similarly, genome-wide estimates of the STR mutational load in **Chapter 5** suggest that at least 70 de novo STR mutations occur every generation. Coupled with STRs' putative phenotypic contributions, we believe that future studies should quantify the levels of de novo STR variations using large cohorts of deeply sequenced trios. As has been done for SNPs, these analyses could uncover the effects of paternal and maternal age on de novo rates [58, 228] and lead to insights about the involvement of these mutations in psychiatric disorders [229] and other diseases.

6.7 Conclusion

Short tandem repeats are an extremely polymorphic and mutable class of genetic variant. Though originally widely used in human genetics, difficulties in characterizing STRs from high-throughput sequencing data have led to gaps in the understanding of STRs. Using population-scale sequencing datasets, we have sought to address these limitations by characterizing STR variability genome-wide, assessing their contribution to gene expression variation and estimating their de novo mutation rates. However, our analyses are just the tip of the iceberg, as many aspects of STRs remain understudied and poorly understood. Our hope is that the work outlined here, coupled with the described algorithms for accurately characterizing STRs, will motivate their inclusion in future large-scale genetic studies.

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

[1] C. International HapMap, "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, pp. 1299–320, 2005.

[2] C. Genomes Project, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[3] U. K. Consortium, K. Walter, J. L. Min, J. Huang, L. Crooks, Y. Memari, S. McCarthy, J. R. Perry, C. Xu, M. Futema, D. Lawson, V. Iotchkova, S. Schiffels, A. E. Hendricks, P. Danecek, R. Li, J. Floyd, L. V. Wain, I. Barroso, S. E. Humphries, M. E. Hurles, E. Zeggini, J. C. Barrett, V. Plagnol, J. B. Richards, C. M. Greenwood, N. J. Timpson, R. Durbin, and N. Soranzo, "The uk10k project identifies rare variants in health and disease," *Nature*, vol. 526, no. 7571, pp. 82–90, 2015.

[4] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson, "The nhgri gwas catalog, a curated resource of snp-trait associations," *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D1001–6, 2014.

[5] J. Yang, T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso, J. M. Cunningham, M. de Andrade, B. Feenstra, E. Feingold, M. G. Hayes, W. G. Hill, M. T. Landi, A. Alonso, G. Lettre, P. Lin, H. Ling, W. Lowe, R. A. Mathias, M. Melbye, E. Pugh, M. C. Cornelis, B. S. Weir, M. E. Goddard, and P. M. Visscher, "Genome partitioning of genetic variation for complex traits using common snps," *Nat Genet*, vol. 43, no. 6, pp. 519–25, 2011.

[6] P. R. Loh, G. Bhatia, A. Gusev, H. K. Finucane, B. K. Bulik-Sullivan, S. J. Pollack, C. Schizophrenia Working Group of Psychiatric Genomics, T. R. de Candia, S. H. Lee, N. R. Wray, K. S. Kendler, M. C. O'Donovan, B. M. Neale, N. Patterson, and A. L. Price, "Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis," *Nat Genet*, vol. 47, no. 12, pp. 1385–92, 2015.

[7] J. L. Weber and P. E. May, "Abundant class of human dna polymorphisms which can be typed using the polymerase chain reaction," *Am J Hum Genet*, vol. 44, no. 3, pp. 388–96, 1989.

[8] M. Litt and J. A. Luty, "A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene," *Am J Hum Genet*, vol. 44, no. 3, pp. 397–401, 1989.

[9] J. Weissenbach, G. Gyapay, C. Dib, A. Vignal, J. Morissette, P. Millasseau, G. Vays-

seix, and M. Lathrop, "A second-generation linkage map of the human genome," *Nature*, vol. 359, no. 6398, pp. 794–801, 1992.

[10] D. Botstein, R. L. White, M. Skolnick, and R. W. Davis, "Construction of a genetic linkage map in man using restriction fragment length polymorphisms," *Am J Hum Genet*, vol. 32, no. 3, pp. 314–31, 1980.

[11] H. Donis-Keller, P. Green, C. Helms, S. Cartinhour, B. Weiffenbach, K. Stephens, T. P. Keith, D. W. Bowden, D. R. Smith, E. S. Lander, and et al., "A genetic linkage map of the human genome," *Cell*, vol. 51, no. 2, pp. 319–37, 1987.

[12] A. M. Bowcock, A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd, and L. L. Cavalli-Sforza, "High resolution of human evolutionary trees with polymorphic microsatellites," *Nature*, vol. 368, no. 6470, pp. 455–7, 1994.

[13] M. Slatkin, "A measure of population subdivision based on microsatellite allele frequencies," *Genetics*, vol. 139, no. 1, pp. 457–62, 1995.

[14] D. B. Goldstein, A. Ruiz Linares, L. L. Cavalli-Sforza, and M. W. Feldman, "An evaluation of genetic distances for use with microsatellite loci," *Genetics*, vol. 139, no. 1, pp. 463–71, 1995.

[15] A. Edwards, A. Civitello, H. A. Hammond, and C. T. Caskey, "Dna typing and genetic mapping with trimeric and tetrameric tandem repeats," *Am J Hum Genet*, vol. 49, no. 4, pp. 746–56, 1991.

[16] B. Budowle, T. R. Moretti, S. J. Niezgoda, and B. L. Brown, "Codis and pcr-based short tandem repeat loci: law enforcement tools," vol. 1998, pp. 73–88, Promega Corporation, Madison, Wisconsin, 1998.

[17] M. Kayser, M. Vermeulen, H. Knoblauch, H. Schuster, M. Krawczak, and L. Roewer, "Relating two deep-rooted pedigrees from central germany by high-resolution y-str haplotyping," *Forensic Sci Int Genet*, vol. 1, no. 2, pp. 125–8, 2007.

[18] E. A. Foster, M. A. Jobling, P. G. Taylor, P. Donnelly, P. de Knijff, R. Mieremet, T. Zerjal, and C. Tyler-Smith, "Jefferson fathered slave's last child," *Nature*, vol. 396, no. 6706, pp. 27–8, 1998.

[19] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying personal genomes by surname inference," *Science*, vol. 339, no. 6117, pp. 321–4, 2013.

[20] K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, and J. L. Weber, "Comprehensive human genetic maps: individual and sex-specific variation in recombination," *Am J Hum Genet*, vol. 63, no. 3, pp. 861–9, 1998.

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

[22] 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*, vol. 44, no. 10, pp. 1161–5, 2012.

[23] J. L. Weber and C. Wong, "Mutation of human short tandem repeats," *Hum Mol Genet*, vol. 2, no. 8, pp. 1123–8, 1993.

[24] R. Chakraborty, M. Kimmel, D. N. Stivers, L. J. Davison, and R. Deka, "Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci," *Proc Natl Acad Sci U S A*, vol. 94, no. 3, pp. 1041–6, 1997.

[25] D. Bachtrog, M. Agis, M. Imhof, and C. Schlotterer, "Microsatellite variability differs between dinucleotide repeat motifs-evidence from drosophila melanogaster," *Mol Biol Evol*, vol. 17, no. 9, pp. 1277–85, 2000.

[26] Y. D. Kelkar, S. Tyekucheva, F. Chiaromonte, and K. D. Makova, "The genome-wide determinants of human and chimpanzee microsatellite evolution," *Genome Res*, vol. 18, no. 1, pp. 30–8, 2008.

[27] M. Gymrek, D. Golan, S. Rosset, and Y. Erlich, "lobstr: A short tandem repeat profiler for personal genomes," *Genome Res*, vol. 22, no. 6, pp. 1154–62, 2012.

[28] S. M. Mirkin, "Expandable dna repeats and human disease," *Nature*, vol. 447, no. 7147, pp. 932–40, 2007.

[29] "A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. the huntington's disease collaborative research group," *Cell*, vol. 72, no. 6, pp. 971–83, 1993.

[30] A. J. Verkerk, M. Pieretti, J. S. Sutcliffe, Y. H. Fu, D. P. Kuhl, A. Pizzuti, O. Reiner, S. Richards, M. F. Victoria, F. P. Zhang, and et al., "Identification of a gene (fmr-1) containing a cgg repeat coincident with a breakpoint cluster region exhibiting length variation in fragile x syndrome," *Cell*, vol. 65, no. 5, pp. 905–14, 1991.

[31] G. Bates, "Huntingtin aggregation and toxicity in huntington's disease," *Lancet*, vol. 361, no. 9369, pp. 1642–4, 2003.

[32] B. M. Davis, M. E. McCurrach, K. L. Taneja, R. H. Singer, and D. E. Housman, "Expansion of a cug trinucleotide repeat in the 3' untranslated region of myotonic dystrophy protein kinase transcripts results in nuclear retention of transcripts," *Proc Natl Acad Sci U S A*, vol. 94, no. 14, pp. 7388–93, 1997.

[33] F. Tassone, R. J. Hagerman, A. K. Taylor, L. W. Gane, T. E. Godfrey, and P. J. Hagerman, "Elevated levels of fmr1 mrna in carrier males: a new mechanism of involvement in the fragile-x syndrome," *Am J Hum Genet*, vol. 66, no. 1, pp. 6–15, 2000.

[34] K. Sathasivam, A. Neueder, T. A. Gipson, C. Landles, A. C. Benjamin, M. K. Bondulich, D. L. Smith, R. L. Faull, R. A. Roos, D. Howland, P. J. Detloff, D. E. Housman, and G. P.

Bates, "Aberrant splicing of htt generates the pathogenic exon 1 protein in huntington disease," *Proc Natl Acad Sci U S A*, vol. 110, no. 6, pp. 2366–70, 2013.

[35] M. T. Raijmakers, P. L. Jansen, E. A. Steegers, and W. H. Peters, "Association of human liver bilirubin udp-glucuronyltransferase activity with a polymorphism in the promoter region of the ugt1a1 gene," *Journal of hepatology*, vol. 33, no. 3, pp. 348–351, 2000.

[36] G. Monaghan, M. Ryan, R. Hume, B. Burchell, and R. Seddon, "Genetic variation in bilirubin udp-glucuronosyltransferase gene promoter and gilbert's syndrome," *The Lancet*, vol. 347, no. 9001, pp. 578–581, 1996.

[37] K. J. Verstrepen, A. Jansen, F. Lewitter, and G. R. Fink, "Intragenic tandem repeats generate functional variability," *Nat Genet*, vol. 37, no. 9, pp. 986–90, 2005.

[38] M. D. Vinces, M. Legendre, M. Caldara, M. Hagihara, and K. J. Verstrepen, "Unstable tandem repeats in promoters confer transcriptional evolvability," *Science*, vol. 324, no. 5931, pp. 1213–6, 2009.

[39] r. Fondon, J. W. and H. R. Garner, "Molecular origins of rapid and continuous morphological evolution," *Proc Natl Acad Sci U S A*, vol. 101, no. 52, pp. 18058–63, 2004.

[40] S. Sureshkumar, M. Todesco, K. Schneeberger, R. Harilal, S. Balasubramanian, and D. Weigel, "A genetic defect caused by a triplet repeat expansion in arabidopsis thaliana," *Science*, vol. 323, no. 5917, pp. 1060–3, 2009.

[41] A. Contente, A. Dittmer, M. C. Koch, J. Roth, and M. Dobbelstein, "A polymorphic microsatellite that mediates induction of pig3 by p53," *Nat Genet*, vol. 30, no. 3, pp. 315–20, 2002.

[42] 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*, vol. 101, no. 10, pp. 3504–9, 2004.

[43] 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*, vol. 47, no. 9, pp. 1073–8, 2015.

[44] T. Lappalainen, M. Sammeth, M. R. Friedlander, P. A. t Hoen, J. Monlong, M. A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlof, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Don-

nelly, M. I. McCarthy, P. Flicek, T. M. Strom, C. Geuvadis, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Hasler, A. C. Syvanen, G. J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigo, I. G. Gut, X. Estivill, and E. T. Dermitzakis, "Transcriptome and genome sequencing uncovers functional variation in humans," *Nature*, vol. 501, no. 7468, pp. 506–11, 2013.

[45] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavare, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis, "Relative impact of nucleotide and copy number variation on gene expression phenotypes," *Science*, vol. 315, no. 5813, pp. 848–53, 2007.

[46] J. L. Weber and K. W. Broman, "Genotyping for human whole-genome scans: past, present, and future," *Adv Genet*, vol. 42, pp. 77–96, 2001.

[47] J. M. Butler, "Genetics and genomics of core short tandem repeat loci used in human identity testing," *J Forensic Sci*, vol. 51, no. 2, pp. 253–65, 2006.

[48] M. Gymrek, "Pcr-free library preparation greatly reduces stutter noise at short tandem repeats," *bioRxiv*, 2016.

[49] C. 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*, vol. 526, no. 7571, pp. 68–74, 2015.

[50] E. R. Mardis, "Next-generation sequencing platforms," *Annual review of analytical chemistry*, vol. 6, pp. 287–303, 2013.

[51] M. L. Metzker, "Sequencing technologiesâĂŤthe next generation," *Nature reviews genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[52] G. Highnam, C. Franck, A. Martin, C. Stephens, A. Puthige, and D. Mittelman, "Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles," *Nucleic Acids Res*, vol. 41, no. 1, p. e32, 2013.

[53] 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*, vol. 43, no. 5, pp. 491–8, 2011.

[54] A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S. R. Twigg, W. G. S. Consortium, A. O. Wilkie, G. McVean, and G. Lunter, "Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications," *Nat Genet*, vol. 46, no. 8, pp. 912–8, 2014.

[55] L. Roewer, "Y chromosome str typing in crime casework," *Forensic Sci Med Pathol*, vol. 5,

no. 2, pp. 77–84, 2009.

[56] K. N. Ballantyne, M. Goedbloed, R. Fang, O. Schaap, O. Lao, A. Wollstein, Y. Choi, K. van Duijn, M. Vermeulen, S. Brauer, R. Decorte, M. Poetsch, N. von Wurmb-Schwark, P. de Knijff, D. Labuda, H. Vezina, H. Knoblauch, R. Lessig, L. Roewer, R. Ploski, T. Dobosz, L. Henke, J. Henke, M. R. Furtado, and M. Kayser, "Mutability of y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications," *Am J Hum Genet*, vol. 87, no. 3, pp. 341–53, 2010.

[57] C. Burgarella and M. Navascues, "Mutation rate estimates for 110 y-chromosome strs combining population and father-son pair data," *Eur J Hum Genet*, vol. 19, no. 1, pp. 70–5, 2011.

[58] A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, W. S. Wong, G. Sigurdsson, G. B. Walters, S. Steinberg, H. Helgason, G. Thorleifsson, D. F. Gudbjartsson, A. Helgason, O. T. Magnusson, U. Thorsteinsdottir, and K. Stefansson, "Rate of de novo mutations and the importance of father's age to disease risk," *Nature*, vol. 488, no. 7412, pp. 471–5, 2012.

[59] R. Rahbari, A. Wuster, S. J. Lindsay, R. J. Hardwick, L. B. Alexandrov, S. Al Turki, A. Dominiczak, A. Morris, D. Porteous, B. Smith, M. R. Stratton, U. K. Consortium, and M. E. Hurles, "Timing, rates and spectra of human germline mutation," *Nat Genet*, vol. 48, no. 2, pp. 126–33, 2016.

[60] A. Itsara, H. Wu, J. D. Smith, D. A. Nickerson, I. Romieu, S. J. London, and E. E. Eichler, "De novo rates and selection of large copy number variation," *Genome Res*, vol. 20, no. 11, pp. 1469–81, 2010.

[61] W. P. Kloosterman, L. C. Francioli, F. Hormozdiari, T. Marschall, J. Y. Hehir-Kwa, A. Abdellaoui, E. W. Lameijer, M. H. Moed, V. Koval, I. Renkens, M. J. van Roosmalen, P. Arp, L. C. Karssen, B. P. Coe, R. E. Handsaker, E. D. Suchiman, E. Cuppen, D. T. Thung, M. McVey, M. C. Wendl, C. Genome of Netherlands, A. Uitterlinden, C. M. van Duijn, M. A. Swertz, C. Wijmenga, G. B. van Ommen, P. E. Slagboom, D. I. Boomsma, A. Schonhuth, E. E. Eichler, P. I. de Bakker, K. Ye, and V. Guryev, "Characteristics of de novo structural changes in the human genome," *Genome Res*, vol. 25, no. 6, pp. 792–801, 2015.

[62] H. Ellegren, "Microsatellites: simple sequences with complex evolution," *Nat Rev Genet*, vol. 5, no. 6, pp. 435–45, 2004.

[63] C. E. Pearson, K. Nichol Edamura, and J. D. Cleary, "Repeat instability: mechanisms of dynamic mutations," *Nat Rev Genet*, vol. 6, no. 10, pp. 729–42, 2005.

[64] B. Brais, J. P. Bouchard, Y. G. Xie, D. L. Rochefort, N. Chretien, F. M. Tome, R. G. Lafreniere, J. M. Rommens, E. Uyama, O. Nohira, S. Blumen, A. D. Korczyn, P. Heutink,

J. Mathieu, A. Duranceau, F. Codere, M. Fardeau, and G. A. Rouleau, "Short gcg expansions in the pabp2 gene cause oculopharyngeal muscular dystrophy," *Nat Genet*, vol. 18, no. 2, pp. 164–7, 1998.

[65] J. Amiel, D. Trochet, M. Clement-Ziza, A. Munnich, and S. Lyonnet, "Polyalanine expansions in human," *Hum Mol Genet*, vol. 13 Spec No 2, pp. R235–43, 2004.

[66] R. Gemayel, M. D. Vinces, M. Legendre, and K. J. Verstrepen, "Variable tandem repeats accelerate evolution of coding and regulatory sequences," *Annu Rev Genet*, vol. 44, pp. 445–77, 2010.

[67] L. A. Sawyer, J. M. Hennessy, A. A. Peixoto, E. Rosato, H. Parkinson, R. Costa, and C. P. Kyriacou, "Natural variation in a drosophila clock gene and temperature compensation," *Science*, vol. 278, no. 5346, pp. 2117–20, 1997.

[68] S. A. Tishkoff, R. Varkonyi, N. Cahinhinan, S. Abbes, G. Argyropoulos, G. Destro-Bisol, A. Drousiotou, B. Dangerfield, G. Lefranc, J. Loiselet, A. Piro, M. Stoneking, A. Tagarelli, G. Tagarelli, E. H. Touma, S. M. Williams, and A. G. Clark, "Haplotype diversity and linkage disequilibrium at human g6pd: recent origin of alleles that confer malarial resistance," *Science*, vol. 293, no. 5529, pp. 455–62, 2001.

[69] M. Kayser and P. de Knijff, "Improving human forensics through advances in genetics, genomics and molecular biology," *Nat Rev Genet*, vol. 12, no. 3, pp. 179–92, 2011.

[70] R. Khan and D. Mittelman, "Rumors of the death of consumer genomics are greatly exaggerated," *Genome Biol*, vol. 14, no. 11, p. 139, 2013.

[71] G. Tamiya, M. Shinya, T. Imanishi, T. Ikuta, S. Makino, K. Okamoto, K. Furugaki, T. Matsumoto, S. Mano, S. Ando, Y. Nozaki, W. Yukawa, R. Nakashige, D. Yamaguchi, H. Ishibashi, M. Yonekura, Y. Nakami, S. Takayama, T. Endo, T. Saruwatari, M. Yagura, Y. Yoshikawa, K. Fujimoto, A. Oka, S. Chiku, S. E. Linsen, M. J. Giphart, J. K. Kulski, T. Fukazawa, H. Hashimoto, M. Kimura, Y. Hoshina, Y. Suzuki, T. Hotta, J. Mochida, T. Minezaki, K. Komai, S. Shiozawa, A. Taniguchi, H. Yamanaka, N. Kamatani, T. Gojobori, S. Bahram, and H. Inoko, "Whole genome association study of rheumatoid arthritis using 27 039 microsatellites," *Hum Mol Genet*, vol. 14, no. 16, pp. 2305–21, 2005.

[72] C. M. Ruitberg, D. J. Reeder, and J. M. Butler, "Strbase: a short tandem repeat dna database for the human identity testing community," *Nucleic Acids Res*, vol. 29, no. 1, pp. 320–2, 2001.

[73] E. Jorgenson and J. S. Witte, "Microsatellite markers for genome-wide association studies," *Nat Rev Genet*, vol. 8, no. 2, 2007.

[74] B. A. Payseur, P. Jing, and R. J. Haasl, "A genomic portrait of human microsatellite variation," *Mol Biol Evol*, vol. 28, no. 1, pp. 303–12, 2011.

[75] M. Molla, A. Delcher, S. Sunyaev, C. Cantor, and S. Kasif, "Triplet repeat length bias

and variation in the human transcriptome," *Proc Natl Acad Sci U S A*, vol. 106, no. 40, pp. 17095–100, 2009.

[76] J. Duitama, A. Zablotskaya, R. Gemayel, A. Jansen, S. Belet, J. R. Vermeesch, K. J. Verstrepen, and G. Froyen, "Large-scale analysis of tandem repeat variability in the human genome," *Nucleic Acids Res*, vol. 42, no. 9, pp. 5728–41, 2014.

[77] T. J. Treangen and S. L. Salzberg, "Repetitive dna and next-generation sequencing: computational challenges and solutions," *Nat Rev Genet*, vol. 13, no. 1, pp. 36–46, 2012.

[78] X. Y. Hauge and M. Litt, "A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the pcr," *Hum Mol Genet*, vol. 2, no. 4, pp. 411–5, 1993.

[79] J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, G. O. Broad, G. O. Seattle, and N. E. S. Project, "Evolution and functional impact of rare coding variation from deep sequencing of human exomes," *Science*, vol. 337, no. 6090, pp. 64–9, 2012.

[80] S. B. Montgomery, D. L. Goode, E. Kvikstad, C. A. Albers, Z. D. Zhang, X. J. Mu, G. Ananda, B. Howie, K. J. Karczewski, K. S. Smith, V. Anaya, R. Richardson, J. Davis, C. Genomes Project, D. G. MacArthur, A. Sidow, L. Duret, M. Gerstein, K. D. Makova, J. Marchini, G. McVean, and G. Lunter, "The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes," *Genome Res*, vol. 23, no. 5, pp. 749–61, 2013.

[81] L. J. McIver, J. F. McCormick, A. Martin, r. Fondon, J. W., and H. R. Garner, "Population-scale analysis of human microsatellites reveals novel sources of exonic variation," *Gene*, vol. 516, no. 2, pp. 328–34, 2013.

[82] G. Ananda, E. Walsh, K. D. Jacob, M. Krasilnikova, K. A. Eckert, F. Chiaromonte, and K. D. Makova, "Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome," *Genome Biol Evol*, vol. 5, no. 3, pp. 606–20, 2013.

[83] S. Leclercq, E. Rivals, and P. Jarne, "Detecting microsatellites within genomes: significant variation among algorithms," *BMC Bioinformatics*, vol. 8, p. 125, 2007.

[84] r. Fondon, J. W., A. Martin, S. Richards, R. A. Gibbs, and D. Mittelman, "Analysis of microsatellite variation in drosophila melanogaster with population-scale genome sequencing," *PLoS One*, vol. 7, no. 3, p. e33036, 2012.

[85] E. Schaper, A. V. Kajava, A. Hauser, and M. Anisimova, "Repeat or not repeat?–statistical validation of tandem repeat prediction in genomic sequences," *Nucleic Acids Res*, vol. 40, no. 20, pp. 10005–17, 2012.

[86] E. Buschiazzo and N. J. Gemmell, "The rise, fall and renaissance of microsatellites in eukaryotic genomes," *Bioessays*, vol. 28, no. 10, pp. 1040–50, 2006.

[87] E. J. Oliveira, J. G. PÃądua, M. I. Zucchi, R. Vencovsky, and M. L. C. Vieira, "Origin, evolution and genome distribution of microsatellites," *Genetics and Molecular Biology*, vol. 29, pp. 294–307, 2006.

[88] Y. D. Kelkar, K. A. Eckert, F. Chiaromonte, and K. D. Makova, "A matter of life or death: how microsatellites emerge in and vanish from the human genome," *Genome Res*, vol. 21, no. 12, pp. 2038–48, 2011.

[89] G. Benson, "Tandem repeats finder: a program to analyze dna sequences," *Nucleic Acids Res*, vol. 27, no. 2, pp. 573–80, 1999.

[90] Y. Lai and F. Sun, "The relationship between microsatellite slippage mutation rate and the number of repeat units," *Mol Biol Evol*, vol. 20, no. 12, pp. 2123–31, 2003.

[91] Y. D. Kelkar, N. Strubczewski, S. E. Hile, F. Chiaromonte, K. A. Eckert, and K. D. Makova, "What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at a/t and gt/ac repeats," *Genome Biol Evol*, vol. 2, pp. 620–35, 2010.

[92] N. A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman, "Clines, clusters, and the effect of study design on the inference of human population structure," *PLoS Genet*, vol. 1, no. 6, p. e70, 2005.

[93] F. Gebhardt, K. S. Zanker, and B. Brandt, "Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1," *J Biol Chem*, vol. 274, no. 19, pp. 13176–80, 1999.

[94] 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*, vol. 455, no. 1-2, pp. 70–4, 1999.

[95] F. Pompanon, A. Bonin, E. Bellemain, and P. Taberlet, "Genotyping errors: causes, consequences and solutions," *Nat Rev Genet*, vol. 6, no. 11, pp. 847–59, 2005.

[96] B. Budowle, T. R. Moretti, A. L. Baumstark, D. A. Defenbaugh, and K. M. Keys, "Population data on the thirteen codis core short tandem repeat loci in african americans, u.s. caucasians, hispanics, bahamians, jamaicans, and trinidadians," *J Forensic Sci*, vol. 44, no. 6, pp. 1277–86, 1999.

[97] M. Stoneking and J. Krause, "Learning about human population history from ancient and modern genomes," *Nat Rev Genet*, vol. 12, no. 9, pp. 603–14, 2011.

[98] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, no. 2, pp. 945–59, 2000.

[99] H. Ellegren, "Heterogeneous mutation processes in human microsatellite dna sequences," *Nat Genet*, vol. 24, no. 4, pp. 400–2, 2000.

[100] X. Xu, M. Peng, and Z. Fang, "The direction of microsatellite mutations is dependent upon allele length," *Nat Genet*, vol. 24, no. 4, pp. 396–9, 2000.

[101] J. C. Whittaker, R. M. Harbord, N. Boxall, I. Mackay, G. Dawson, and R. M. Sibly, "Likelihood-based estimation of microsatellite mutation rates," *Genetics*, vol. 164, no. 2, pp. 781–7, 2003.

[102] N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman, "Genetic structure of human populations," *Science*, vol. 298, no. 5602, pp. 2381–5, 2002.

[103] B. Dumont, D. Lasne, C. Rothschild, M. Bouabdelli, V. Ollivier, C. Oudin, N. Ajzenberg, B. Grandchamp, and M. Jandrot-Perrus, "Absence of collagen-induced platelet activation caused by compound heterozygous gpvi mutations," *Blood*, vol. 114, no. 9, pp. 1900–3, 2009.

[104] C. Hermans, C. Wittevrongel, C. Thys, P. A. Smethurst, C. Van Geet, and K. Freson, "A compound heterozygous mutation in glycoprotein vi in a patient with a bleeding disorder," *J Thromb Haemost*, vol. 7, no. 8, pp. 1356–63, 2009.

[105] B. A. Payseur, M. Place, and J. L. Weber, "Linkage disequilibrium between strps and snps across the human genome," *Am J Hum Genet*, vol. 82, no. 5, pp. 1039–50, 2008.

[106] A. D. Ewing and J. Kazazian, H. H., "Whole-genome resequencing allows detection of many rare line-1 insertion alleles in humans," *Genome Res*, vol. 21, no. 6, pp. 985–90, 2011.

[107] F. Hormozdiari, C. Alkan, M. Ventura, I. Hajirasouliha, M. Malig, F. Hach, D. Yorukoglu, P. Dao, M. Bakhshi, S. C. Sahinalp, and E. E. Eichler, "Alu repeat discovery and characterization within human genomes," *Genome Res*, vol. 21, no. 6, pp. 840–9, 2011.

[108] A. Kirby, A. Gnirke, D. B. Jaffe, V. Baresova, N. Pochet, B. Blumenstiel, C. Ye, D. Aird, C. Stevens, J. T. Robinson, M. N. Cabili, I. Gat-Viks, E. Kelliher, R. Daza, M. DeFelice, H. Hulkova, J. Sovova, P. Vylet'al, C. Antignac, M. Guttman, R. E. Handsaker, D. Perrin, S. Steelman, S. Sigurdsson, S. J. Scheinman, C. Sougnez, K. Cibulskis, M. Parkin, T. Green, E. Rossin, M. C. Zody, R. J. Xavier, M. R. Pollak, S. L. Alper, K. Lindblad-Toh, S. Gabriel, P. S. Hart, A. Regev, C. Nusbaum, S. Kmoch, A. J. Bleyer, E. S. Lander, and M. J. Daly, "Mutations causing medullary cystic kidney disease type 1 lie in a large vntr in muc1 missed by massively parallel sequencing," *Nat Genet*, vol. 45, no. 3, pp. 299–303, 2013.

[109] J. Hui, K. Stangl, W. S. Lane, and A. Bindereif, "Hnrnp l stimulates splicing of the enos gene by binding to variable-length ca repeats," *Nat Struct Biol*, vol. 10, no. 1, pp. 33–7, 2003.

[110] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–53, 2009.

[111] M. O. Press, K. D. Carlson, and C. Queitsch, "The overdue promise of short tandem repeat variation for heritability," *Trends Genet*, vol. 30, no. 11, pp. 504–12, 2014.

[112] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and S. Genome Project Data Processing, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–9, 2009.

[113] D. Falush, M. Stephens, and J. K. Pritchard, "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies," *Genetics*, vol. 164, no. 4, pp. 1567–87, 2003.

[114] A. R. Quinlan and I. M. Hall, "Bedtools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–2, 2010.

[115] J. C. Barrett, S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, J. D. Rioux, S. R. Brant, M. S. Silverberg, K. D. Taylor, M. M. Barmada, A. Bitton, T. Dassopoulos, L. W. Datta, T. Green, A. M. Griffiths, E. O. Kistner, M. T. Murtha, M. D. Regueiro, J. I. Rotter, L. P. Schumm, A. H. Steinhart, S. R. Targan, R. J. Xavier, N. I. G. Consortium, C. Libioulle, C. Sandor, M. Lathrop, J. Belaiche, O. Dewit, I. Gut, S. Heath, D. Laukens, M. Mni, P. Rutgeerts, A. Van Gossum, D. Zelenika, D. Franchimont, J. P. Hugot, M. de Vos, S. Vermeire, E. Louis, I. B. D. C. Belgian-French, C. Wellcome Trust Case Control, L. R. Cardon, C. A. Anderson, H. Drummond, E. Nimmo, T. Ahmad, N. J. Prescott, C. M. Onnie, S. A. Fisher, J. Marchini, J. Ghori, S. Bumpstead, R. Gwilliam, M. Tremelling, P. Deloukas, J. Mansfield, D. Jewell, J. Satsangi, C. G. Mathew, M. Parkes, M. Georges, and M. J. Daly, "Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease," *Nat Genet*, vol. 40, no. 8, pp. 955–62, 2008.

[116] M. F. Moffatt, M. Kabesch, L. Liang, A. L. Dixon, D. Strachan, S. Heath, M. Depner, A. von Berg, A. Bufe, E. Rietschel, A. Heinzmann, B. Simma, T. Frischer, S. A. Willis-Owen, K. C. Wong, T. Illig, C. Vogelberg, S. K. Weiland, E. von Mutius, G. R. Abecasis, M. Farrall, I. G. Gut, G. M. Lathrop, and W. O. Cookson, "Genetic variants regulating ormdl3 expression contribute to the risk of childhood asthma," *Nature*, vol. 448, no. 7152, pp. 470–3, 2007.

[117] G. T. Consortium, "Human genomics. the genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans," *Science*, vol. 348, no. 6235, pp. 648–60, 2015.

[118] A. C. Nica, S. B. Montgomery, A. S. Dimas, B. E. Stranger, C. Beazley, I. Barroso, and

E. T. Dermitzakis, "Candidate causal regulatory effects by integration of expression qtls with complex trait genetic associations," *PLoS Genet*, vol. 6, no. 4, p. e1000895, 2010.

[119] D. L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M. E. Dolan, and N. J. Cox, "Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas," *PLoS Genet*, vol. 6, no. 4, p. e1000888, 2010.

[120] L. D. Ward and M. Kellis, "Interpreting noncoding genetic variation in complex traits and human disease," *Nat Biotechnol*, vol. 30, no. 11, pp. 1095–106, 2012.

[121] E. P. Consortium, "An integrated encyclopedia of dna elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[122] E. Grundberg, K. S. Small, A. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T. P. Yang, E. Meduri, A. Barrett, J. Nisbett, M. Sekowska, A. Wilk, S. Y. Shin, D. Glass, M. Travers, J. L. Min, S. Ring, K. Ho, G. Thorleifsson, A. Kong, U. Thorsteindottir, C. Ainali, A. S. Dimas, N. Hassanali, C. Ingle, D. Knowles, M. Krestyaninova, C. E. Lowe, P. Di Meglio, S. B. Montgomery, L. Parts, S. Potter, G. Surdulescu, L. Tsaprouni, S. Tsoka, V. Bataille, R. Durbin, F. O. Nestle, S. O'Rahilly, N. Soranzo, C. M. Lindgren, K. T. Zondervan, K. R. Ahmadi, E. E. Schadt, K. Stefansson, G. D. Smith, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, T. D. Spector, and C. Multiple Tissue Human Expression Resource, "Mapping cis- and trans-regulatory effects across multiple tissues in twins," *Nat Genet*, vol. 44, no. 10, pp. 1084–9, 2012.

[123] F. A. Wright, P. F. Sullivan, A. I. Brooks, F. Zou, W. Sun, K. Xia, V. Madar, R. Jansen, W. Chung, Y. H. Zhou, A. Abdellaoui, S. Batista, C. Butler, G. Chen, T. H. Chen, D. D'Ambrosio, P. Gallins, M. J. Ha, J. J. Hottenga, S. Huang, M. Kattenberg, J. Kochar, C. M. Middeldorp, A. Qu, A. Shabalin, J. Tischfield, L. Todd, J. Y. Tzeng, G. van Grootheest, J. M. Vink, Q. Wang, W. Wang, W. Wang, G. Willemsen, J. H. Smit, E. J. de Geus, Z. Yin, B. W. Penninx, and D. I. Boomsma, "Heritability and genomics of gene expression in peripheral blood," *Nat Genet*, vol. 46, no. 5, pp. 430–7, 2014.

[124] P. Martin, K. Makepeace, S. A. Hill, D. W. Hood, and E. R. Moxon, "Microsatellite instability regulates transcription factor binding and gene expression," *Proc Natl Acad Sci U S A*, vol. 102, no. 10, pp. 3800–4, 2005.

[125] R. Willems, A. Paul, H. G. van der Heide, A. R. ter Avest, and F. R. Mooi, "Fimbrial phase variation in bordetella pertussis: a novel mechanism for transcriptional regulation," *EMBO J*, vol. 9, no. 9, pp. 2803–9, 1990.

[126] D. Yogev, R. Rosengarten, R. Watson-McKown, and K. S. Wise, "Molecular basis of mycoplasma surface antigenic variation: a novel set of divergent genes undergo spontaneous mutation of periodic coding regions and 5' regulatory sequences," *EMBO J*, vol. 10, no. 13, pp. 4069–79, 1991.

[127] 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*, vol. 24, no. 11, pp. 1988–98, 2005.

[128] S. Rothenburg, F. Koch-Nolte, A. Rich, and F. Haag, "A polymorphic dinucleotide repeat in the rat nucleolin gene forms z-dna and inhibits promoter activity," *Proc Natl Acad Sci U S A*, vol. 98, no. 16, pp. 8985–90, 2001.

[129] J. N. Weiser, J. M. Love, and E. R. Moxon, "The molecular mechanism of phase variation of h. influenzae lipopolysaccharide," *Cell*, vol. 59, no. 4, pp. 657–65, 1989.

[130] E. A. Hammock and L. J. Young, "Microsatellite instability generates diversity in brain and sociobehavioral traits," *Science*, vol. 308, no. 5728, pp. 1630–4, 2005.

[131] J. O. Yanez-Cuna, C. D. Arnold, G. Stampfel, L. M. Boryn, D. Gerlach, M. Rath, and A. Stark, "Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features," *Genome Res*, vol. 24, no. 7, pp. 1147–56, 2014.

[132] S. Sawaya, A. Bagshaw, E. Buschiazzo, P. Kumar, S. Chowdhury, M. A. Black, and N. Gemmell, "Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements," *PLoS One*, vol. 8, no. 2, p. e54710, 2013.

[133] T. Bilgin Sonay, T. Carvalho, M. D. Robinson, M. P. Greminger, M. Krutzen, D. Comas, G. Highnam, D. Mittelman, A. Sharp, T. Marques-Bonet, and A. Wagner, "Tandem repeat variation in human and great ape populations and its impact on gene expression divergence," *Genome Res*, vol. 25, no. 11, pp. 1591–9, 2015.

[134] 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*, vol. 33, no. 8, pp. 1302–9, 2012.

[135] M. V. Rockman and G. A. Wray, "Abundant raw material for cis-regulatory evolution in humans," *Mol Biol Evol*, vol. 19, no. 11, pp. 1991–2004, 2002.

[136] K. M. Warpeha, W. Xu, L. Liu, I. G. Charles, C. C. Patterson, F. Ah-Fat, S. Harding, P. M. Hart, U. Chakravarthy, and A. E. Hughes, "Genotyping and functional analysis of a polymorphic (ccttt)(n) repeat of nos2a in diabetic retinopathy," *FASEB J*, vol. 13, no. 13, pp. 1825–32, 1999.

[137] C. Genomes Project, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–73, 2010.

[138] T. Willems, M. Gymrek, G. Highnam, C. Genomes Project, D. Mittelman, and Y. Erlich, "The landscape of human str variation," *Genome Res*, vol. 24, no. 11, pp. 1894–904, 2014.

[139] A. R. La Spada, D. B. Roling, A. E. Harding, C. L. Warner, R. Spiegel, I. Hausmanowa-

Petrusewicz, W. C. Yee, and K. H. Fischbeck, "Meiotic stability and genotype-phenotype correlation of the trinucleotide repeat in x-linked spinal and bulbar muscular atrophy," *Nat Genet*, vol. 2, no. 4, pp. 301–4, 1992.

[140] M. Duyao, C. Ambrose, R. Myers, A. Novelletto, F. Persichetti, M. Frontali, S. Folstein, C. Ross, M. Franz, M. Abbott, and et al., "Trinucleotide repeat length instability and age of onset in huntington's disease," *Nat Genet*, vol. 4, no. 4, pp. 387–92, 1993.

[141] P. Flicek, I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. GarcÃŋa-GirÃșn, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A. K. KÃďhÃďri, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T. J. P. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, and S. M. J. Searle, "Ensembl 2013," *Nucleic Acids Research*, vol. 41, no. Database issue, pp. D48–D55, 2013.

[142] B. E. Stranger, S. B. Montgomery, A. S. Dimas, L. Parts, O. Stegle, C. E. Ingle, M. Sekowska, G. D. Smith, D. Evans, M. Gutierrez-Arcelus, A. Price, T. Raj, J. Nisbett, A. C. Nica, C. Beazley, R. Durbin, P. Deloukas, and E. T. Dermitzakis, "Patterns of cis regulatory variation in diverse human populations," *PLoS Genet*, vol. 8, no. 4, p. e1002639, 2012.

[143] S. Sawaya, M. Jones, and M. Keller, "Linkage disequilibrium between single nucleotide polymorphisms and hypermutable loci," *bioRxiv*, 2015.

[144] C. Lamina, M. Haun, S. Coassin, A. Kloss-Brandstatter, C. Gieger, A. Peters, H. Grallert, K. Strauch, T. Meitinger, L. Kedenko, B. Paulweber, and F. Kronenberg, "A systematic evaluation of short tandem repeats in lipid candidate genes: riding on the snp-wave," *PLoS One*, vol. 9, no. 7, p. e102113, 2014.

[145] A. Gusev, S. H. Lee, B. M. Neale, G. Trynka, B. J. Vilhjalmsson, H. Finucane, H. Xu, C. Zang, S. Ripke, E. Stahl, n. a. Schizophrenia Working Group of the Pgc, n. a. Swe-Scz Consortium, A. K. Kahler, C. M. Hultman, S. M. Purcell, S. A. McCarroll, M. Daly, B. Pasaniuc, P. F. Sullivan, N. R. Wray, S. Raychaudhuri, and A. L. Price, "Regulatory variants explain much more heritability than coding variants across 11 common diseases," *bioRxiv*, 2014.

[146] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher, "Common snps explain a large proportion of the heritability for human height," *Nat Genet*, vol. 42, no. 7, pp. 565–9, 2010.

[147] J. P. Ioannidis, "Why most discovered true associations are inflated," *Epidemiology*, vol. 19, no. 5, pp. 640–8, 2008.

[148] D. J. Gaffney, J. B. Veyrieras, J. F. Degner, R. Pique-Regi, A. A. Pai, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard, "Dissecting the regulatory architecture of gene expression qtls," *Genome Biol*, vol. 13, no. 1, p. R7, 2012.

[149] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, "Detection of nonneutral substitution rates on mammalian phylogenies," *Genome Res*, vol. 20, no. 1, pp. 110–21, 2010.

[150] G. Trynka, H. J. Westra, K. Slowikowski, X. Hu, H. Xu, B. E. Stranger, R. J. Klein, B. Han, and S. Raychaudhuri, "Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci," *Am J Hum Genet*, vol. 97, no. 1, pp. 139–52, 2015.

[151] J. Ernst and M. Kellis, "Chromhmm: automating chromatin-state discovery and characterization," *Nat Methods*, vol. 9, no. 3, pp. 215–6, 2012.

[152] H. Zeng, T. Hashimoto, D. D. Kang, and D. K. Gifford, "Gerv: a statistical method for generative evaluation of regulatory variants for transcription factor binding," *Bioinformatics*, vol. 32, no. 4, pp. 490–6, 2016.

[153] A. Doring, C. Gieger, D. Mehta, H. Gohlke, H. Prokisch, S. Coassin, G. Fischer, K. Henke, N. Klopp, F. Kronenberg, B. Paulweber, A. Pfeufer, D. Rosskopf, H. Volzke, T. Illig, T. Meitinger, H. E. Wichmann, and C. Meisinger, "Slc2a9 influences uric acid concentrations with pronounced sex-specific effects," *Nat Genet*, vol. 40, no. 4, pp. 430–6, 2008.

[154] V. Vitart, I. Rudan, C. Hayward, N. K. Gray, J. Floyd, C. N. Palmer, S. A. Knott, I. Kolcic, O. Polasek, J. Graessler, J. F. Wilson, A. Marinaki, P. L. Riches, X. Shu, B. Janicijevic, N. Smolej-Narancic, B. Gorgoni, J. Morgan, S. Campbell, Z. Biloglav, L. Barac-Lauc, M. Pericic, I. M. Klaric, L. Zgaga, T. Skaric-Juric, S. H. Wild, W. A. Richardson, P. Hohenstein, C. H. Kimber, A. Tenesa, L. A. Donnelly, L. D. Fairbanks, M. Aringer, P. M. McKeigue, S. H. Ralston, A. D. Morris, P. Rudan, N. D. Hastie, H. Campbell, and A. F. Wright, "Slc2a9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout," *Nat Genet*, vol. 40, no. 4, pp. 437–42, 2008.

[155] C. Wallace, S. J. Newhouse, P. Braund, F. Zhang, M. Tobin, M. Falchi, K. Ahmadi, R. J. Dobson, A. C. Marcano, C. Hajat, P. Burton, P. Deloukas, M. Brown, J. M. Connell, A. Dominiczak, G. M. Lathrop, J. Webster, M. Farrall, T. Spector, N. J. Samani, M. J. Caulfield, and P. B. Munroe, "Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia," *Am J Hum Genet*, vol. 82, no. 1, pp. 139–49, 2008.

[156] S. Y. Shin, E. B. Fauman, A. K. Petersen, J. Krumsiek, R. Santos, J. Huang, M. Arnold,

I. Erte, V. Forgetta, T. P. Yang, K. Walter, C. Menni, L. Chen, L. Vasquez, A. M. Valdes, C. L. Hyde, V. Wang, D. Ziemek, P. Roberts, L. Xi, E. Grundberg, C. Multiple Tissue Human Expression Resource, M. Waldenberger, J. B. Richards, R. P. Mohney, M. V. Milburn, S. L. John, J. Trimmer, F. J. Theis, J. P. Overington, K. Suhre, M. J. Brosnan, C. Gieger, G. Kastenmuller, T. D. Spector, and N. Soranzo, "An atlas of genetic influences on human blood metabolites," *Nat Genet*, vol. 46, no. 6, pp. 543–50, 2014.

[157] M. J. Chaisson, J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig, F. Hormozdiari, F. Antonacci, U. Surti, R. Sandstrom, M. Boitano, J. M. Landolin, J. A. Stamatoyannopoulos, M. W. Hunkapiller, J. Korlach, and E. E. Eichler, "Resolving the complexity of the human genome using single-molecule sequencing," *Nature*, vol. 517, no. 7536, pp. 608–11, 2015.

[158] G. Bhatia, A. Gusev, P.-R. Loh, B. J. VilhjÃąlmsson, S. Ripke, S. Purcell, E. Stahl, M. Daly, T. R. de Candia, K. S. Kendler, M. C. O'Donovan, S. H. Lee, N. R. Wray, B. M. Neale, M. C. Keller, N. A. Zaitlen, B. Pasaniuc, J. Yang, and A. L. Price, "Haplotypes of common snps can explain missing heritability of complex diseases," *bioRxiv*, 2015.

[159] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, "Plink: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, vol. 81, no. 3, pp. 559–75, 2007.

[160] A. Guilmatre, G. Highnam, C. Borel, D. Mittelman, and A. J. Sharp, "Rapid multiplexed genotyping of simple tandem repeats using capture and high-throughput sequencing," *Hum Mutat*, vol. 34, no. 9, pp. 1304–11, 2013.

[161] D. Karolchik, G. P. Barber, J. Casper, H. Clawson, M. S. Cline, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, A. S. Hinrichs, K. Learned, B. T. Lee, C. H. Li, B. J. Raney, B. Rhead, K. R. Rosenbloom, C. A. Sloan, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn, and W. J. Kent, "The ucsc genome browser database: 2014 update," *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D764–70, 2014.

[162] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, "The human genome browser at ucsc," *Genome Res*, vol. 12, no. 6, pp. 996–1006, 2002.

[163] C. Trapnell, L. Pachter, and S. L. Salzberg, "Tophat: discovering splice junctions with rna-seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–11, 2009.

[164] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, "Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks," *Nat Protoc*, vol. 7, no. 3, pp. 562–78, 2012.

[165] N. L. Barbosa-Morais, M. J. Dunning, S. A. Samarajiwa, J. F. Darot, M. E. Ritchie, A. G. Lynch, and S. Tavare, "A re-annotation pipeline for illumina beadarrays: improving the interpretation of gene expression data," *Nucleic Acids Res*, vol. 38, no. 3, p. e17, 2010.

[166] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "Gcta: a tool for genome-wide complex trait analysis," *Am J Hum Genet*, vol. 88, no. 1, pp. 76–82, 2011.

[167] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," *PLoS Genet*, vol. 2, no. 12, p. e190, 2006.

[168] K. Wang, H. Zhang, D. Ma, M. Bucan, J. T. Glessner, B. S. Abrahams, D. Salyakina, M. Imielinski, J. P. Bradfield, P. M. Sleiman, C. E. Kim, C. Hou, E. Frackelton, R. Chiavacci, N. Takahashi, T. Sakurai, E. Rappaport, C. M. Lajonchere, J. Munson, A. Estes, O. Korvatska, J. Piven, L. I. Sonnenblick, A. I. Alvarez Retuerto, E. I. Herman, H. Dong, T. Hutman, M. Sigman, S. Ozonoff, A. Klin, T. Owley, J. A. Sweeney, C. W. Brune, R. M. Cantor, R. Bernier, J. R. Gilbert, M. L. Cuccaro, W. M. McMahon, J. Miller, M. W. State, T. H. Wassink, H. Coon, S. E. Levy, R. T. Schultz, J. I. Nurnberger, J. L. Haines, J. S. Sutcliffe, E. H. Cook, N. J. Minshew, J. D. Buxbaum, G. Dawson, S. F. Grant, D. H. Geschwind, M. A. Pericak-Vance, G. D. Schellenberg, and H. Hakonarson, "Common genetic variants on 5p14.1 associate with autism spectrum disorders," *Nature*, vol. 459, no. 7246, pp. 528–33, 2009.

[169] I. Iossifov, B. J. O'Roak, S. J. Sanders, M. Ronemus, N. Krumm, D. Levy, H. A. Stessman, K. T. Witherspoon, L. Vives, K. E. Patterson, J. D. Smith, B. Paeper, D. A. Nickerson, J. Dea, S. Dong, L. E. Gonzalez, J. D. Mandell, S. M. Mane, M. T. Murtha, C. A. Sullivan, M. F. Walker, Z. Waqar, L. Wei, A. J. Willsey, B. Yamrom, Y. H. Lee, E. Grabowska, E. Dalkic, Z. Wang, S. Marks, P. Andrews, A. Leotta, J. Kendall, I. Hakker, J. Rosenbaum, B. Ma, L. Rodgers, J. Troge, G. Narzisi, S. Yoon, M. C. Schatz, K. Ye, W. R. McCombie, J. Shendure, E. E. Eichler, M. W. State, and M. Wigler, "The contribution of de novo coding mutations to autism spectrum disorder," *Nature*, vol. 515, no. 7526, pp. 216–21, 2014.

[170] N. Krumm, T. N. Turner, C. Baker, L. Vives, K. Mohajeri, K. Witherspoon, A. Raja, B. P. Coe, H. A. Stessman, Z. X. He, S. M. Leal, R. Bernier, and E. E. Eichler, "Excess of rare, inherited truncating mutations in autism," *Nat Genet*, vol. 47, no. 6, pp. 582–8, 2015.

[171] S. Ripke, C. O'Dushlaine, K. Chambert, J. L. Moran, A. K. Kahler, S. Akterin, S. E. Bergen, A. L. Collins, J. J. Crowley, M. Fromer, Y. Kim, S. H. Lee, P. K. Magnusson, N. Sanchez, E. A. Stahl, S. Williams, N. R. Wray, K. Xia, F. Bettella, A. D. Borglum, B. K. Bulik-Sullivan, P. Cormican, N. Craddock, C. de Leeuw, N. Durmishi, M. Gill, V. Golimbet, M. L. Hamshere, P. Holmans, D. M. Hougaard, K. S. Kendler, K. Lin, D. W. Morris, O. Mors, P. B. Mortensen, B. M. Neale, F. A. O'Neill, M. J. Owen, M. P. Milovancevic, D. Posthuma, J. Powell, A. L. Richards, B. P. Riley, D. Ruderfer, D. Rujescu, E. Sigurdsson, T. Silagadze, A. B. Smit, H. Stefansson, S. Steinberg, J. Suvisaari,

S. Tosato, M. Verhage, J. T. Walters, C. Multicenter Genetic Studies of Schizophrenia, D. F. Levinson, P. V. Gejman, K. S. Kendler, C. Laurent, B. J. Mowry, M. C. O'Donovan, M. J. Owen, A. E. Pulver, B. P. Riley, S. G. Schwab, D. B. Wildenauer, F. Dudbridge, P. Holmans, J. Shi, M. Albus, M. Alexander, D. Campion, D. Cohen, D. Dikeos, J. Duan, P. Eichhammer, S. Godard, M. Hansen, F. B. Lerer, K. Y. Liang, W. Maier, J. Mallet, D. A. Nertney, G. Nestadt, N. Norton, F. A. O'Neill, G. N. Papadimitriou, R. Ribble, A. R. Sanders, J. M. Silverman, D. Walsh, N. M. Williams, B. Wormley, C. Psychosis Endophenotypes International, M. J. Arranz, S. Bakker, S. Bender, E. Bramon, D. Collier, B. Crespo-Facorro, *et al.*, "Genome-wide association analysis identifies 13 new risk loci for schizophrenia," *Nat Genet*, vol. 45, no. 10, pp. 1150–9, 2013.

[172] V. Marx, "The dna of a nation," *Nature*, vol. 524, no. 7566, pp. 503–505, 2015.

[173] 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*, 2015.

[174] H. Li, "Toward better understanding of artifacts in variant calling from high-coverage samples," *Bioinformatics*, vol. 30, no. 20, pp. 2843–51, 2014.

[175] O. Delaneau and J. F. Zagury, "Haplotype inference," *Methods Mol Biol*, vol. 888, pp. 177–96, 2012.

[176] 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*, vol. 81, no. 5, pp. 1084–97, 2007.

[177] B. N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genet*, vol. 5, no. 6, p. e1000529, 2009.

[178] B. L. Browning and S. R. Browning, "Genotype imputation with millions of reference samples," *Am J Hum Genet*, vol. 98, no. 1, pp. 116–26, 2016.

[179] E. Garrison and G. Marth, "Haplotype-based variant detection from short-read sequencing," *arXiv preprint arXiv:1207.3907*, 2012.

[180] D. Shinde, Y. Lai, F. Sun, and N. Arnheim, "Taq dna polymerase slippage mutation rates measured by pcr and quasi-likelihood analysis: (ca/gt)n and (a/t)n microsatellites," *Nucleic Acids Res*, vol. 31, no. 3, pp. 974–80, 2003.

[181] P. H. Sudmant, S. Mallick, B. J. Nelson, F. Hormozdiari, N. Krumm, J. Huddleston, B. P. Coe, C. Baker, S. Nordenfelt, M. Bamshad, L. B. Jorde, O. L. Posukh, H. Sahakyan, W. S. Watkins, L. Yepiskoposyan, M. S. Abdullah, C. M. Bravi, C. Capelli, T. Hervig, J. T. Wee, C. Tyler-Smith, G. van Driem, I. G. Romero, A. R. Jha, S. Karachanak-Yankova, D. Toncheva, D. Comas, B. Henn, T. Kivisild, A. Ruiz-Linares, A. Sajantila, E. Metspalu, J. Parik, R. Villems, E. B. Starikovskaya, G. Ayodo, C. M. Beall, A. Di Rienzo,

M. F. Hammer, R. Khusainova, E. Khusnutdinova, W. Klitz, C. Winkler, D. Labuda, M. Metspalu, S. A. Tishkoff, S. Dryomov, R. Sukernik, N. Patterson, D. Reich, and E. E. Eichler, "Global diversity, population stratification, and selection of human copy-number variation," *Science*, vol. 349, no. 6253, p. aab3761, 2015.

[182] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with bwa-mem," *arXiv preprint arXiv:1303.3997*, 2013.

[183] J. M. Zook, D. Catoe, J. McDaniel, L. Vang, N. Spies, A. Sidow, Z. Weng, Y. Liu, C. Mason, N. Alexander, D. Chandramohan, E. Henaff, F. Chen, E. Jaeger, A. Moshrefi, K. Pham, W. Stedman, T. Liang, M. Saghbini, Z. Dzakula, A. Hastie, H. Cao, G. Deikus, E. Schadt, R. Sebra, A. Bashir, R. M. Truty, C. C. Chang, N. Gulbahce, K. Zhao, S. Ghosh, F. Hyland, Y. Fu, M. Chaisson, J. Trow, C. Xiao, S. T. Sherry, A. W. Zaranek, M. Ball, J. Bobe, P. Estep, G. M. Church, P. Marks, S. Kyriazopoulou-Panagiotopoulou, G. Zheng, M. Schnall-Levin, H. S. Ordonez, P. A. Mudivarti, K. Giorda, Y. Sheng, K. B. Rypdal, and M. Salit, "Extensive sequencing of seven human genomes to characterize benchmark reference materials," *bioRxiv*, 2015.

[184] D. F. Conrad, J. E. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang, F. Casals, Y. Idaghdour, C. L. Hartl, C. Torroja, K. V. Garimella, M. Zilversmit, R. Cartwright, G. A. Rouleau, M. Daly, E. A. Stone, M. E. Hurles, P. Awadalla, and P. Genomes, "Variation in genome-wide mutation rates within and between human families," *Nat Genet*, vol. 43, no. 7, pp. 712–4, 2011.

[185] J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies," *Nat Rev Genet*, vol. 11, no. 7, pp. 499–511, 2010.

[186] H. Li, "Fermikit: assembly-based variant calling for illumina resequencing data," *Bioinformatics*, vol. 31, no. 22, pp. 3694–6, 2015.

[187] M. Lek, K. Karczewski, E. Minikel, K. Samocha, E. Banks, T. Fennell, A. O'Donnell-Luria, J. Ware, A. Hill, B. Cummings, T. Tukiainen, D. Birnbaum, J. Kosmicki, L. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, D. Cooper, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. Kurki, A. Levy Moonshine, P. Natarajan, L. Orozco, G. Peloso, R. Poplin, M. Rivas, V. Ruano-Rubio, D. Ruderfer, K. Shakir, P. Stenson, C. Stevens, B. Thomas, G. Tiao, M. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, E. Roberto, J. Florez, S. Gabriel, G. Getz, C. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. McCarthy, D. McGovern, R. McPherson, B. Neale, A. Palotie, S. Purcell, D. Saleheen, J. Scharf, P. Sklar, S. Patrick, J. Tuomilehto, H. Watkins, J. Wilson, M. Daly, and D. MacArthur, "Analysis of protein-coding genetic variation in 60,706 humans," *bioRxiv*, 2015.

[188] J. Huang, B. Howie, S. McCarthy, Y. Memari, K. Walter, J. L. Min, P. Danecek, G. Malerba, E. Trabetti, H. F. Zheng, U. K. Consortium, G. Gambaro, J. B. Richards,

R. Durbin, N. J. Timpson, J. Marchini, and N. Soranzo, "Improved imputation of low-frequency and rare variants using the uk10k haplotype reference panel," *Nat Commun*, vol. 6, p. 8111, 2015.

[189] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[190] C. A. Albers, G. Lunter, D. G. MacArthur, G. McVean, W. H. Ouwehand, and R. Durbin, "Dindel: accurate indel calls from short-read data," *Genome Res*, vol. 21, no. 6, pp. 961–73, 2011.

[191] A. Scally and R. Durbin, "Revising the human mutation rate: implications for understanding human evolution," *Nat Rev Genet*, vol. 13, no. 10, pp. 745–53, 2012.

[192] K. E. Samocha, E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo, L. M. McGrath, J. A. Kosmicki, K. Rehnstrom, S. Mallick, A. Kirby, D. P. Wall, D. G. MacArthur, S. B. Gabriel, M. DePristo, S. M. Purcell, A. Palotie, E. Boerwinkle, J. D. Buxbaum, J. Cook, E. H., R. A. Gibbs, G. D. Schellenberg, J. S. Sutcliffe, B. Devlin, K. Roeder, B. M. Neale, and M. J. Daly, "A framework for the interpretation of de novo mutation in human disease," *Nat Genet*, vol. 46, no. 9, pp. 944–50, 2014.

[193] J. C. Roach, G. Glusman, A. F. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, and D. J. Galas, "Analysis of genetic inheritance in a family quartet by whole-genome sequencing," *Science*, vol. 328, no. 5978, pp. 636–9, 2010.

[194] L. C. Francioli, P. P. Polak, A. Koren, A. Menelaou, S. Chun, I. Renkens, C. Genome of the Netherlands, C. M. van Duijn, M. Swertz, C. Wijmenga, G. van Ommen, P. E. Slagboom, D. I. Boomsma, K. Ye, V. Guryev, P. F. Arndt, W. P. Kloosterman, P. I. de Bakker, and S. R. Sunyaev, "Genome-wide patterns and properties of de novo mutations in humans," *Nat Genet*, vol. 47, no. 7, pp. 822–6, 2015.

[195] L. A. Zhivotovsky, P. A. Underhill, C. Cinnioglu, M. Kayser, B. Morar, T. Kivisild, R. Scozzari, F. Cruciani, G. Destro-Bisol, G. Spedini, G. K. Chambers, R. J. Herrera, K. K. Yong, D. Gresham, I. Tournev, M. W. Feldman, and L. Kalaydjieva, "The effective mutation rate at y chromosome short tandem repeats, with application to human population-divergence time," *Am J Hum Genet*, vol. 74, no. 1, pp. 50–61, 2004.

[196] E. Heyer, J. Puymirat, P. Dieltjes, E. Bakker, and P. de Knijff, "Estimating y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees," *Hum Mol Genet*, vol. 6, no. 5, pp. 799–803, 1997.

[197] D. H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. Larue, J. L. King, and B. Budowle, "Strait razor: a length-based forensic str allele-calling tool for use with second generation sequencing data," *Forensic Sci Int Genet*, vol. 7, no. 4, pp. 409–17, 2013.

[198] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and G. Genomes Project Analysis, "The variant call format and vcftools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–8, 2011.

[199] A. Stamatakis, "Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, no. 9, pp. 1312–3, 2014.

[200] D. H. Huson and C. Scornavacca, "Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks," *Syst Biol*, vol. 61, no. 6, pp. 1061–7, 2012.

[201] G. D. Poznik, Y. Xue, F. L. Mendez, T. F. Willems, A. Massaia, M. A. Wilson Sayres, Q. Ayub, S. A. McCarthy, A. Narechania, S. Kashin, Y. Chen, and R. Banerjee, "Punctuated bursts in human male demography inferred from 1,244 worldwide y-chromosome sequences," *Nature Genetics*, vol. (in press), 2016.

[202] A. Helgason, A. W. Einarsson, V. B. Guethmundsdottir, A. Sigurethsson, E. D. Gunnarsdottir, A. Jagadeesan, S. S. Ebenesersdottir, A. Kong, and K. Stefansson, "The y-chromosome point mutation rate in humans," *Nat Genet*, vol. 47, no. 5, pp. 453–7, 2015.

[203] Y. Xue, Q. Wang, Q. Long, B. L. Ng, H. Swerdlow, J. Burton, C. Skuce, R. Taylor, Z. Abdellah, Y. Zhao, Asan, D. G. MacArthur, M. A. Quail, N. P. Carter, H. Yang, and C. Tyler-Smith, "Human y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree," *Curr Biol*, vol. 19, no. 17, pp. 1453–7, 2009.

[204] S. Willuweit, L. Roewer, and Y. C. U. G. International Forensic, "Y chromosome haplotype reference database (yhrd): update," *Forensic Sci Int Genet*, vol. 1, no. 2, pp. 83–7, 2007.

[205] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent, "The ucsc genome browser database: update 2006," *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D590–8, 2006.

[206] E. K. Hanson and J. Ballantyne, "Comprehensive annotated str physical map of the human y chromosome: Forensic implications," *Leg Med (Tokyo)*, vol. 8, no. 2, pp. 110–20, 2006.

[207] M. Vermeulen, A. Wollstein, K. van der Gaag, O. Lao, Y. Xue, Q. Wang, L. Roewer, H. Knoblauch, C. Tyler-Smith, P. de Knijff, and M. Kayser, "Improving global and regional resolution of male lineage differentiation by simple single-copy y-chromosomal short tandem repeat polymorphisms," *Forensic Sci Int Genet*, vol. 3, no. 4, pp. 205–13, 2009.

[208] J. Purps, S. Siegert, S. Willuweit, M. Nagy, C. Alves, R. Salazar, S. M. Angustia, L. H. Santos, K. Anslinger, B. Bayer, Q. Ayub, W. Wei, Y. Xue, C. Tyler-Smith, M. B. Bafal-

luy, B. Martinez-Jarreta, B. Egyed, B. Balitzki, S. Tschumi, D. Ballard, D. S. Court, X. Barrantes, G. Bassler, T. Wiest, B. Berger, H. Niederstatter, W. Parson, C. Davis, B. Budowle, H. Burri, U. Borer, C. Koller, E. F. Carvalho, P. M. Domingues, W. T. Chamoun, M. D. Coble, C. R. Hill, D. Corach, M. Caputo, M. E. D'Amato, S. Davison, R. Decorte, M. H. Larmuseau, C. Ottoni, O. Rickards, D. Lu, C. Jiang, T. Dobosz, A. Jonkisz, W. E. Frank, I. Furac, C. Gehrig, V. Castella, B. Grskovic, C. Haas, J. Wobst, G. Hadzic, K. Drobnic, K. Honda, Y. Hou, D. Zhou, Y. Li, S. Hu, S. Chen, U. D. Immel, R. Lessig, Z. Jakovski, T. Ilievska, A. E. Klann, C. C. Garcia, P. de Knijff, T. Kraaijenbrink, A. Kondili, P. Miniati, M. Vouropoulou, L. Kovacevic, D. Marjanovic, I. Lindner, I. Mansour, M. Al-Azem, A. E. Andari, M. Marino, S. Furfuro, L. Locarno, P. Martin, G. M. Luque, A. Alonso, L. S. Miranda, H. Moreira, N. Mizuno, Y. Iwashima, R. S. Neto, T. L. Nogueira, R. Silva, M. Nastainczyk-Wulf, J. Edelmann, M. Kohl, S. Nie, X. Wang, B. Cheng, *et al.*, "A global analysis of y-chromosomal haplotype diversity for 23 str loci," *Forensic Sci Int Genet*, vol. 12, pp. 12–23, 2014.

[209] L. A. Zhivotovsky, P. A. Underhill, and M. W. Feldman, "Difference between evolutionarily effective and germ line mutation rate due to stochastically varying haplogroup size," *Mol Biol Evol*, vol. 23, no. 12, pp. 2268–70, 2006.

[210] J. Felsenstein, "Evolutionary trees from dna sequences: a maximum likelihood approach," *J Mol Evol*, vol. 17, no. 6, pp. 368–76, 1981.

[211] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[212] B. Brinkmann, M. Klintschar, F. Neuhuber, J. Huhne, and B. Rolf, "Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat," *Am J Hum Genet*, vol. 62, no. 6, pp. 1408–15, 1998.

[213] K. N. Ballantyne, A. Ralf, R. Aboukhalid, N. M. Achakzai, M. J. Anjos, Q. Ayub, J. Balazic, J. Ballantyne, D. J. Ballard, B. Berger, C. Bobillo, M. Bouabdellah, H. Burri, T. Capal, S. Caratti, J. Cardenas, F. Cartault, E. F. Carvalho, M. Carvalho, B. Cheng, M. D. Coble, D. Comas, D. Corach, M. E. D'Amato, S. Davison, P. de Knijff, M. C. De Ungria, R. Decorte, T. Dobosz, B. M. Dupuy, S. Elmrghni, M. Gliwinski, S. C. Gomes, L. Grol, C. Haas, E. Hanson, J. Henke, L. Henke, F. Herrera-Rodriguez, C. R. Hill, G. Holmlund, K. Honda, U. D. Immel, S. Inokuchi, M. A. Jobling, M. Kaddura, J. S. Kim, S. H. Kim, W. Kim, T. E. King, E. Klausriegler, D. Kling, L. Kovacevic, L. Kovatsi, P. Krajewski, S. Kravchenko, M. H. Larmuseau, E. Y. Lee, R. Lessig, L. A. Livshits, D. Marjanovic, M. Minarik, N. Mizuno, H. Moreira, N. Morling, M. Mukherjee, P. Munier, J. Nagaraju, F. Neuhuber, S. Nie, P. Nilasitsataporn, T. Nishi, H. H. Oh, J. Olofsson, V. Onofri, J. U. Palo, H. Pamjav, W. Parson, M. Petlach, C. Phillips, R. Ploski, S. P. Prasad, D. Primorac, G. A. Purnomo, J. Purps, H. Rangel-Villalobos, K. Rebala, B. Rerkamnuaychoke, D. R. Gonzalez, C. Robino, L. Roewer, A. Rosa, A. Sajantila, A. Sala, J. M. Salvador, P. Sanz, C. Schmitt, A. K. Sharma, D. A. Silva, K. J. Shin, *et al.*, "Toward male individualization

with rapidly mutating y-chromosomal short tandem repeats," *Hum Mutat*, vol. 35, no. 8, pp. 1021–32, 2014.

[214] R. Nielsen, "A likelihood approach to populations samples of microsatellite alleles," *Genetics*, vol. 146, no. 2, pp. 711–6, 1997.

[215] I. J. Wilson and D. J. Balding, "Genealogical inference from microsatellite data," *Genetics*, vol. 150, no. 1, pp. 499–510, 1998.

[216] I. J. Wilson, M. E. Weale, and D. J. Balding, "Inferences from dna data: population histories, evolutionary processes and forensic match probabilities," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 166, no. 2, pp. 155–188

[217] M. Kayser, L. Roewer, M. Hedman, L. Henke, J. Henke, S. Brauer, C. Kruger, M. Krawczak, M. Nagy, T. Dobosz, R. Szibor, P. de Knijff, M. Stoneking, and A. Sajantila, "Characteristics and frequency of germline mutations at microsatellite loci from the human y chromosome, as revealed by direct observation in father/son pairs," *Am J Hum Genet*, vol. 66, no. 5, pp. 1580–8, 2000.

[218] Q. Y. Huang, F. H. Xu, H. Shen, H. Y. Deng, Y. J. Liu, Y. Z. Liu, J. L. Li, R. R. Recker, and H. W. Deng, "Mutation patterns at dinucleotide microsatellite loci in humans," *Am J Hum Genet*, vol. 70, no. 3, pp. 625–34, 2002.

[219] W. Wei, Q. Ayub, Y. Xue, and C. Tyler-Smith, "A comparison of y-chromosomal lineage dating using either resequencing or y-snp plus y-str genotyping," *Forensic Sci Int Genet*, vol. 7, no. 6, pp. 568–72, 2013.

[220] S. Kruglyak, R. T. Durrett, M. D. Schug, and C. F. Aquadro, "Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations," *Proc Natl Acad Sci U S A*, vol. 95, no. 18, pp. 10774–8, 1998.

[221] R. Sainudiin, R. T. Durrett, C. F. Aquadro, and R. Nielsen, "Microsatellite mutation models: insights from a comparison of humans and chimpanzees," *Genetics*, vol. 168, no. 1, pp. 383–95, 2004.

[222] T. D. Petes, P. W. Greenwell, and M. Dominska, "Stabilization of microsatellite sequences by variant repeats in the yeast saccharomyces cerevisiae," *Genetics*, vol. 146, no. 2, pp. 491–8, 1997.

[223] A. L. Bacon, S. M. Farrington, and M. G. Dunlop, "Sequence interruptions confer differential stability at microsatellite alleles in mismatch repair-deficient cells," *Hum Mol Genet*, vol. 9, no. 18, pp. 2707–13, 2000.

[224] J. Shao and C. F. J. Wu, "A general theory for jackknife variance estimation," *The Annals of Statistics*, vol. 17, no. 3, pp. 1176–1197, 1989.

[225] R. Durbin, "Efficient haplotype matching and storage using the positional burrows-wheeler transform (pbwt)," *Bioinformatics*, vol. 30, no. 9, pp. 1266–72, 2014.

[226] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, "Real-time dna sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133–8, 2009.

[227] J. Clarke, H. C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley, "Continuous base identification for single-molecule nanopore dna sequencing," *Nat Nanotechnol*, vol. 4, no. 4, pp. 265–70, 2009.

[228] W. S. Wong, B. D. Solomon, D. L. Bodian, P. Kothiyal, G. Eley, K. C. Huddleston, R. Baker, D. C. Thach, R. K. Iyer, J. G. Vockley, and J. E. Niederhuber, "New observations on maternal age effect on germline de novo mutations," *Nat Commun*, vol. 7, p. 10486, 2016.

[229] M. Ronemus, I. Iossifov, D. Levy, and M. Wigler, "The role of de novo mutations in the genetics of autism spectrum disorders," *Nat Rev Genet*, vol. 15, no. 2, pp. 133–41, 2014.