Characterizing variation at short tandem repeats and their role in human genome regulation

by

Melissa A. Gymrek

B.S. in Computer Science and Engineering and in Mathematics, Massachusetts Institute of Technology (2011)

Submitted to the Harvard-MIT Program in Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Author .....
Harvard-MIT Program in Health Sciences and Technology
February 1, 2016

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

Certified by .....
Mark J. Daly, PhD
Associate Professor Medicine, Harvard Medical School
Thesis Supervisor

Accepted by .....
Emery N. Brown, MD, PhD
Director, Harvard-MIT Program in Health Sciences and Technology/Professor of Computational Neuroscience and Health Sciences and Technology

Characterizing variation at short tandem repeats and their role in human genome regulation

by

Melissa A. Gymrek

Submitted to the Harvard-MIT Program in Health Sciences and Technology
on February 1, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

A central goal in genomics is to understand the genetic variants that underlie molecular changes and lead to disease. Recent studies have identified thousands of genetic loci associated with human phenotypes. These have primarily analyzed point mutations, ignoring more complex types of variation. Here we focus on Short Tandem Repeats (STRs) as a model for complex variation. STRs are comprised of repeating motifs of 1-6bp that span over 1% of the human genome. The level of STR variation and its effect on phenotypes remains mostly uncharted, mainly due to the difficulty in accurately genotyping STRs on a large scale.

To overcome bioinformatic challenges in high throughput STR genotyping, we developed lobSTR, an algorithm for profiling STRs from high throughput sequencing data. lobSTR employs a unique mapping strategy to rapidly align repetitive reads, and uses statistical learning techniques to account for STR-specific noise patterns. We applied lobSTR to generate the largest and highest quality STR catalog to date. This provided the first characterization of more than a million loci and gave novel insights into population-wide trends of STR variation. We used this catalog to conduct a genome-wide analysis of the contribution of STRs to gene expression in humans. This revelated that STRs explain 10-15% of the *cis* heritability of expression mediated by common variants and potentially play a role in various clinically relevant conditions.

Overall these studies highlight the contribution of STRs to the genetic architecture of quantitative traits. We anticipate that integrating the analysis of repetitive elements, specifically STRs, will lead to the discovery of new genetic variants relevant to human conditions.

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

Thesis Supervisor: Mark J. Daly, PhD
Title: Associate Professor Medicine, Harvard Medical School

Acknowledgments

This is the acknowledgements section. You should replace this with your own acknowledgements.

Contents

Chapter 1

Introduction

1.1 Overview

A central goal in genomics is to understand the genetic variants that underlie phenotypic changes and lead to disease. Recent studies have identified thousands of genetic loci associated with human phenotypes. Since the advent of next generation sequencing, the majority of genomic studies have focused on single nucleotide polymorphisms (SNPs). These are both the simplest type of variation to genotype and one of the easiest to model. However, a wide range of other classes of variants is important for controlling phenotypes.

Short tandem repeats (STRs) consist of period DNA motifs of 1-6bp and comprise of more than 1% of the human genome. Their repetitive structure induces DNA polymerase slippage events that add or delete repeat units, resulting in mutation rates that are orders of magnitude higher than those for most other variant types. STRs are implicated in more than 40 human diseases, mostly consisting of Mendelian disorders caused by large expansions of trinucleotide repeats. Additionally, several dozen single gene studies have shown that STRs can be involved in quantitative traits including gene expression. Because of their abundance, high polymorphism rates, and previous implication of functional roles, we focused on STRs as a model to investigate the role of complex variants in human phenotypes.

Here I present our contributions to enable the first large scale studies of STR variation and reveal that these loci play a significant role in complex traits in humans. In the first part of this thesis, we develop novel tools for high-throughput STR analysis from next generation sequencing data (**chapter 2**, **Appendix A**). We then apply these methods to large sequencing cohorts consisting of thousands of samples with diverse origins to provide the first population-wide catalog of hundreds of thousands of previously uncharacterized STR loci (**chapter 3**, **Appendix C**). An important aspect of this work has been a commitment to maximize utility of our results for the wider genomics community through providing open-source software packages

and online visualization tools that are alredy being utilized by other researchers. In the second part of this thesis, we interrogate the role of STRs in complex traits, focusing on gene expression as an initial phenotype (**chapter 4**). This study reveals more than 2,000 STRs whose lengths are correlated with gene expression (termed "expression STRs", or eSTRs) and shows that STRs make a significant contribution to regulating expression of nearby genes. These loci are enriched in putative regulatory regions and are predicted to modulate regulatory activity. These results highlight the contribution of STRs to the genetic architecture of gene expression and complex traits in humans.

To frame this work, I first review properties and applications of STRs and what we know about their contribution to disease and molecular phenotypes in humans. Next, I describe challenges in developing high throughput methods for STR analysis and state of the art experimental and bioinformatic methods for doing so. I summarize what we have learned to date about patterns of STR variation in humans using these methods. Then, I review what has been revealed about the genetic architecture of gene expression through SNP studies and the contribution of gene regulation to human conditons. I examine evidence that suggests variants such as STRs that are not well tagged by common SNPs may play an important role in these traits. Finally, I summarize the contributions of this thesis toward enabling large scale STR analysis and highlighting an important role fo STRs in human phenotypes.

1.2 STRs are an abundant source of genetic variability

For the purposes of this thesis, STRs are defined as short motifs of 1-6bp repeated in tandem. Note however that much of our work has excluded "homopolymers" with a motif length of 1 due to the difficult in genotyping these loci. chapter 3 gives a more precise definition of the requirements for a sequence to be called an STR and Table C.2 shows the relative abundance of each motif length. Using our definition, the human genome harbors more than one million STRs, enriched near genes and promoters [**?**], comprising over 1% of our genomes. This is likely an underestimate, as large regions inaccessible to current sequencing technologies are highly enriched for long STRs (see chapter 5).

The repetitive nature of STRs induces DNA polymerase slippage events that add or delete repeat units, resulting in mutation rates that are orders of magnitude higher than those for most other variant types [**?**, **?**]. Classical theoretical models have assumed that STRs mutate via a random

walk, or "simple stepwise model" whereby a locus is equally likely to gain or lose one or more repeats at each mutation [?]. A study by Sun *et al* [?] analyzing hundreds of individual STR mutations in trios found clear evidence that longer STR alleles are more mutagenic and more likely to mutate to shorter alleles, whereas shorter alleles were likely to expand, showing allele length plays a key role in governing the mutation process. Additional studies have found links between sequence features including STR length, motif length, base composition, and presence of STR sequence imperfections on STR heterozygosity [?], suggesting these features affect the mutation process.

1.3 Applications of STRs

STRs are an extremely abundant source of genetic polymorphism between individuals. Additionally, before the era of next generation sequencing, it was far cheaper to genotype STR lengths rather than simple sequence variations. As a result STRs have been the marker of choice for a number of human genetics applications:

- **Linkage analysis**: Linkage analysis is a method to map the chromosomal location of genetic variants associated with disease and other phenotypes. It relies on the fact that loci that are nearer to each other are less likely to be separated by recombination. Given a set of markers spaced along a chromosome and observed crossover events, one can narrow down an interval in which the locus of interest lies. This method was used for decades, before the widespread use of DNA sequencing, and so relied on either STR markers or restriction fragment length polymorphisms (RFLPs). The Marshfield Panel [?] contains thousands of STRs widely used in linkage analysis.
- **Forensics**: Unlike SNPs, which almost always have only two alleles segregating in the population, STRs often have ten or or more common alleles consisting of different repeat numbers. Therefore, the information content in a single STR is quite high, and genotypes for a small number of STRs can often identify a single individual (except of course in the case of identical twins). During the 1980s, the FBI Laboratory developed the CODIS set, consisting of 13 STR loci plus the AMEL marker to determine sex [?]. The National DNA Index database consists of more than 10 million CODIS profiles collected from arrested or convicted offenders.
- **Genetic genealogy**: Due to their high information content, STRs have also been widely

used by genealogists to infer genetic relationships between individuals. In particular, Y-chromosome STRs have proven useful for studying patrilineal ancestry due to the co-inheritance of Y chromosomes with surnames in Western societies. For example, if a man with the last name "Smith" has a son, he generally passes on both his Y chromosome and his surname. Genealogists have taken advantage of this fact and companies such as Family Tree DNA and others have compiled large databases connecting Y-STR profiles with surnames. By searching these databases with a male's Y-STR profile, one can reveal surnames of patrilineally related individuals. Y-STR profiles are also widely used in paterntiy testing or to identify male subjects in forensics cases.

Because of their identifying information, genotyping STRs and storing large databases of STR profiles raises significant privacy concerns. For instance, searching genealogy databases with the Y-STR profile of an unidentified male may identify a unique surname. Combined with additional information such as age or state of residence, this can in some cases narrow down the male's identity to a single individual. Indeed, there are several documented cases of male children conceived by anonymous sperm donation finding their biological fathers by genotyping their own Y-STRs [**?**, **?**]. This has important implications for human genetics studies, in which sample donors are assumed anonymous. Infact, we have shown that in many cases we could use Y-STRs obtained from publicly available whole genome sequencing datasets in addition to pedigree and geographic information to uniquely identify these individuals. This work is described in Appendix B.

1.4 Evidence of a regulatory role for STRs

Multiple *in vitro* studies have shown that STRs may regulate transcription. For instance, they have been shown to modulate transcription factor binding [**?**, **?**]. distance between promoter elements [**?**, **?**], and splicing efficiency [**?**, **?**]. It has also been shown that certain STRs may induce DNA to form noncanonical "Z"-DNA secondary structure, which can have an effect on transcriptional regulation of nearby genes [**?**].

Additionally, *in vivo* studies in model organisms have reported specific examples of STRs that modulate gene expression. For instance, experimental modification of the number of repeats in the promoter of the *FLO1* gene in yeast shows a direct quantitative effect on the cells' adherence to plastic [**?**]. An intronic expanded GAA repeat in *Arabidopsis thaliana* has an effect

on growth [?]. A 5' UTR STR in *avpr1a* predicts differences in socio-behavioral traits in prairie vole and modulates gene expression *in vitro* [?]. Finally, STR length variations in or nearby coding regiosn of *Alx-4* and *Runx-2* were shown to correlate with facial and limb lengths in certain canine species [?].

These single gene studies suggest STRs could provide a substrate for rapid evolution of fine-tunable gene expression regulation. Indeed, comparative genomics studies have found that the presence of STRs in promoters or transcribed regions is strongly associated with divergence of gene expression profiles across great apes [?]. This agrees with an earlier study showing similar trends across yeast strains [?].

Taken together, these anecdotes suggest that gene regulation by STRs is a widespread phenomenon that occurs across a range of taxa.

1.5 STRs in human disease and phenotypic variation

Dozens of STRs have been implicated in both disease and molecular phenotypes in humans. In nearly all of these cases, the resulting phenotype showed a quantitative relationship with the number of repeats, strongly suggesting STRs may play an important role in complex traits.

1.5.1 Dozens of disorders are caused by STR expansions

STR expansions are known to cause dozens of human single-gene, Mendelian disorders [?], affecting more than hundreds of thousands of patients in the U.S. [?]. The majority of these are due to expansions of trinucleotide repeats, and nearly all affect neurological function, most with a late onset disease course. For instance, an exonic CAG repeat expansion encoding polyglutamine results in Huntington's Disease, a devastating neurological disorder [?]; CGG expansion disrupts a methylation site on the X chrmosome leading to Fragile X Syndrome, one of the leading cuases of mental retardation in males [?]; a CUG expansion in the 3' untranslated (UTR) region *DPMK* results in myotonic dystrophy [?], a severe multisystemic form of muscular dystrophy. Other classes of repeats have been implicated in STR expansion diseases. Recently, high throughput sequencing scans for causative SNPs fortuitously revealed that a hexanucleotide expansion in *C9orf72* is responsible for 9p21-linked amyotrophic lateral sclerosis-frontotemporal dementia (ALS-FTD) [?]. A summary of STR expansion diseases is given by Mirkin [?].

The mechanisms by which most of these expansions lead to disease are still poorly understood. In cases of exonic repeat expansions, particularly polyglutamine expansions such as Huntington's Disease, it is thought that the expanded amino acid tracts form toxic aggregates that accumulate over time, consistent with the late-onset nature of these diseases [**?**]. Alternatively, it was recently shown that a key factor in Huntington's Disease may be repeat-length dependent aberrant splicing of *HTT* [**?**]. Other proposed pathogenic mechanisms include loss of protein expression, over-expression of the wildtype protein copy, and toxic gain of function of RNAs encoding expanded repeats [**?**]. Recent studies have found evidence that an alternative mechanism, "Repeat-associated non-ATG translation", termed "RAN-translation", may be responsible for rendering CAG repeats toxic. Under this phenomenon, expanded CAG repeats in RNA may be translated in the absence of an ATG start codon, and may produce transcripts under all seven possible reading frames [**?**]. These "RAN" transcripts have already been identified in patients with a variety of repeat disorders, including spinocerebellar ataxia type 8 (SCA8), myotonic dystrophy type 1 (DM1), Fragile-X tremor ataxia syndrome (FXTAS), and ALS-FTD [**?**].

One hallmark of repeat disorders is the phenomenon of "anticipation" in which the severity of the condition tends to increase and the age of onset tends to decrease with each generation. This generally corresponds to an in increase in repeat number for the expanded allele in each generation. Many of these repeats, experience a bi-phasic mutation process: under a certain number of repeats the region is stable with relatively low mutation rate. However, once a threshold length has been passed, the repeat will have an extremely high mutation rate and tend toward massive expansions [**?**]. In many cases, the number of repeats is directly related to phenotype. For instance, there is a negative linear relationship between CAG repeat number and age of onset of Huntington's Disease [**?**]. Interestingly, this suggests that unlike point mutations, which can serve as an "on/off" switch for Mendelian and other disorders, repeats possess a more fine-tunable mechanism to affect phenotype by adjusting the number of repeats on a quantitative, rather than a binary, scale.

Many such pathogenic STR expansions have been characterized, clearly indicating a biological function for at least a subset of repetitive elements. However, the majority of repeats remain uncharacterized and little is known about the extent of polymorphism and allele ranges at these and other STRs in healthy individuals. As mentioned above, STRs are prone to replication slippage events that cause them to mutate rapdily. Whereas intermediate length STRs of $\sim<$200bp tend to mutate in a stepwise fashion. longer repeats such as those involved in expansion disorders may for unusual non-B DNA structures [**?**] that cause rapid expansions, leading to unusual behavior

in pathogenic conditions. We hypothesize that whereas longer unstable repeats radically disrupt function of a given locus, leading to severe mono-genic disorders, STRs in the intermediate range length may be responsible for fine-tuning genomic regulation and for contributing to more incremental variation, which could make an important contribution to more complex traits in humans.

1.5.2 STRs in complex human traits

Little is known about the role of STRs in more complex, polygenic traits. This is largely due to the fact that until recently there was limited ability to systematically profile these loci on a large scale (see 1.6) and because SNP-based studies have limited ability to capture STR associations (see 1.8). However, a small number of STR associations with complex traits have been reported: repeat length in the first exon of the androgen receptor correlates with risk of hepatocellular carcinoma risk in women [?]; a TC repeat in *HMGA2* is associated with uterine leiomyomata and decreased height [?]; a CAG repeat in *KCNN3* is associated with cognitive performance in schizophrenia [?]. However, no study has systematically evaluated the role of STRs in complex traits.

1.5.3 Mechanisms for STR involvement in genome regulation

STRs are found in at least 5% of human protein-coding genes [?] and are abundant in intragenic regions and UTRs [?]. These rapidly evolving elements provide an evolutionary substrate to incrementally affect gene activity without introducing major sequences changes. STRs have been demonstrated or hypothesized to affect gene regulation in several ways, summarized below.

Transcribed STRs

As described above, many putative pathogenic STRs lie in coding regions and may lead to disease through mechanisms such as protein or RNA aggregation, RAN-translation, or aberrant splicing. Alternative mechanisms could allow exonic STRs to affect protin function. For instance, STR mutations in coding regions could function to silence genes by introducing frameshift mutations leading to premature stop codons [?], especially in the case of non tri- or hexa-nucleotide repeats. Additionally, expansions or contractions could alter spacing between protein domains, leading to a change in function.

Intronic repeats may also affect gene regulation. Lengths of intronic STRs have been shown to affect gene expression [**?**], in some cases through altering transcription factor binding sites. For example, a TATC repeat in *TH* affects binding of the transcription factor ZNF191 [**?**]. Several studies have also shown that intronic STRs may regulate splicing efficiency in a repeat-specific manner. For instance, the length of an intronic TG repeat in *CFTR* is directly related with inclusion of the adjacent exon [**?**], which is likely to be due to effects of the repeat on RNA secondary structure. An intronic CA repeat in *eNOS* was shown to regulate splicing by affecting binding of the splicing factor HnRNP L [**?**], and it was later shown that intronic CA repeats may have a widespread effect on splicing [**?**].

STRs in UTRs may affect gene regulation. For example, a CTG/CAG repeat in the 3'UTR of *DMPK* is implicated in the repeat expansion disorder myotonic dystrophy 1, and is thought to sequester splicing factors leading to aberrant splicing [**?**]. 3'UTR repeats may also harbor microRNA binding sites, affecting gene regulation.

STR lengths influence promoter and enhancer activity

In addition to transcribed regions, STRs can affect the function of genomic regulatory elements through several mechanisms. AC dinucleotides are over-represented in predicted *cis*-regulatory elements [**?**]. Recently, it was shown that dinucleotide repeats are a hallmark of enhancer elements in *Drosophila* and human cell lines [**?**]. This suggests they may provide an abundant source of transcriptional regulation in these elements. Heidari, *et al.* [**?**] demonstrated that a GA repeat in the promoter of *SOX5* can affect nucleosome processing, affecting downstream gene expression. This finding is supported by a study in yeast that found variation in repeat length had a strong effect on nucleosome positioning and gene expression in 25 of 33 randomly chosen STR-containing promoters [**?**]. STRs may also bind transcription factors and create a number of binding sites dependent on the number of repeats. Guillon *et al.* found that the oncogenic EWSR1-FLI1 fusion protein formed in Ewing Sarcoma preferentially binds GGAA repeats [**?**]. Interestingly, a recently study found that a reported genome-wide association study (GWAS) signal for Ewing Sarcoma actually points to a SNP for which the alternate allele joins two adjacent GGAA repeats into one long repeat tract, resulting in overexpression of the nearby gene *EGR2* [**?**]. In another example, a pentanucleotide repeat in the promoter of *PIG3* creates a varying number of *p53* binding sites, affecting downstream expression [**?**].

Together, these examples show clear evidence that STRs may regulate quantitative traits, such

as gene expression, transcription factor binding, and splicing efficiency, in a repeat-dependent manner and therefore are prime candidates for contributing to complex traits in humans.

1.6 Methods for genotyping STRs

1.6.1 Capillary electrophoresis

The current gold standard technique for STR typing relies on cumbersome capillary electrophoresis methods [**?**], which require PCR amplification of the locus of interest, followed by size separation via electrophoresis to determine the alleles present. The method is widely used for typing STRs used as genetic markers for linkage analyses (Marshfield set [**?**]), the FBI CODIS set, markers for genealogical studies [**?**, **?**] and for determining alleles present at known pathogenic loci [**?**, **?**].

Capillary electrophoresis uses a separate reaction per locus, and requires time-consuming optimization of conditions for each reaction. With a cost that is upwards of several dollars per STR and lengthy preparation, current panels can consist of up to several hundreds of STRs - only a fraction of the hundreds of thousands of STR loci in the human genome [**?**]. While the method is robust and can achieve accuracy above 99% [**?**], it has several technical limitations. First, many loci, especially dinucleotide STRs, are plagued by "stutter peaks" due to errors introduced during PCR amplification that may complicate calling. Second, this technique can only return the size of the allele: it is unable to distinguish homoplasmic alleles [**?**] (two different alleles with the same size but distinct sequences). Third, because capillary techniques simply return the length of the amplified region, it will be sensitive to the presence of linked insertions or deletions that are not part of the STR itself being genotyped. Due to these limitations, electrophoresis is unable to provide an accurate genome-wide picture of sequence variation at STR loci.

1.6.2 Genotyping STRs from next-generation sequencing

Challenges of genotyping STRs from short reads

Although theoretically any type of variation should be captured by DNA sequencing, STRs have proven challenging to genotype from short reads produced by high-throughput sequencing platforms. Major challenges include:

1. Reads must entirely span an STR region to be informative about the number of repeats present.
2. STRs with a large length difference from the reference sequence present as a gapped alignment problem. The run time of mainstream aligners such as BWA [?] increases rapidly with the number of insertions or deletions allows.
3. As in capillary techniques, sequencing technologies require PCR amplification of the DNA sample, which can introduce false "stutter peaks" due to the same polymerase slippage process leading to germline STR mutations.

As a result of these challenges, STRs are not routinely analyzed in sequencing studies [?], and loci containing repetitive loci are frequently filtered out due to the high presence of genotyping errors in these regions.

A major contribution of our work was develop the first efficient algorithm for generating accurate STR genotypes, called lobSTR [?] and described in chapter 2. Over the last several years, additional bioinformatic tools and long read sequencing technologies have arisen that have the potential to greatly increase STR calling. These are discussed in the conclusion (chapter 5)

Challenges in visualizing complex variatns

A key aspect of developing and using tools for variant analysis is visualization of sequence alignments. Often, inspecting raw read alignments can be informative of systamic sequencing artifacts leading to errneous genotype calls.Currently, the UCSC Genome Browser [?] and the Integrative Genomics Viewer (IGV) [?] are the most widely used genome browsers for alignment visualization.

Visualization of insertions and deletions are key to analyzing reads containing STR variations from the reference genome. However, UCSC, IGV, and similar tools are limited in their ability to display insertions from the reference genome. Because the display is based entirely on the reference genome, insertions are simply displayed as a vertical bar, with no information about the length of insertions. As a result, reads containing different insertions consisting of different lengths, such as a diploid locus where both alleles are longer than the referene allele, will be displayed identically. To overcome this and other genome browser challenges, I have created a novel application, PyBamView, for visualizing sequence alignments at complex variants. This contribution is described in Appendix A.

1.7 Population-wide characterization of STR variation

Several large panels of STR variation have been previously generated using capillary electrophoresis. To the best of our knowledge, the largest panels are from the Rosenberg Lab (https://rosenberglab.stanford.edu/data/rosenbergEtAl2005/) which contains genotypes for 993 STRs in 1,048 individuals from the Human Genome Diversity Project (HGDP) (http://www.hagsc.org/hgdp/) and the Payseur Lab (http://payseur.genetics.wisc.edu/strpData.htm), with 721 STRs in 201 individuals from the HapMap Project. Both panels use a subset of the Marshfield marker set originally used for linkage analysis. In addition to autosomal STRs, 61 Y-STRs have been genotyped in 669 HGDP samples (ftp://ftp.cephb.fr/hgdp_supp9/) and for 16 Y-STRs in 49 HapMap Samples [**?**].

These panels have shown that STRs are informative of ancestry and can accurately capture population structure. Rosenberg *et al.* found that STRs could be used to cluster individuals by geographic regions and individual populations [**?**], in strong agreement with self-reported ancestries. He *et al.* showed that Y-STRs were informative of geographic region of origin of HapMap samples [**?**]. Additionally, Y-STRs can be informative of specific historical events. For instance, Y-STR analysis has identified haplotypes likely to have descended from Genghis Khan [**?**] and haplotypes specific to the Cohen group of Jewish priests [**?**].

Capillary electrophoresis panels have been valuable in inferring a number of features of STR variability, such as the dependence of mutation rate on repeat unit length [**?**] and repeat tract length [**?**]. However, the STRs used in these panels represent a small fraction of all STRs in the genome, and consist almost entirely of di- and tetranucleotide repeats. Moreover, they have been particularly chosen due to their high polymorphism rates and because they are straightforward to genotype, likely imposing ascertainment biases that may confound analyses.

1.8 Genetic architecture of gene expression and complex traits

1.8.1 Expression quantitative trait loci

challenges in doing this GTEx, Pritchard, etc.

1.8.2 Additional types of eQTLs

1.8.3 Heritability of gene expression

Wright, UK10K, Price

1.8.4 The role of gene regulation in complex traits

1.8.5 Power of SNP studies to capture STRs

Gaurav (here or below)

1.9 Contributions of this thesis

Taken together, the studies described above provide strong evidence that STRs may play a widespread role in quantitative traits in humans. At the outset of this work, STRs were not amenable to large scale studies due to limitations in bioinformatic and sequencing technologies. Furthermore, repetitive regions were largely considered neutral, or "junk DNA" that were simply filtered out of most analyses.

In this thesis, I present our work to enable and perform genome-wide analysis of the effect of STRs on gene expression, and ultimately complex traits, in humans. These efforts have shown that STRs indeed contribute significantly to the heritability of gene expression and likely have a widespread effect on genome regulation. Importantly, the tools and results generated by this work have helped to renew interest in variants beyond point mutations and have highlighted the importance of including repetitive variations in genome-wide studies of complex traits.

Specifically, I present the following studies, briefly summarized here:

1.9.1 Tools for STR analysis from short reads

STRs pose major challenges to current bioinformatics pipelines. They are often extremely polymorphic and exhibit large length differences from the reference sequence. Additionally, they are prone to stutter errors that occur due to polymerase slippage during PCR amplification. As the first step to studying STRs on a large scale, we developed lobSTR [?] chapter 2, an

algorithm for profiling short tandem repeats from high throughput sequencing data. lobSTR employs a unique mapping strategy to rapidly align repetitive STR containing reads, and uses statistical learning techniques to account for stutter noise.

We have invested significant effort in making this tool user friendly, and lobSTR now has an active online user community on Github and Google groups. lobSTR has been used to generate large scale catalogs of STR variation from the 1000 Genomes and other projects (see below). We continue to improve lobSTR to adapt to improvements in sequencing technology and bioinformatics tools. For instance, lobSTR now handles input from the BWA-MEM aligner, has a specific noise model for PCR-free protocols, and uses joint calling to simultaneously genotype thousands of samples at once.

In addition to lobSTR itself, we are actively developing additional tools for analysis and visualization of STR and other complex variant data. To deal with challenges in visualization STRs and other complex variants, I developed PyBamView [**?**] Appendix A, an interactive alignment visualization tool. PyBamView allows accurate visualization of insertions, which are not displayed well by current browsers such as IGV and UCSC, and allows researchers to share alignment views with others through a web browser.

1.9.2 The first genome-wide catalogs of STR variation

Until recently, studies of STR variation have been limited to several highly ascertained sets of loci encompassing several thousand highly polymorphic autosomal and Y-chromosome STRs used in linkage studies, genetic genealogy, and forensics. Little is known about patterns of polymorphism across the majority of the 1.6 million STR loci in the human genome. Building a catalog of STR variation is an important first step in determining their contribution to genetic and phenotypic variation.

To build an initial genome-wide STR catalog, we applied lobSTR to more than 3,000 low coverage whole genome sequencing datasets generated by the 1000 Genomes Project [**?**] chapter 3. We used this call set to analyze sequence determinants of STR variation, assess patterns of variation in coding regions, and find common loss of function alleles. We found that hundreds of thousands of STRs are polymorphic in the number of repeats across individuals. This catalog has already served as a valuable resource for researchers interested in variation at specific STR loci that may be implicated in human phenotype.

While our initial catalog gave accurate information on the distribution of STR alleles across populations, the quality of individual genotypes was poor due to very low sequencing coverage. Additionally, we were not able to obtain accurate genotypes for homopolymers, which show notoriously high error rates due to polymerase slippage. We recently created the most comprehensive STR call set to date based on 300 deeply sequenced genomes from the Simons Genome Diversity Project Appendix C. These calls show 93% concordance with gold standard STR genotypes from capillary electrophoresis, and accurately capture population structure. This dataset provides unprecedented opportunities to study STR variation that were not possible using previous studies either due to the small number of markers or to the low quality of individual genotypes, including in-depth study of homopolymers and population dynamics of STR variation.

1.9.3 Abundant contribution of STRs to gene expression in humans

With a robust pipeline for STR genotyping and a large catalog of STR genotypes, we were in a position to assess the genome-wide contribution of STRs to phenotypes for the first time. We focused on the role of STRs in regulating gene expression. Multiple single-gene studies in humans and other organisms have shown that STRs may modulate gene expression in *cis*. However there has been no systematic study of their effect on expression in humans.

We conducted a genome-wide analysis of STRs that affect expression of nearby genes, which we termed expression STRs (eSTRs), in lymphoblastoid cell lines (LCLs) [**?**] chapter 4. This well-studied model permitted the integration of whole genome sequencing data, expression profiles from RNA-sequencing and arrays, and functional genomics data. We tested for association in close to 190,000 STR$\times$gene pairs and found over 2,000 significant eSTRs. Using a multitude of statistical genetic and functional genomics analyses, we show that hundreds of these eSTRs are predicted to be functional, and that 10-15% of *cis* heritability of gene expression in LCLs can be attributed to STRs, uncovering a new class of regulatory variants.

1.9.4 Identifying personal genomes by surname inference

Sequencing datasets are often shared without identifiers under the assumption that the identities of the samples will not be revealed. We showed that we can recover genealogical STRs from the Y chromosome by applying lobSTR to whole genome sequencing datasets. We used the resulting genotypes to query recreational genealogical databases to recover the surnames of anonymous

sample donors. Combining the surname with additional metadata such as age and state can lead to a complete breach of anonymity [?] Appendix B.

Due to the widespread use of STRs in forensics and genealogy, this work had important implications for genetic privacy of human research subjects. As a result I have become more aware of ethical concerns and have been an active participant in discussions of social aspects of genetic research in the wider genomics community. Importantly, it is now recognized that it is impossible to promise anonymity, and that this should be clarified to participants. As a result of our study many data usage agreements have been augmented to prevent researchers from attempting to identify participants. We believe this project paved the way toward greater transparency about data privacy issues and will ultimately help protect and inform participants in genetic research.

Chapter 2

lobSTR: A short tandem repeat profiler for personal genomes

Most of this chapter was first published as:

> **Gymrek M**, Golan D, Rosset S, Erlich Y. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Research*. (2012).
>
> **Gymrek M**, Erlich Y. Profiling short tandem repeats from short reads. Book chapter in *Deep Sequencing Data Analysis* by *Methods Mol Biol* (2013).

Part of this work is covered by a U.S. patent:

> Analyzing short tandem repeats from high throughput sequencing data for genetic applications. (2013). Inventors: Erlich Y, **Gymrek M**. US 2014/0163900 A1.

Abstract: Short Tandem Repeats (STRs) have a wide range of applications, including medical genetics, forensics, and genetic genealogy. High throughput sequencing (HTS) has the potential to profile hundreds of thousands of STR loci. However, mainstream bioinformatics pipelines are inadequate for the task. These pipelines treat STR mapping as gapped alignment, which results in cumbersome processing times and a biased sampling of STR alleles. Here, we present lobSTR, a novel method for profiling STRs in personal genomes. lobSTR harnesses concepts from signal processing and statistical learning to avoid gapped alignment and to address the specific noise patterns in STR calling. The speed and reliability of lobSTR exceed the performance of current mainstream algorithms for STR profiling. We validated lobSTR's accuracy by measuring its consistency in calling STRs from whole genome sequencing of two biological replicates from the same individual, by tracing Mendelian inheritance patterns in STR alleles in whole-genome sequencing of a HapMap trio, and by comparing lobSTR results to traditional molecular techniques. Encouraged by the speed and accuracy of lobSTR, we used the algorithm to conduct a

comprehensive survey of STR variations in a deeply sequenced personal genome. We traced the mutation dynamics of close to 100,000 STR loci and observed more than 50,000 STR variations in a single genome. lobSTR's implementation is an end-to-end solution. The package accepts raw sequencing reads and provides the user with the genotyping results. It is written in C/C++, includes multi-threading capabilities, and is compatible with the BAM format.

lobSTR is available at https://github.com/mgymrek/lobstr-code.

2.1 Introduction

Short tandem repeats (STRs), also known as microsatellites, are a class of genetic variations with repetitive elements of 2 to 6 nucleotides that consists of approximately a quarter million loci in the human genome [?]. The repetitive structure of those loci creates unusual secondary DNA conformations that are prone to replication slippage events and result in high variability in the number of repeat elements [?]. The spontaneous mutation rate of STRs exceeds that of any other type of known genetic variation, and can reach 1/500 mutations per locus per generation [?, ?], 200 fold higher than the rate of spontaneous copy number variations (CNV) [?] and 200,000 fold higher than the rate of de novo SNPs [?].

STR variations have been instrumental in wide-ranging areas of human genetics. STR expansions are implicated in the etiology of a variety of genetic disorders, such as Huntington's Disease and Fragile-X Syndrome [?, ?]. Forensics DNA-fingerprinting relies on profiling autosomal STR markers and Y-chromosome STR (Y-STR) loci [?]. STRs have been extensively used in genetic anthropology, where their high mutation rates create a unique capability to link recent historical events to DNA variations, including the well-known Cohen Modal Haplotype that segregates in patrilineal lines of Jewish priests [?, ?]. Another relatively recent application of STR analysis is tracing cell lineages in cancer samples [?].

Despite the plurality of applications, STR variations are not routinely analyzed in whole genome sequencing studies, mainly due to a lack of adequate tools [?]. STRs pose a remarkable challenge to mainstream HTS analysis pipelines. First, not all reads that align to an STR locus are informative (**Supplemental Figure 2.8.1A**). If a single or paired-end read partially encompasses an STR locus, it provides only a lower bound on the number of repeats. Only reads that fully encompass an STR can be used for exact STR allelotyping. Second, mainstream aligners, such as BWA, generally exhibit a trade-off between run time and tolerance to insertions/deletions

(indels) [?]. Thus, profiling STR variations – even for an expansion of three repeats in a trinucleotide STR – would require a cumbersome gapped alignment step and lengthy processing times (**Supplemental Figure 2.8.1B**). Third, PCR amplification of an STR locus can create stutter noise, in which the DNA amplicons show false repeat lengths due to successive slippage events of DNA polymerase during amplification [?, ?] (**Supplemental Figure 2.8.1C**). Since PCR amplification is a standard step in library preparation for whole genome sequencing, an STR profiler should explicitly model and attempt to remove this noise to enhance accuracy.

Here, we present lobSTR, a rapid and accurate algorithm for STR profiling in whole genome sequencing datasets (**Figure 2.1**). Briefly, the algorithm has three steps. The first step is sensing: lobSTR swiftly scans genomic libraries, flags informative reads that fully encompass STR loci, and characterizes their STR sequence. This ab initio procedure relies on a signal processing approach that uses rapid entropy measurements to find informative STR reads followed by a Fast Fourier Transform to characterize the repeat sequence. The second step is alignment: lobSTR uses a divide and conquer strategy that anchors the non-repetitive flanking regions of STR reads to the genome to reveal the STR position and length. We use a modified reference that takes advantage of the information extracted from the sensing step to increase the alignment specificity. This step avoids a cumbersome gapped alignment and, importantly, is virtually indifferent to the magnitude of STR variations. Finally, in the third step, the pipeline allelotypes the STRs using a statistical learning approach that models the stutter noise in order to enhance the signal of the true allelic configuration. See **Supplemental Text 2.6**, **Supplemental Figures 2.8.2**, **2.8.3**, **2.8.4**, **2.8.5** and **Supplemental Table 2.9.1** for full details about the lobSTR algorithm.

lobSTR implementation offers a complete solution that takes raw sequencing data and reports the alleles present at each profiled STR locus. The program's input is one or more sequencing libraries in FASTA/FASTQ or BAM format. The output is the alignment of STR reads in BAM format and the most likely alleles for each STR locus in a custom tab-delimited text format. lobSTR supports multi-threaded processing.

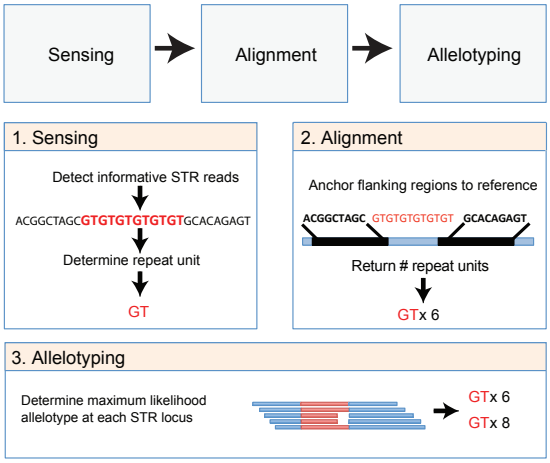


Figure 2-1: **lobSTR algorithm overview** lobSTR consists of three steps. The sensing step detects informative STR reads and determines their repeat motif. The alignment step maps the STRs' flanking regions to the reference. The allelotyping step determines the STR alleles present at each locus.

2.2 Results

2.2.1 Comparing lobSTR to mainstream aligners

We benchmarked lobSTR's alignment performance with reads from an Illumina whole genome sequencing library with 101bp reads (**Methods** 2.4). To demonstrate its added value for STR profiling over mainstream aligners, we also ran BWA, Novoalign, and Bowtie on the same input data with and without the GATK local indel realignment tool. In addition, we ran BLAT [**?**] to characterize STR alignment by a tool that is centered on sensitivity rather than speed. BWA and Novoalign were tested with the default parameters that can detect up to 5bp and 7bp indels, respectively. Bowtie has no indel tolerance and was evaluated as a control condition with tolerance of up to two mismatches. BLAT was tested with the default parameters that can tolerate up to 10% divergence from the reference, which corresponds to approximately 10bp indels. To focus on the pure algorithm speed-up, all tests were executed on a single CPU.

lobSTR excelled in all the parameters required for efficient STR alignment (**Table** 2.2.1). First, lobSTR processed the reads 2.2 times faster than Bowtie, 22 times faster than BWA, 70 times faster than Novoalign, and almost 1000 times faster than BLAT (**Figure** 2.2.1**A**). These

results indicate that there is a minimal computational payment in running lobSTR in parallel to mainstream aligners in order to augment variation calling to include STR polymorphisms. Second, as required, lobSTR reported only informative reads that fully encompass STR loci. On the other hand, the mainstream aligners reported between 2,000 to 5,000 non-informative STR reads per million input sequences, which may confound downstream calling algorithms if not removed. Third, lobSTR detected the largest number of informative reads with STR variations compared to mainstream aligners (**Figure 2.2.1B**). The other aligners showed a strong tendency to report STR reads with the reference allele vis-a-vis with their indel tolerance. Bowtie did not report any STR variation. After GATK local realignment, BWA and Novoalign, respectively, reported that 20% and 25% of the informative reads have STR variations. BLAT reported that 37% of the informative reads have STR variations, compared to 50% in lobSTR. Analyzing data collected from a large number of randomly ascertained STR loci [**?**] (Utah Marker Development Group) demonstrates that 33-66% of STR sequence reads should exhibit a non-reference allele (see Methods). This suggests that lobSTR's results are more representative of the true rate of STR variations than mainstream alignment tools.

Algorithm	Indel tolerance (bp)	Time (sec)	# Non-informative reads	# Informative reads	# Var. reads[a]	Ratio[b]	Peak memory (Gbyte)
lobSTR	-	109	0	973	485	0.5	0.3
Bowtie	0	258	2,193	523	0	0	2.2
BWA	5	2,450	3,026	883	174	0.19	2.5
BWA+GATK	5	2,943	2,691	869	172	0.20	2.5
Novoalign	7	7,601	4,947	1,024	208	0.2	13.8
Novoalign+GATK	7	8,123	4,906	1,047	259	0.25	13.8
BLAT	10	108,862	19,919	1,611	602	0.37	3.7

Table 2.1: **STR Alignment performance across different algorithms.** All results are per million 101bp Illumina reads. [a]Number of informative reads that show a non-reference allele. [b]Ratio of reads with the non-reference allele versus total informative STR reads.

Reporting STR reads with non-reference alleles is crucial for profiling pathogenic mutations. We further explored whether lobSTR can correctly detect disease alleles of dominant trinucleotide repeat expansion disorders. As test cases, we focused on two conditions that can be theoretically profiled using standard Illumina runs. The first condition was a GCN expansion in *PABPN1* that causes oculopharyngeal muscular dystrophy (OPMD) [**?**], where the normal allele exhibits 10 repeats and the pathogenic allele spectrum for the dominant form is between 12 to 17 repeats [**?**]. The second condition was a GCG expansion in *HOXD13* that is implicated in synpolydactyly [**?**], a severe limb malformation, where the normal allele is 15 repeats and the documented pathogenic allele spectrum is between 22-29 repeats [**?**]. To simulate each condition, we generated 100 reads of length 101bp that were equally sampled from the disease locus consisting of a normal and pathogenic allele with 100bp flanking upstream and downstream regions with 1% sequencing

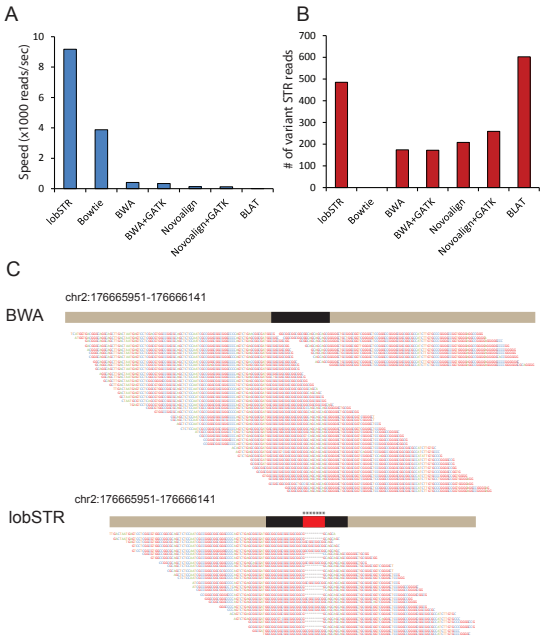


Figure 2-2: **lobSTR shows an added value for STR profiling over mainstream techniques (A) Alignment speed (reads per second) of lobSTR, mainstream alignment and BLAT**. lobSTR processes reads between 2.5 and 1000 times faster than alternative methods **(B) The sensitivity of detecting STR variations of different alignment strategies.** Only BLAT detected more STR variations than lobSTR **(C) lobSTR accurately detects pathogenic trinucleotide expansions that are discarded by mainstream aligners.** The figure shows simulation results of the *HOXD13* heterozygous locus with a normal and a pathogenic allele that contains 7 additional alanine insertions. BWA reports only the normal allele. Reads exhibiting a pathogenic STR expansion are not detected. lobSTR identifies both alleles present at the simulated locus. All positions are according to hg18.

error rate. For both simulated disease conditions, lobSTR accurately aligned the normal and pathogenic reads to the correct location in the genome. All aligned reads were informative and the allelotyping step correctly assigned a heterozygous state to the disease loci with the correct repeat lengths: (10,15) for *PABPN1* and (15, 22) for *HOXD13*. In stark contrast, BWA failed to correctly align reads from the pathogenic alleles of both loci. Only reference reads were reported (**Figure 2.2.1C**).

2.2.2 Measuring lobSTR concordance using biological replicates

To explore the precision of lobSTR, we conducted genome-wide STR profiling of blood and saliva samples from the same individual [?]. These samples were sequenced using Illumina HiSeq2000 with 101bp PE to a mean autosomal coverage of 50x and 102x, respectively. lobSTR ran with default parameters on 20 CPUs and analyzed the two datasets within 12 and 22 hours respectively. After filtering loci with low quality calls, 143,793 shared STRs were covered in the two datasets with at least one read and 79,771 STRs were covered with 10 reads or more (**Figure 2.2.2A**).

We quantified the rate of discordant autosomal calls between the two samples. We focused on two measurements: the genotype discordance rate and the allelic discordance rate [?]. The former reports as an error any mismatch between corresponding calls, whereas the latter reports only the fraction of discordant alleles in corresponding calls. For example, consider a locus that is called (A,B) in the saliva sample and (A,C) in the blood sample. This locus shows a single genotype discordance, but only 0.5 allelic discordance, since the A allele was correct.

Both types of error greatly diminished with sufficient coverage (**Figure 2.2.2B,C**). At 5x coverage, the genotype discordance rate was 11% and the allelotype discordance was 5%; At 21x coverage, the genotype discordance rate was 3% and the allelotype discordance rate was 2%. Similar to STR studies with capillary platforms [?], most of the errors were generated in dinucleotide STR loci, whereas other types of STRs showed moderate and similar error rates. The dinucleotide error rates presumably stem from two factors: first, these loci usually show the highest heterozygosity rates [?, ?, ?]. Therefore, they require on average more sequence reads to be correctly called. Second, dinucleotide STRs are more prone to stutter noise [?] and their higher error rates might be due to residual noise after lobSTR stutter deconvlution.

We further analyzed the STR length differences in discordant calls. To avoid analyzing errors that are simply due to allele drop-outs, we focused on discordant calls that were both heterozygous in blood and saliva. At a coverage of âĽě5x, more than 90% of the errors showed a single repeat unit difference and 99% of the errors were within two repeat units (**Figure 2.2.2D**). This indicates that incorrect alignment of STRs has a minimal effect on allelotyping results, and that stutter is likely the main source of noise. We also found that only 0.8% of calls at heterozygous loci showed a difference due to an incomplete repeat unit. This highlights that lobSTR can determine STR alleles at a single base-pair resolution.

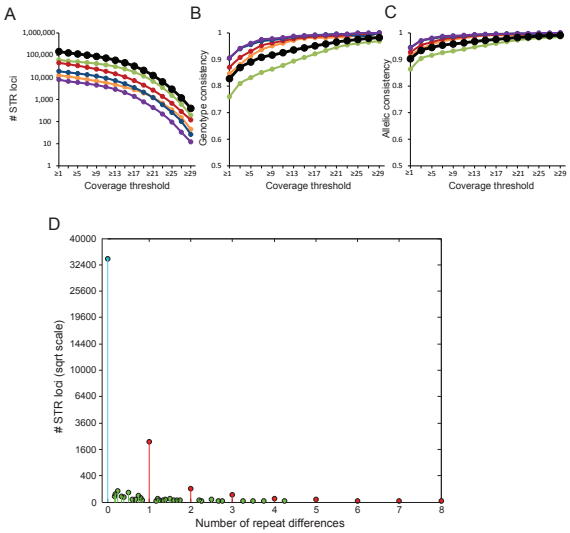


Figure 2-3: **Measuring lobSTR consistency from two samples of the same individual** (A-C) (green - period 2, orange - period 3, red - period 4, blue - period 5, purple - period 6, black - all) **(A) Loci covered in both samples at increasing coverage thresholds. (B) The genotype discordance rate as a function of coverage threshold. (C) The allelic discordance rate as a function of coverage threshold. (D) Number of repeat differences at heterozygous loci.** Blue - no difference; red - integer numbers of repeat differences; green - non-integer numbers of repeat differences. Most discordance calls consist of a single repeat unit difference between calls in the two samples. Distance was measured as the second minimum distance between alleles of the two samples. The y-axis is given in a square root scale.

2.2.3 Tracing Mendelian inheritance using lobSTR

To further explore lobSTR performance, we conducted a genome-wide STR profiling of a HapMap trio - a father (NA12877), mother (NA12878), and son (NA12882) – from the CEU population that were sequenced using 100PE reads on a HiSeq2000 (**Table 2.2.3**). The average autosomal coverage was 50x and average STR coverage was 14x. At $\geq$10x coverage threshold, 57% of the STRs in the CEU trio had a non-reference allele.

Individual	Relationship	Input reads	STR Aligned reads	Mean STR Coverage
NA12878	Mother	1,708,169,546	3,398,933	14.8
NA12877	Father	1,637,816,924	3,212,073	14.1
NA12882	Son	1,625,404,856	3,183,795	14.0

Table 2.2: **Profiling STRs in Illumina reads from a HapMap trio.**

In general, deviations of offspring's STR alleles from Mendelian inheritance (MI) indicate a potential calling error [**?**]. With 5x coverage across all trio members, the MI rate was 95%; with 10x coverage, MI was 97%; and with coverage threshold of 15 or more, MI was 99%. (**Figure 2.2.3A**). We also repeated the analysis only with discordant parental sites (for example, A/B call in one parent and A/C call in another parent). We noticed a drop to 93% in the MI patterns with a low coverage threshold of 5x, which is mainly because of partial coverage of heterozygous sites in the parents. The MI rate was recovered to the same level with higher coverage threshold. At 17x coverage 99% of sites showed a perfect Mendelian segregation pattern (**Figure 2.2.3B**).

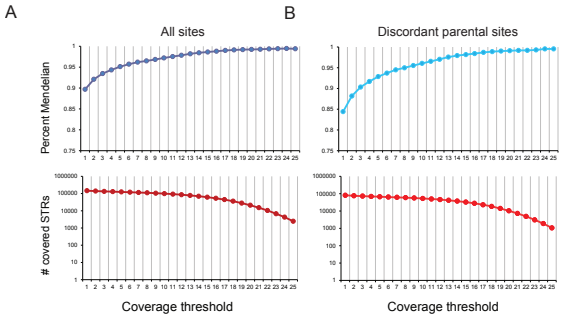


Figure 2-4: **Validating lobSTR by Mendelian inheritance in a HapMap trio.** Mendelian inheritance (blue) rose to 99% above 17x coverage. The number of covered loci at each coverage threshold is shown in red. **(A) Mendelian inheritance of all covered loci. (B) Mendelian inheritance of loci with discordant parental allelotypes.**

2.2.4 Validating lobSTR accuracy with DNA electrophoresis

We sought to compare lobSTR calls to the results of DNA electrophoresis, which is considered the gold standard for STR allelotyping. First, we focused on a set of STR markers that are used for forensic DNA fingerprinting. As an input for lobSTR, we sequenced a male genome from our lab collection with three runs of Illumina GAIIx for 101PE cycles that yielded 740 million reads. The autosomal sequencing coverage was 36x according to alignment with mainstream algorithms. lobSTR identified 1.6 million informative reads that mapped to ~140,000 STR loci, with an average of 4.91x coverage of diploid STR loci. In parallel, we used a commercial forensic kit to genotype 14 autosomal STR markers on a capillary electrophoresis platform. Thirteen out of 14 markers were covered by at least a single sequence read and 8 markers were covered by at least 3 sequence reads. The marker that was not covered spanned more than 129bp, exceeding

the limit for detecting informative reads with the 101bp sequence reads.

We observed good concordance between lobSTR and the capillary results (**Table 2.2.4**). lobSTR correctly called all but one of the 8 markers that were covered by at least 3 reads and most of the alleles in loci that were covered with 2 or less reads. Remarkably, some of these markers, such as D8S1179, displayed two heterozygous alleles that did not match the reference. Other alleles, such as in Penta D and Penta E, correctly returned 20bp and 25bp length differences from the reference allele, respectively. The capillary results of one tetranucleotide marker, THO1, exhibited a non-integer number of copies (9 repeats + 3bp). lobSTR reported exactly the same results, further demonstrating that STRs can be called within a single base pair resolution. lobSTR also correctly called a homozygous STR that was covered by a single read. In another 4 markers with coverage of âĽd'2x, lobSTR correctly called one allele and missed the other allele due to sequencing coverage. We observed only a single erroneous call due to stutter noise in the D5S818 locus. This homozygous locus was covered by three sequence reads: two correct and one with a single repeat expansion. With such a low sequencing coverage, the allelotyping algorithm was not able to identify the noisy read and assigned a heterozygous state to the locus.

STR locus	lobSTR (bp)	Converted lobSTR	Capillary	Hg18	Repeat	Coverage	Result[a]
D8S1179	-8/8	11/15	11/15	13	[TCTA]n	13	Y
CSF1PO	-12/-4	10/12	10/12	13	[AGAT]n	13	Y
TPOX	0/12	8/11	8/11	8	[AATG]n	12	Y
THO1	11/11	9.3/9.3	9.3	7	[AATG]n	11	Y
D16S539	4/12	12/14	12/14	11	[GATA]n	5	Y
D7S820	-20/-8	8/11	8/11	13	[GATA]n	3	Y
Penta D	-20/0	9/13	9/13	13	[AAAGA]n	3	Y
D5S818	0/4	11/12	11	11	[AGAT]n	3	E
D3S1358	-4/-4	15/15	15/17	16	[TCTN]n	2	P
PentaE	25/25	10/10	10/15	5	[AAAGA]n	1	P
FGA	-4/-4	21/21	21/24	22	[TTTC]n	1	P
D18S51	-12/-12	15/15	15	18	[AGAA]n	1	Y
D13S317	4/4	12/12	11/12	11	[TATC]n	1	P

Table 2.3: **Capillary platform results versus lobSTR results for the CODIS set.** [a]Y - both platforms agree. P - lobSTR reported only one allele out of two. E - lobSTR reported an allele that does not exist.

Next, we evaluated lobSTR calls made in 12 low-pass sequenced genomes from the Human Genome Diversity Project (HGDP) [**?**, **?**]. Five genomes had coverage of 1.4x-1.9x with 109bp reads, and the other seven had coverage of 4.8x-7.7x with 77bp reads (**Supplemental Table 2.9.3**). One hundred and ninety five STRs with equivalent entries in the lobSTR reference have

been genotyped in these genomes using DNA electrophoresis as part of the CEPH-HGDP panel [?, ?]. Combining lobSTR results from all datasets gave 59 comparable markers with coverage of 3-5 reads with a median coverage of 3x (**Supplemental Table 2.9.4**). Despite the low coverage, lobSTR correctly returned 75% of the genotypes and 85% of the allele calls. Most of the alleles showed at least 5bp difference from the reference and some alleles showed a difference of 24bp and were correctly called. We did not observe a significant correlation between errors and the size of the variation.

2.2.5 Genome-wide STR profiling confirms previously locus-centric observations

Encouraged by the accuracy and speed of lobSTR, we harnessed our pipeline to establish a reliable reference for future studies. Our input dataset was a male individual that, as of today, has been sequenced to highest coverage of 126-fold from a blood sample [?]. Fourteen billion sequencing reads were obtained from 100bp PE runs on Illumina GAIIx and HiSeq 2000. lobSTR ran for 26 hours using 25 CPUs. It aligned ∼6 million reads to approximately 180,000 STR loci out of the 249,000 in the Tandem Repeat Table reference with average coverage 20.82 for autosomal loci. The average reference allele length of undetected loci was 150bp, whereas the mean reference length of detected loci was 41bp. Therefore, in most cases, the undetected loci could not physically be spanned by a single read of the current sequencing length.

We assigned each autosomal STR to one of four allelotype categories: both alleles match the reference (homozygous reference), one allele matches the reference (heterozygous reference), both alleles do not match the reference but are the same (homozygous non-reference), and both alleles are different and do not match the reference (heterozygous non-reference). In all previous experiments, a coverage threshold of 20x resulted with near-perfect STR calling even for dinucleotide loci. To increase the reliability of our results, we focus the analysis on the 97,844 loci that were called with at least this sequencing coverage. The length distribution of these alleles in the reference was mainly between 25-50bp with a low number of very long STRs (**Figure 2.2.5A**).

Similar to the other genomes in this study, 55% (52,338) of the STR loci differed from the reference: 22,271 (23%) loci were heterozygous reference, 15,515 (16%) loci were homozygous non-reference, and 14,552 (15%) loci were heterozygous non-reference. The other 43,335 (45%) loci were homozygous reference. Some of the variations reached to 49bp difference from the

reference allele. On average, STR variations showed 6.3bp difference from the reference allele and 41% of the variations were more than 5bp away from the reference (**Figure 2.2.5B**). Thus, mainstream-dependent analysis pipelines that can tolerate only a few nucleotide indels, such as BWA, are likely to miss most STR variations.

The genome-wide STR dynamics reported by lobSTR confirm previous findings of locus-centric studies. The rate of STR polymorphism showed a striking correlation with the repeat unit length (**Figure 2.2.5C**). Dinucleotide STRs are nearly equally likely to fall into any of the above four categories, whereas hexanucleotide STRs are most likely to match the reference. This trend matches results of a previous study that measured the mutation rate of a few hundreds of Y-STR loci as a function of repeat unit length [**?**]. Similar to our results, the authors showed that penta- and hexanucleotide repeats mutate at half the rate of tri- and tetra-nucleotide repeats. We also found that the rate of STR polymorphism is significantly correlated to the length of the STR allele in the reference (**Figure 2.2.5D**). The non-reference loci (n=52,338) had significantly greater lengths than loci that are homozygous reference (n=43,335) ($p<0.05$, one-sided Mann-Whitney test for each allelotype category versus reference) as previously reported in studies that analyzed a few dozen STRs [**?**, **?**].

We also used lobSTR to determine genome-wide trends of STRs at single base pair resolution (**Figure 2.2.5E**). Overall, 99% of alleles varying from the reference allele showed differences that were complete multiples of the STR unit. This trend varied by period, with dinucleotide STRs least likely (0.3%) to differ by an incomplete motif unit and hexanucleotide most likely (4.7%).

Finally, lobSTR reported significant differences between repeat variations in intronic and exonic regions (**Figure 2.2.5F**). Intronic trinucleotide STRs were twice as likely to exhibit at least one non-reference allele than exonic regions (0.480-0.502 95% CI and 0.179-0.336 95% CI for introns and exons, respectively), and nearly five times as likely to exhibit two non-reference alleles (0.107-0.119 95% CI and 0-0.047 95% CI for introns and exons, respectively). Significantly, lobSTR reported that 1.9% (62 out of 3276) of the intronic trinucleotide STRs showed length differences that were not a multiple of three nucleotides. On the other hand, all reported exonic trinucleotide variants retained the reading frame. In addition, lobSTR allelotyped 34,667 intronic and 7 exonic non-trinucleotide STRs. Of the intronic non-trinucleotide STRs, 18,277 (53%) showed at least one allele with a frameshift deviation, and 8,686 (25%) showed two frameshifted alleles. Surprisingly, 3 of the 7 exonic loci, all tetranucleotides, showed expansions by units of 4bp, which would result in a frameshift mutation. In one case, in exon 8 of DCHS2,

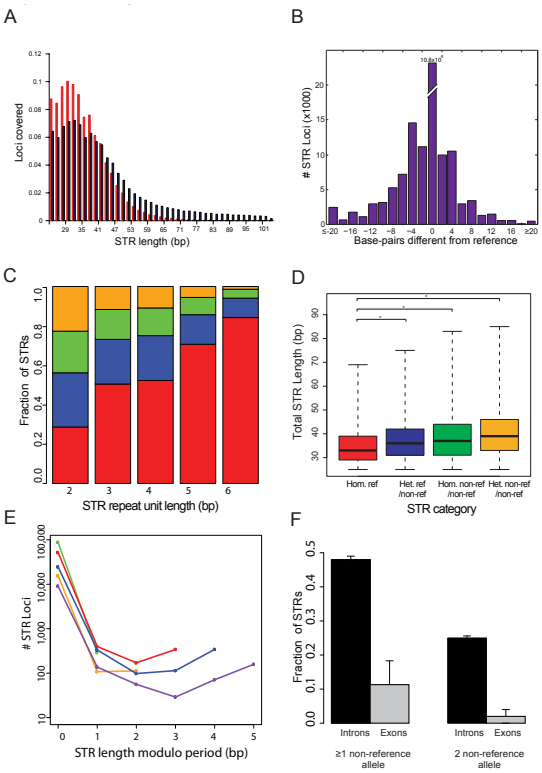


Figure 2-5: **Genome-wide STR profile of an individual. (A) Distribution of STRs with 20x coverage or more as a function of the allele size in hg18. (B) Distribution of allele size differences from reference in lobSTR calls.** The average difference was 6.3bp away from the reference. **(C) STR polymorphism as function of period.** The number of STR alleles matching the reference sequence increases with increasing repeat unit length (the colors in the entire panel are as follows: red - homozygous reference, blue - heterozygous non-reference/reference, green - homozygous non-reference/non-reference, orange - heterozygous non-reference/non-reference **(D) Longer STR regions are more polymorphic.** The median STR length (thick black line) increases with the number of variant alleles. * denotes a significant ($p<0.05$) difference according to a one-sided MannâĂŞWhitney test. Boxes denote the interquartile range and whiskers denote 3 times the interquartile range. **(E) lobSTR shows mutational trends at single base pair resolution.** The number of base pairs difference from reference modulo the period size versus the number of alleles detected (in logarithmic scale) is shown for each period (green - period 2, orange - period 3, red - period 4, blue - period 5, purple - period 6). Incomplete STR unit differences tend to differ by a full unit +/- 1bp from the reference. **(F) Fraction of trinucleotide STRs with non-reference alleles in introns versus exons.** 95% confidence intervals are given by the error bars.

the frameshift variation was homozygous. This call was supported by 33 independent reads, showing a potential loss of function in this gene.

Taken together, the overall findings of lobSTR in this genome serve as a biological validation for the accuracy and utility of genome-wide STR profiling using our technique.

2.3 Discussion

STR profiling techniques have changed very little in the past two decades, relying on the faithful yet cumbersome capillary electrophoresis technique to scan a few dozen loci at a time. The advent of HTS has ushered in the opportunity to conduct genome-wide STR variation analyses. Here, we presented an end-to-end solution for this task. Our solution bypasses the gapped alignment problem, has no inherent indel limitation, and can reliably profile highly polymorphic STRs at a single base pair resolution. We provided a detailed comparison between lobSTR and popular mainstream aligners and showed that even with long reads, these aligners are significantly biased towards the detection of the reference allele. We have established the feasibility of lobSTR to profile STR loci from a total of 20 genomic datasets and demonstrated the strategy's accuracy by analyzing its consistency, ability to trace Mendelian inheritance, and comparing its results to orthogonal molecular techniques. Moreover, our genome-wide STR analysis confirms previous biological observations, which further highlights the algorithmic validity.

lobSTR results from the trio genomes and the Ajay et al. genome consistently showed genome-wide polymorphism rates of 55%-57% for STRs with lengths 25bp and over. A recent study by McIver et al. [**?**] evaluated the performance of STR calling using post-BWA alignment files with a set of quality rules. Using a mixture of Illumina 45-100bp reads and 454 reads from two trios in the 1000Genoems project, they reported that 1.1% of the STRs with lengths of 20bp and over were polymorphic. We wondered if the polymorphism discrepancy between the studies could be explained by the shorter reading lengths in the McIver study that biased their calls to very short, less polymorphic STRs. However, when we ran lobSTR on the 1000Genomes CEU trio datasets (Methods), we found again that 57% of the STRs were polymorphic (25,885 out of 45,461 STRs that were called with âĽě5x coverage at the three genomes). These results suggest that STR profiling that is restricted by the default BWA indel tolerance – 5bp for the Illumina datasets in the McIver et al. study – can significantly reduce the sensitivity for observing STR variations.

We envision that lobSTR will be used in parallel to conventional analysis pipelines in order to augment variation calling to STR loci. The fast running time of our algorithm should not impose a significant computational burden on users. A low coverage genome of 5x takes about an hour on a standard server with 25 CPUs, a high coverage genome of 30x takes eight hours using the same settings, and a ultra-high covered genome of 126x takes 26 hours (**Supplemental Table 2.9.2**).

Currently, the major barrier for STR profiling is the sequencing read length, as the number of detectable STRs is limited to those that are entirely spanned by a single read. To test the effect of genomic coverage on STR profiling, we sampled reads from the 126x genome and calculated the amount of reported STRs (**Supplemental Figure 2.8.6**). With genome-wide coverage of 40x, there are more than 100,000 STRs that will be pass an STR-coverage threshold of 10x. However, higher genomic coverage does not linearly improve the number of STRs that pass this threshold, marking a potential upper bound of sequencing read lengths of 100bp. We also explored the utility of the longer reads by Sanger, 454, and IonTorrent for STR profiling of personal genomes using lobSTR (**Supplemental Table 2.9.5** and **Supplemental Text 2.6**). Longer reads indeed increased the number of reported STR loci compared to the same autosomal coverage by Illumina. However, out of these, Sanger seemed to be the only method to produce reliable STR reads. We expect that as sequencing reads continue to increase in both length and quality, lobSTR's performance will further improve and allow inclusion of a larger number of STR variations. Ultimately, these will include large pathogenic expansion, such as those in Huntington's disease, which can span more than 100bp.

As of today, sequence analysis algorithms can detect almost any type of genetic variations, from SNPs [**?**] and indels [**?**, **?**] to CNVs and chromosomal translocations [**?**]. lobSTR adds a new layer of information with tens of thousands of highly polymorphic genetic variations that have a multitude of applications, from personal genomics, to population studies, forensics analysis, and cancer genome profiling.

2.4 Methods

2.4.1 Comparing lobSTR to mainstream aligners

All alignment strategies were tested in a Linux environment, on a server with 4 twelve-core AMD Opteron 6100 and 128Gbyte of RAM. The following software versions were tested: BWA version 0.5.7, Bowtie version 0.12.7, and Novoalign freeware version 2.7.13, BLAT version 34, and GATK version 1.3-21.

The input was 5 million Illumina reads of the male sample from our lab collection. BLAT results were filtered to include only the top hit for each read. We suppressed multi-mappers in all other tools. Informative STR reads were identified by the intersectBed tool of the Bedtools packages [**?**]. We converted CIGAR scores to the number of base pairs difference from the reference allele by counting any insertions or deletions falling within and directly adjacent to the STR region. Simulating reads from pathogenic STR loci was conducted using a simple Python script available by request from the authors.

2.4.2 Determining the expected number of non-reference reads

A previous study by The Utah Marker Development Group has shown that 70% of thousands of randomly chosen tetranucleotide STR loci are polymorphic. We also re-analyzed Payseur et al. data to infer the polymorphism rate in STRs with length âĽě25bp in the assembled genome of Craig Venter using results reported in their Supplementary Tables 1-5. Concordant with the Utah study, this rate was 66%.

The rate of non-reference STR reads is bounded between two extreme cases. The lower bound is that all polymorphic STRs are heterozygous with a reference allele. Thus, only half of the reads from variable loci will show a non-reference allele, which gives 33% as a lower bound. The upper bound is that all polymorphic STRs are different from the reference in their two alleles. In this case, every read from a variable locus will show a non-reference allele, which gives 66% as an upper bound.

2.4.3 Biological replicates analysis

Raw reads for blood-derived and saliva-derived genomic DNA from the same individual were downloaded from the NCBI short read archive (www.ncbi.nlm.nih.gov/sra) with accessions SRX097307 and SRX097312, respectively. Loci in which (1) less than 75% of reads agreed with the allelotype call in both samples or (2) the locus was covered in either sample by more than three times the mean coverage level were removed from the analysis.

2.4.4 CEU trio data for Mendelian inheritance

The HapMap CEU trio were NA12877 (father), NA12878 (mother), and NA12882 (son). Raw reads were downloaded from the European short read archive (www.ebi.ac.uk/ena/) with accessions ERP001228, ERP001229, and ERP001230, respectively. To determine if an STR followed Mendelian inheritance, we required that the alleles detected in the son could be explained by inheriting one allele from each parent. Low quality loci were filtered as described in Biological replicates analysis.

2.4.5 Validating lobSTR accuracy using capillary electrophoresis

Four Catch-All buccal swabs (Epicenter, QEC89100) were used to collect the DNA sample according to the manufacturer protocol. gDNA was extracted by QuickExtract (Epicentre), followed by phenol-chloroform purification and ethanol precipitation. Library preparation was performed according to the standard Illumina protocol. Three runs of 101bp paired-end with a GAIIx platform were used for sequencing. The study was approved by MIT's Committee on the Use of Humans as Experimental Subjects (COUHES). The general sequencing coverage was analyzed as previously reported [**?**].

Capillary electrophoresis results were obtained from Sorenson Genomics laboratory using the commercial Promega PowerPlex 16 system. To find the genomic positions of these loci, we downloaded corresponding primers that target these loci from the Short Tandem Repeats Internet Database (STRBASE) website (http://www.cstl.nist.gov/strbase/) of the US National Institute of Standards and Technology (NIST) and used the In Silico PCR tool on the UCSC genome browser to reveal their location. Two loci had proprietary primers and their genomic location could not be identified. The STR repeats in the sequencing file were converted to the

PowerPlex allele nomenclature using the NIST definitions.

2.4.6 Obtaining CEPH-HGDP STR allelotypes

STR allelotypes along with a table of RefSeq reference alleles were downloaded from the Rosenberg lab site (www.stanford.edu/group/rosenberglab/repeatsDownload.html). The allelotypes were given as the number of repeats converted from PCR product size as described in Pemberton et al. [?]. The repeat number is given as reference repeat number plus the difference in product size from the reference divided by the motif size. Sequence data were downloaded from the NCBI Short Read Archive with accessions: ERX004003, ERX004002, ERX004001, ERX004000, ERX0039999, ERX004007, ERX007978, ERX007977, ERX007976, ERX007975, ERX007974, ERX007973, ERX007972.

Using the STS marker table available from the UCSC Genome Browser, we converted the Pemberton et al. markers to hg18 genomic coordinates and annotated them using the TRF table. lobSTR calls that are supported by three or more reads were converted to the Pemberton results. Non-integer repeats reported by lobSTR were rounded to the smallest integer for compatibility with Pemberton data. Markers that could not be faithfully annotated were removed from the analysis.

2.4.7 Genome-wide STR profiling of a deeply sequenced personal genome

Raw sequencing reads for accession number ERP000765 were downloaded from the European Nucleotide Archive's Sequence Read Archive (http://www.ebi.ac.uk/ena/). The Mann-Whitney test was performed using the wilcox.test function in R. Confidence intervals were calculated using a normal approximation to the Poisson distribution, with a 95% confidence interval of $\lambda \pm 1.96\sqrt{\lambda}$, where λ is the estimated mean of the distribution. Only loci with greater than 20-fold coverage were included in the analysis. Exon and intron coordinates were obtained from the UCSC table browser for human genome build hg18.

2.4.8 1000 Genomes data analysis for the McIver Study

The HapMap CEU trio were NA12878 (daughter), NA12891 (father), and NA12892 (mother). Raw sequencing reads for the CEU HapMap trios with length of at least 47bp were downloaded

from the 1000 genomes NCBI ftp site (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/). 228, 274, and 214 files were included for individuals NA12878, NA12891, and NA12892. To accommodate the shorter read lengths, lobSTR was run with non-default parameters âĂŞfft-window-size 20 âĂŞfft-window-step 10, –maxflank 100, and âĂŞextend-flank 5.

2.5 Acknowledgements

Y.E is an Andria and Paul Heafy Family Fellow. This publication was supported by the National Defense Science and Engineering Graduate Fellowship (M.G.) and by a fellowship from the Edmond J. Safra Bioinformatics program at Tel-Aviv University (D.G.). D.G. and S.R. acknowledge support from Israeli Science Foundation grant ISF 1227/09 and an IBM Open Collaborative Research grant. We thank Mona Sheikh, Dina Esposito, and Alon Goren for useful comments on the manuscript, Assaf Gordon for his assistance with multithreading programming, Cole Trapnell for his assistance with preparing lobSTR executables, Mona Sheikh and Sam Sinai for testing lobSTR code, and Dina Esposito for preparing samples for genotyping.

2.6 Supplemental Text

2.6.1 lobSTR algorithm

Sensing

The aim of the sensing step is to find informative reads and characterize their STR sequence. The first task of the algorithm is to detect whether a read contains a repetitive sequence. The algorithm breaks the sequence read into overlapping windows with length of w nucleotides and r nucleotide overlap between consecutive windows. In practice, we use $w = 24$ and $r = 12$. Then, it measures the sequence entropy of each window, according to:

$$E(S_j) = -\sum_{i \in \Sigma} f_i \log_2 f_i \tag{2.1}$$

where E is the entropy, S_j is the sequence of the j-th window, Σ is the alphabet set, i is a symbol in the alphabet, and f_i is the frequency of symbol i. A fully random sequence results

in the maximal entropy score that equals to $\log_2(|\Sigma|)$, whereas a repetitive sequence overuses a few symbols and results in a low entropy score, ideally zero in the case of a perfect homopolymer run.

The entropy score proved extremely powerful in discriminating STR sequences from other genomic sequences (**Supplemental Figure 2.8.2A**). We calculated the entropy score of sliding windows of 24bp from all documented human STR sequences of repeat unit length of 2-6bp that span up to 100bp. In parallel we scored one million randomly sampled human genomic sequences of 24bp. Then, we classified the input sequences according to their entropy score. The area under the receiver-operating curve (ROC) was 98.3% when the entropy measurement used the four nucleotides as the input alphabet. We further boosted the classification performance by calculating the entropy using dinucleotide symbols, meaning that âĂIJAAâĂİ maps to one symbol, âĂIJACâĂİ maps to a different symbol, and so forth. In this case, the area under the ROC climbed to 99.4%, which renders it a nearly perfect classifier. Accordingly, lobSTR uses the dinucleotide symbols for the entropy measure, and we empirically found that an entropy threshold of 2.2 bits provides the optimal performance in terms of speed and number of aligned STR reads.

lobSTR uses the pattern of entropy scores to identify informative reads that fully encompass STR regions. These reads display a series of windows with entropy score below-threshold (the STR region) that are flanked by one or more windows with entropy score above-threshold (the non-repetitive regions) (**Supplemental Figure 2.8.2B**). The algorithm only retains reads that follow this pattern. Approximately 97% of whole genome sequence reads are excluded by this rapid procedure. This significantly contributes to the algorithm's speed, since only a few simple entropy calculations are required to identify the informative STR reads in massive sequencing datasets.

The next task of the sensing step is to determine the length of the repeat unit. Most STR loci do not contain a perfect series of the same repeat unit [**?**]. We took a spectral analysis approach that quickly integrates information over the entire STR region to reliably identify the repeat consensus even in imperfect repeats [**?**]. Starting from the window with the lowest entropy score, consecutive windows scoring below the threshold are merged. The sequence of the merged repetitive region is represented as M, an $n \times 4$ binary matrix, where n is the number of nucleotides in the repetitive region, the i-th row of the matrix corresponds to the i-th position of the sequence, and each column corresponds to a different nucleotide type (A,C,G,T). For instance, the DNA sequence ACCGT is represented as:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2}$$

The power spectrum of the STR matrix is calculated by performing a Fast Fourier Transform along the columns of the matrix:

$$S(f) = \sum_{y=1}^{4} \left(\sum_{x=1}^{n} M_{x,y} \cdot e^{-\frac{2\pi i x f}{n}} \right)^2 \tag{2.3}$$

Where $M_{x,y}$ is the element in the x-th row and y-th column of M.

STRs have a unique fingerprint in the frequency domain [**?**, **?**]. Similar to repetitive signals in the time domain, the spectral response of STR elements is characterized by harmonics - a strong signal in recurrent frequency bins and a weak signal in other bins (**Supplemental Figure 2.8.2C**). The peaks of the harmonics for a repeat unit of length k dwell in the $\frac{ni}{k}$ mod n bins, $i = \{0, \pm 1, \pm 2, \ldots\}$. For instance, in a case of n=24, a dinucleotide STR generates a strong signal in bins 0 and $\pm$12. A trinucleotide STR generates a strong signal in bins 0, $\pm$8, and $\pm$16. The algorithm integrates over the normalized energy of the first harmony (i.e. i=1) of each possible repeat unit between 2 to 6bp. The consensus repeat unit length is selected according to the highest energy of the corresponding frequency bin (**Supplemental Figure 2.8.2D**). Some STR regions may show strong signals in more than one energy bin (i.e., repeats of period 4 show strong energy in both the second and fourth harmonics, and repeats with several insertions or deletions may have more than one strong harmonic). If the second highest harmonic has energy within 30% of the highest, lobSTR will attempt alignment using the second best period choice if alignment using the first choice fails.

Finally, the algorithm determines the actual STR sequence. It uses a rolling hash function to record all possible k-mers in the STR region, where k is the reported repeat unit length described above. The most frequently occurring k-mer is set to be the repeat unit of the STR. The output of the sensing step is (a) the consensus sequence of the STR's repeat unit in the canonical form (see Supplemental Methods for the canonical form definition) (b) the sequence read divided into three regions: the STR region, and upstream and downstream flanking regions that correspond

to the location of the above threshold windows.

Alignment

The aim of the alignment step is to reveal the identity and the repeat length of an STR-containing read. We do not attempt to align the entire sequence read to the genome to avoid time-prohibitive gapped alignment. Instead, lobSTR employs a divide-and-conquer approach. It separately anchors the upstream and downstream flanking regions of STR-containing sequence reads, without mapping the STR region itself. This procedure identifies the genomic location of the STR and reveals the repeat length by measuring the distance between the flanking regions.

A major challenge of the divide-and-conquer approach is to specifically map the short flanking regions to the genome. To circumvent this problem, we restrict the alignment to STR loci with the same repeat sequence that was reported by the sensing step. We built a reference that holds the flanking regions of all the 240,351 STR loci with repeat unit 2-6bp in the human genome according to the Tandem Repeat Finder table [**?**]. The flanking strings are compressed using a Burrows Wheeler Transform [**?**] (BWT) to allow efficient searching. All flanking strings of STRs with the same repeat structure are organized under the same BWT structure (Supplemental Methods). Thus, lobSTR only searches a single BWT data structure that corresponds to the repeat sequence, which typically holds up to a few thousand loci. Then, lobSTR intersects the potential mapping positions of the upstream and downstream regions to identify a single compatible location and excludes multiple mappers. This procedure not only speeds up the alignment, but enables higher rates of unique mapping even when the flanking regions are only a few nucleotides long.

To determine the length of the STR in the read, the algorithm uses the following equation:

$$L = s - (d - u) + L_{ref} \tag{2.4}$$

Where L is the observed STR length, s is the length of the sequence read, d is the genomic coordinate of the last nucleotide in the downstream region, u is the genomic coordinate of the first nucleotide of the upstream region, and L_{ref} is the length of the STR region in the reference genome.

One important aspect of Eq. 4 is that inaccuracies in the sensing step regarding the exact

boundaries of the STR do not affect the reported length of the STR. However, insertions or deletions in the flanking regions might be reported as STR differences, although they are not actual differences in the STR region itself. To mitigate this issue, lobSTR performs local realignment of the entire read once a match is found using the [?]. Indels that are detected in the flanking regions are not taken into account and removed from Eq. 2.4, providing accurate length calling of the STR region. In addition, the local realignment is used to produce a CIGAR string with the locations of the indels in the read. Downstream genotype callers can use the output of lobSTR to call SNPs and indels in the STR region and its flanking regions.

The output of the alignment step is the genomic coordinates of the aligned read, the strand, the STR region extracted from the read, the STR motif, the nucleotide length difference compared to the reference genome as described above, the CIGAR string, and the realignment score. We report the alignments in a custom tab-delimited format, as well as in the BAM format [?] to ensure compatibility with other downstream bioinformatics tools.

Allelotyping

The aim of the allelotyping step is to determine the most likely alleles of each STR locus by integrating information from all aligned reads and the expected stutter noise, which is created due to in vitro slippage events during sample preparation. This part of the program uses a BAM file as input and reports the allelotype calls.

By analyzing real sequencing data, we found that the length of the repeat unit is a major determinant of the stutter noise distribution (Supplemental Methods). In accordance with the mutation dynamics of STRs previously reported [?], short repeat units are associated with higher stutter noise and long repeat units are more immune to noise (**Supplemental Figure 2.8.3A**). We did not find a significant association between stutter noise and the length of the STR (**Supplemental Figure 2.8.3B**) as was observed in past studies [?].

We developed a generative model for stutter noise that consists of two steps: (a) with probability $\pi(k)$, the read is a product of stutter noise, where k denotes the repeat unit length (b) if the read is a product of stutter noise, then with probability $\mu(s;\lambda_k)$, the noisy read deviates by s base pairs from the original allele, where $\mu(s;\lambda_k)$ is a Poisson distribution with parameter λ_k. The probabilities that the deviation is positive (repeat expansion) or negative (repeat contraction) are equal.

The user has two options to estimate the model parameters $\pi(k)$ and λ_k. In the case of a male genome, the user can instruct lobSTR to scan the hemizygous sex chromosomes to accumulate unambiguous data about stutter noise distribution. The algorithm observes the stutter probability for each repeat unit length and uses a logistic regression to infer $\pi(k)$ (**Supplemental Figure 2.8.3A**) and a Poission regression to learn λ_k. In the case of a female genome, users can use pre-computed values either from our observations or analyze male data in their collection.

Overall, the probability of generating a read with L bp in the STR region from a hemizygous locus with an STR with A bp in the STR region is:

$$P(L|A,k) = \begin{array}{cc} 1-\pi(k) & \text{if } L = A \\ \frac{\pi k}{2}\mu(|A-L|-1,\lambda_k) & \text{otherwise} \end{array} \tag{2.5}$$

In a diploid STR locus with A and B repeat lengths, we use the following heuristic to approximate the likelihood of observing a read with length L:

$$P(L|A,B,k) = max(P(L|A,k), P(L|B,k)) \tag{2.6}$$

This heuristic was found to be more robust when the two STR alleles have large length differences.

Let $\vec{R}$ be a vector that describes the STR lengths of sequence reads from the same locus after removing PCR duplicates (Supplemental Methods). Since each remaining sequence read is a product of an independent series of PCR rounds, we assume that the stutter noise of different entries in $\vec{R}$ is independent. Accordingly:

$$\log[P(\vec{R}|A,B,k) = \sum_{L \in \vec{R}} \log[P(L|A,B,k)]] \tag{2.7}$$

Thus, the most likely allelotype call is when Eq. (7) is maximized with respect to A and B. To find the best bi-allelic combination, we simply iterate over all possible pairs of STR lengths observed at the interrogated locus and compute the likelihood of generating the observed data given the noise model. For example, if $\vec{R} = (13, 13, 12, 12, 12)$, we calculate the log likelihood in Eq. 2.7 for the combinations: (A=12,B=12), (A=12,B=13), and (A=13,B=13). In addition to the log likelihood score, we require a minimum threshold of the variant allele in order to

call a locus as heterozygous, with a default threshold of 20% and a minimum percentage of reads supporting the resulting allelotype, which defaults to 50%. In the case of sex chromosome loci for a male sample, only homozygous allelotypes are considered. The most likely (A, B) combination is reported.

For each STR locus, the allelotyping step returns the chromosome, start, and end of the locus, the STR motif and period, the reference repeat number from TRF, the allelotype call given as the number of base pairs difference from reference for each allele, coverage, number of reads agreeing with the allelotype call, the number of reads disagreeing with the allelotype call, and the number of reads supporting each observed allele.

2.6.2 Technical Evaluation of lobSTR

Comparison between lobSTR sensing step and the TRF tool

Tandem Repeat Finder was developed to find repetitive elements in large sequence contigs. Conceptually, it could also process short reads and replace the lobSTR sensing step in characterizing STR repeats. To compare the performance of the two lobSTR sensing step and TRF, we challenged the two tools with a set of 5 million 101bp whole-genome Illumina reads. To make a fair comparison, TRF was restricted to a maximum repeat unit period of six nucleotides and lobSTR ran on a single CPU.

Our results indicate that lobSTR's sensing step is significantly more adequate for high throughput sequencing data. lobSTR running time was just under 8 minutes compared to 6.5 hours for TRF (about 50 times slower, **Supplemental Figure 2.8.4A**). This means that analyzing personal genomes would take weeks instead of half a day of running time. Moreover, 94% percent of reads that were flagged as informative by both methods were reported with the same repeat sequence (**Supplemental Figure 2.8.4B**). Most of the discordant results occurred in STRs of period 5 or 6 where lobSTR and TRF could not reach a consensus regarding the repeat unit of imperfect repeats. Last, lobSTR flagged as informative 75%-85% of the reads that were flagged by TRF, with higher sensitivity with increasing STR purity (**Supplemental Figure 2.8.4C**). Thus, while lobSTR cannot detect every read that is detected by TRF, it does reach high sensitivity with 1/50 of the running time which is more suited to the ultra-exponential trajectory of high throughput sequencing datasets.

lobSTR performance with different sizes of STRs

The size of the flanking regions determines the mappability of STR containing reads. In order to find the minimal flanking regions, we extracted genomic sequences of 100bp upstream and downstream of all STRs in the TRF table and organized them in prefix trees according to their canonical repeat unit. Then, we exhaustively aligned target STRs by allowing increased flanking region lengths and reporting the minimal length when a unique and correct alignment was achieved. Since this step is time prohibitive, we focused our analysis on a set of 2050 STR from the CODIS set, exonic regions, and genealogical Y-STR markers that were covered by 100bp reads. Our results show that a total of 8-9bp of upstream and downstream flanking regions is a lower bound for unique alignment of 80% of tested STRs (**Supplemental Figure 2.8.5A**). This means that with 100bp reads, lobSTR can theoretically detect STR regions of up to 84nt.

We also determined the power of lobSTR to detect reads with very short STR regions due to strong repeat contraction. These reads have higher entropy and might not cross the threshold in the sensing step. To simulate this effect, we ran TRF on a set of 5 million input reads in a setting returning detected STRs as few as 12bp long. We then measured the performance of lobSTR to detect reads from these short STR loci and found that repetitive elements with 12nt were well captured (**Supplemental Figure 2.8.5B**). Our overall results suggest that lobSTR can perform well in detecting STRs of 12-84bp.

lobSTR performance on various sequenecing platforms

To test lobSTR performance on other sequencing platforms than Illumina, we ran the algorithm on publicly available genomes from three different platforms: Sanger (Craig Venter genome) [**?**], 454 (Watson genome) [**?**], and IonTorrent (Moore genome) [**?**]. In the absence of orthogonal information about STRs in these genomes, we estimated the performance of lobSTR by several parameters: (a) the ratio of aligned STR reads to the total input (b) the fraction of reads with a non-integer number of repeat units different from reference (c) the coverage of STR loci (**Supplemental Table 2.9.5**).

As expected from its long read length and high accuracy, Sanger sequencing showed the best performance. It produced the best ratio of reads that aligned to STR loci and showed the lowest fraction (7.3%) of STR reads with a non integer number of repeat units difference from reference. Importantly, the Sanger fraction of non-integer number of repeats was close to the

Illumina fraction (8.0%). 454 produced more STR aligned reads per amount of sequencing data than Illumina but 25% of the STR reads showed a non-integer number of repeat units. IonTorrent showed the worst performance in both the ratio of STR reads and non-integer repeats. The high number of STR reads with non-integer repeat units is presumably because 454 and Ion Torrent exhibit indel error when sequencing homopolymer runs that are abundant in many types of repeats (e.g. AAAAC).

Our results show that lobSTR can process sequencing files from other high throughput sequencing platforms and report STR reads. However, the accuracy of the STR calls is expected to be inferior to that reported for Illumina. We expect that improvement in homopolymer sequencing in 454 and Ion Torrent will make their datasets more amenable to STR profiling.

STR coverage as a function of input libraries

We sought to explore the function of STR coverage by lobSTR to the genome-wide coverage of autosomal regions. Using the genome sequenced to 126x coverage by Ajay, et al. [**?**] described in the main text, we sampled from the BAM file produced by lobSTR for a range of desired coverage levels. We then allelotyped only this subset of reads and counted how many STR loci were covered by at least one informative read (**Supplemental Figure 2.8.6**).

As a rule of thumb for 100bp reads, we found that STRs obtain an average coverage of approximately one-fifth the genome-wide autosomal coverage. In addition, we found that around 60,000 to 80,000 STRs can be covered by at least a single sequence read even with a shallow genomic coverage of less than 5x. The number of STRs covered by at least 1 read rapidly plateaued to 180,000 loci after a genome-wide coverage of around 40x.

Certain STR loci in the TRF table cannot be detected regardless of the coverage. The main limitation is that 100bp reads cannot span 16% of the STR entries in TRF. Other STR regions dwell in repetitive elements and generate non unique alignments, such as the Y-STR marker DYS464a/b/c/d/e/f, which has multiple locations [**?**]. Reads from these loci will be flagged as multi-mappers and will be removed from the analysis. Finally, some STR regions do not pass the entropy threshold due to their imperfect repeat structure and will not be detected using lobSTR default parameters. This can be circumvented by lowering the lobSTR entropy threshold but will require substantial running time.

Coverage bias in heterozygous loci

We found a slight but statistically significant bias of 1:1.06 in the number of reads towards the shorter alleles in heterozygous loci (one sided Mann-Whitney test, $p < 0.05$). For instance, there are on average $\sim$2 more reads that support the shorter allele when an STR is covered by 30 reads. This observation can be explained by a PCR bias as reported by a previous study [?]. Since this small effect only becomes visible in ultra-high coverage STRs, lobSTR does not currently correct for it.

Hardware requirements

With the given TRF reference, lobSTR reaches a peak memory footprint of 0.3Gb regardless of input size and can process about 0.6 million reads per minute on a single processor. On 25 processors, lobSTR took 26 hours to process the genome sequenced to 126x coverage described in the main text, rendering the hardware requirements of lobSTR well within the range of routinely performed bioinformatics tasks such as SNP calling and short read alignment. The processing times for several genomes analyzed in this paper are given in **Supplemental Table 2.9.2**.

2.7 Supplemental Methods

2.7.1 Building an STR reference

The STR reference was built according to the entries of the Simple Tandem Repeat Table for human reference genome build hg18, available from the UCSC genome browser (this reference was used for all other results as well) [?]. The table was filtered to include STRs with repeat unit lengths of 2-6bp. Nearly half of the loci are dinucleotide repeats. The number of STR loci with each repeat unit length is given in **Supplemental Table 2.9.6**. The 10 most common repeat units are given in **Supplemental Table 2.9.7**. The median length of STR regions is near 40bp for each repeat unit length. The distribution of repeat region sizes increases slightly with the repeat period, and less than 6% of STR loci span more than 100bp (**Supplemental Figure 2.8.7**). The majority of reference STRs lie in intergenic regions. 1,221 reference loci overlap exonic regions.

STRs display cyclic ambiguity. For example, consider the following STR: GACGACGACGAC-GAC. This STR can be described in three ways (GAC)5, (ACG)5, or (CGA)5, as well as by (GTC)5, (CGT)5, or (TCG)5 on the reverse strand. The sequence repeats in the TRF table are reported in a redundant format that does not distinguish between cyclic shifts. We converted all repeat sequences in the table to a canonical form in which the repeat sequence is the lexicographically highest among all possible cyclic representations of the sequence and their reverse complements. STRs whose repeat sequences contradicted the canonical repeat unit length, such as TTT listed as period 3 instead of 1, were removed from the reference. lobSTR reports the period of the STR according to the canonical form.

For mapping Illumina reads, the reference consists of the Âś150bp flanking regions of each STR locus. We grouped reference sequences from loci with the same canonical STR repeat unit into a single FASTA file and built a single BWT index using the BWA function "bwa index âĂŞa is" on each file.

2.7.2 PCR duplicate removal

By default all reads with the same 5' coordinate and length are flagged as PCR duplicates and collapsed into a single read. The user has the option to turn PCR duplicate removal off. If a group of PCR duplicate reads are associated with more than one STR length, lobSTR uses a majority vote to determine the STR length of the collapsed read. If the majority vote results in a tie, the STR length of the collapsed read is determined according to the read with the highest average quality score. All reported sequencing coverage numbers are given after removing PCR duplicates.

2.7.3 Building a model for stutter noise

We analyzed Illumina reads from approximately 6,000 hemizygous loci on the sex chromosomes of a male individual from our lab collection. We assumed that the mode of the STR lengths in each locus was the true allele. All reads differing from the modal allele differed by either one (76% of noisy reads) or two (24% of noisy reads) repeats. Initial analysis of the stutter noise was done using R and was implemented in C++/R in the allelotyping script that is part of the lobSTR package.

2.7.4 lobSTR implementation details

lobSTR is written in C/C++ and calls on R for allelotyping step. We made an effort to use existing, highly optimized libraries for lobSTR implementation to increase the speed of the program. The spectral analysis in the detection step was implemented using FFTW [?] and the alignment step uses extensive parts of the BWA code [?] for BWT-indexing and the BamTools library [?] for reading and writing BAM files.

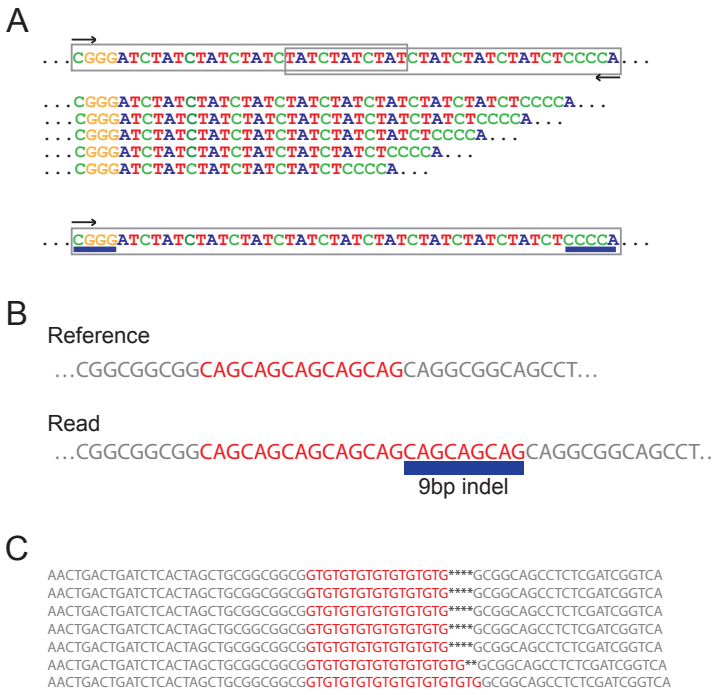
From the user's perspective, lobSTR consists of running two simple programs: one command for sensing and alignment, followed by a command for allelotyping aligned reads from a BAM file. In the simplest setting, the user just needs to specify the input files, the prefix name of the output files, and the location of the reference, which is provided with the software. However, we also provide advanced options that include modification of the detection threshold, re-sizing the FFT windows, and increasing the tolerance to sequencing errors in the flanking regions (**Supplemental Table 2.9.1**). The user can also build a custom reference using a tool in the lobSTR package.

2.7.5 lobSTR comparison across sequencing platforms

Raw reads for the Watson genome were downloaded from the NCBI short read archive with accession SRX000114. Reads for the Moore genome were downloaded from the Europe an short read archive with accession ERS024569. Reads for the Venter genome were downloaded from TraceDB (Genbank accession ABBA00000000). For the Venter genome, we trimmed the first 50bp of every read due to the high error rate at the beginning of Sanger sequence reads and discarded reads whose length after trimming was less than 100bp.

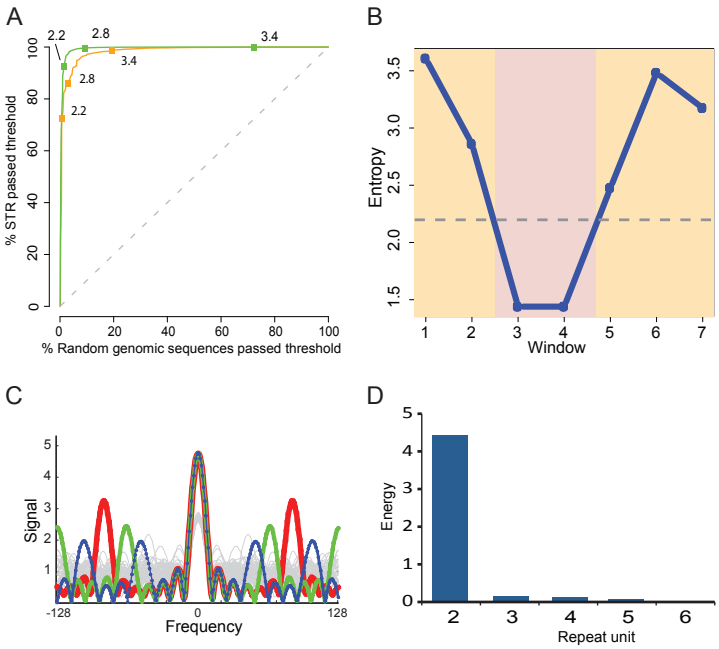
2.8 Supplemental Figures

2.8.1 Supplemental Figure 1



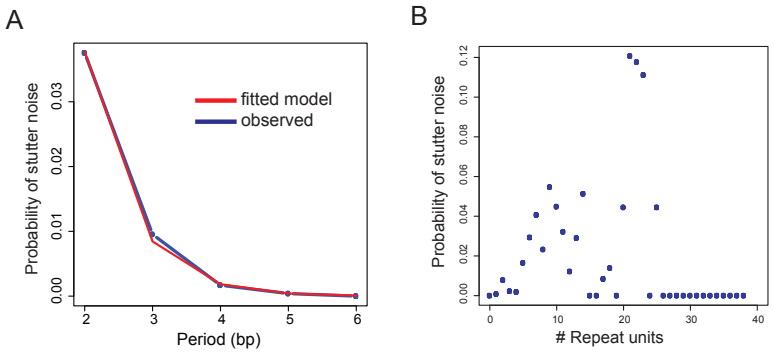
Feasibility of STR profiling with HTS. (A) Long paired-end reads are not a sufficient condition for STR profiling. Each read end (gray box) only partially covers the STR locus (top). Although the ends overlap and together span the entire locus, the number of repeat units is ambiguous (middle), since the exact overlap length is unknown. Single-ends that span the entire STR are informative (bottom). The single-end read (gray box) encompasses the flanking regions outside the STR locus (blue lines). These allow the read to be anchored to obtain unambiguous information about the STR length. **(B) Modest STR polymorphisms translate to large indels.** Thus, detecting non-reference STR reads requires cumbersome processing times by mainstream aligners **(C) An example of stutter noise due to PCR slippage events.** A series of sequence reads were obtained from the same allele. The last two reads have additional repeats due to PCR slippage events. This can confound a naive STR calling strategy to report that the locus is heterozygous.

2.8.2 Supplemental Figure 2



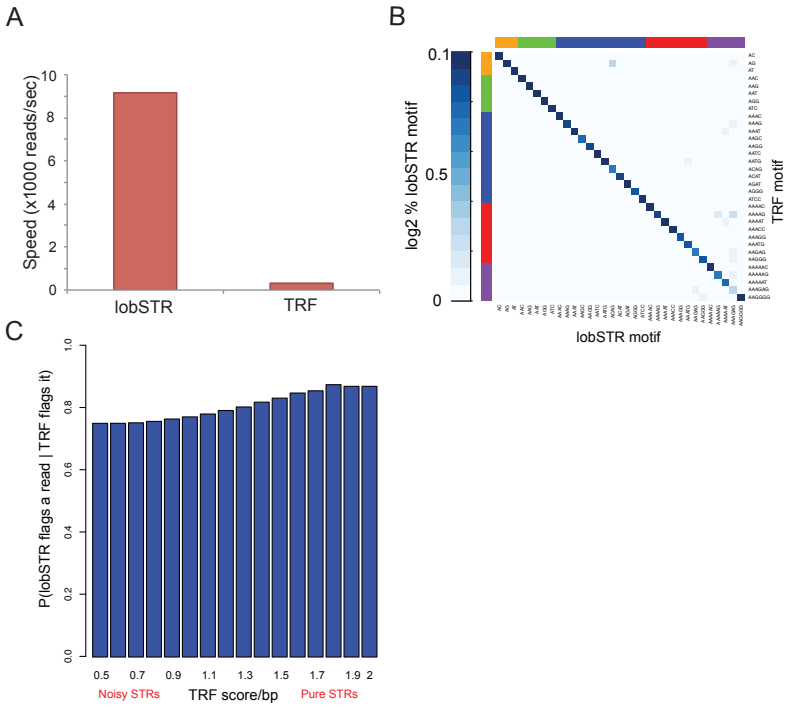
Detection of STR containing reads (A) Entropy scores discriminate between STR sequences and random genomic sequences. The receiver operating curve of the rate of STRs passing the threshold (sensitivity) versus the rate of random genomic sequences passing the threshold (1-specificity). The dinucleotide entropy (green) shows nearly perfect classification and outperforms the mononucleotide entropy (orange). The numbers on the curves denote the corresponding entropy threshold **(B) Informative reads have a unique signature in entropy analysis.** The dinucleotide entropy score is presented for sliding windows along the STR-containing sequence. The flanking regions (yellow) have a high entropy score, whereas the STR-containing region (pink) exhibits a low score. The dashed line depicts lobSTR's default entropy threshold **(C) STR periods create distinct signals in the frequency domain.** The normalized spectral response of STR repeats is characterized by distinct harmonics corresponding to the repeat unit size (blue - 5mer, green - 4mer, red - 3mer, gray - random noise) **(D) Spectral analysis determines the repeat unit length.** The period whose first harmonic shows the maximum energy is chosen as the repeat length. The example displays the normalized energies of periods 2 through 6 of a dinucleotide STR.

2.8.3 Supplemental Figure 3



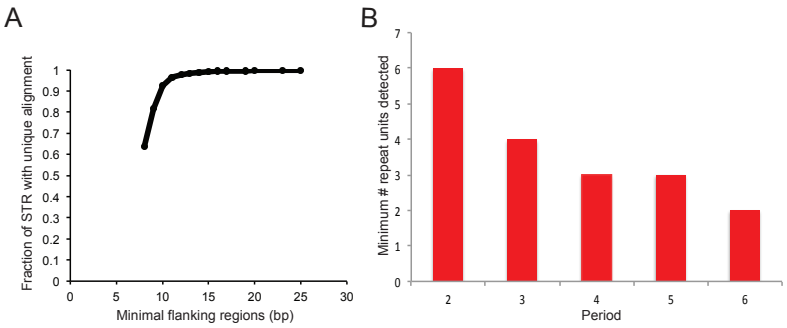
The allelotyping step models the likelihood of PCR stutter noise. (A) Stutter noise as a function of STR period. The probability of PCR stutter noise decreases with period length. Stutter noise (blue) was measured as the percentage of reads from a male sample aligned to sex chromosomes that did not exhibit the mode number of repeat units. We fitted a logistic regression (red) to model noise based on STR period. **(B) Probability of PCR stutter noise as a function of STR number.** There is no strong association between the STR region length and the stutter noise.

2.8.4 Supplemental Figure 4



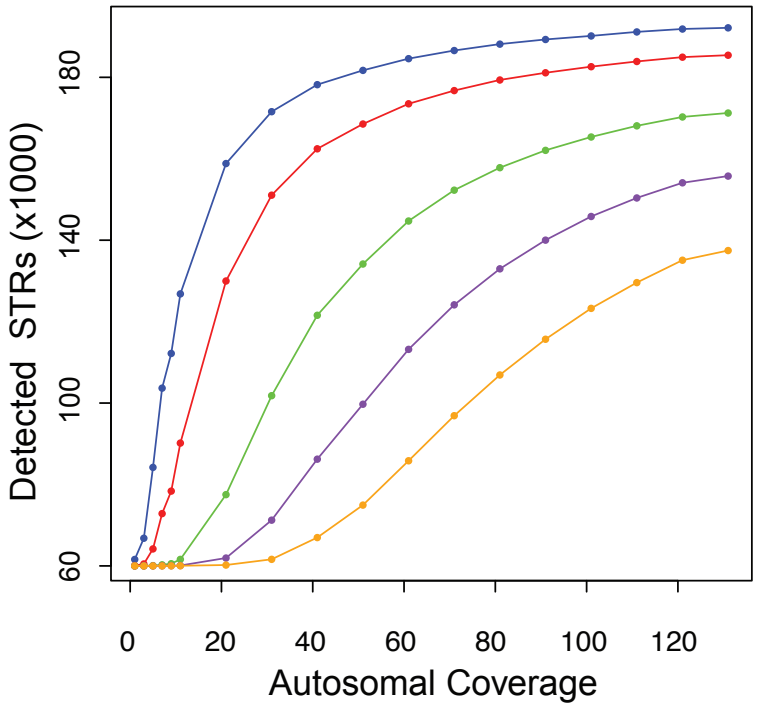
Evaluation of the lobSTR sensing step versus TRF (A) lobSTR senses reads 50 times faster than TRF (B) lobSTR motif detection agrees with Tandem Repeat Finder. In reads where both lobSTR and TRF detect an STR, a comparison of the most represented motifs is shown. Each column is normalized to sum to one, so that values are given in the percentage of instances of a motif detected by lobSTR that were detected as a given motif in TRF. (orange = period 2, green = period 3, blue = period 4, red = period 5, purple = period 6). Overall, lobSTR and TRF agreed in 94% of the calls **(C) lobSTR flagging rate for reads that were flagged by TRF as a function of STR purity.** lobSTR flags about 75% of the reads that were detected from noisy STRs and 85% of the reads from pure STRs.

2.8.5 Supplemental Figure 5



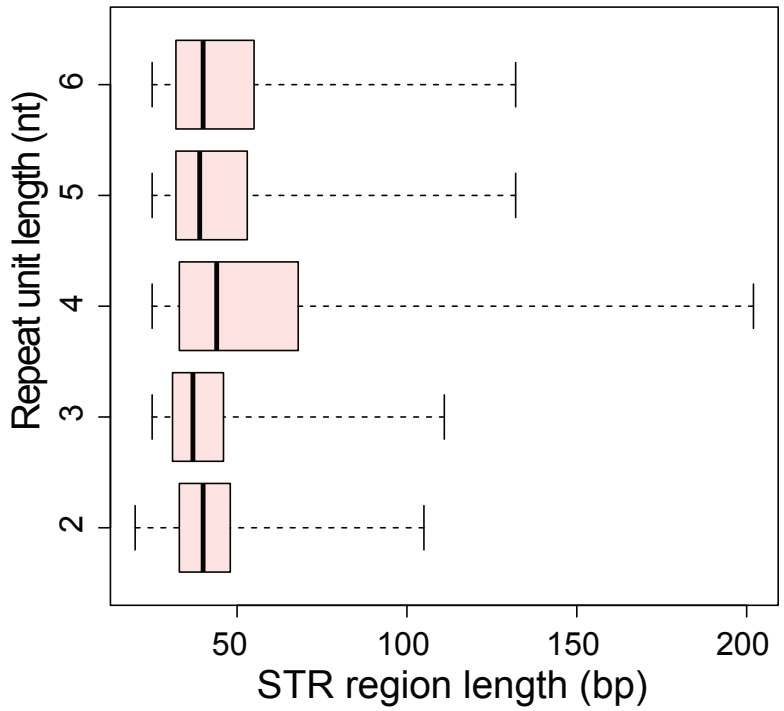
Range of STR lengths that lobSTR can process. lobSTR is able to process STRs with total lengths in the range of 12-84bp from 100bp reads. **(A) Mappability imposes a lower bound on the size of flanking regions for alignment.** We found that above 9bp upstream and downstream flanking regions, nearly 100% of STR regions are uniquely aligned **(B) lobSTR can detect STRs with minimal repeat units.** lobSTR can detect STR regions spanning as few as 12 base pairs.

2.8.6 Supplemental Figure 6



Coverage requirements for STR allelotyping. The number of STR loci at various STR coverage levels (blue - 3x, red - 5x, green - 10x, purple - 15x, orange - 20x) is shown as a function of autosomal genomic coverage. Various coverage levels were simulated by sampling from the aligned reads file of the 126x genome under the assumption that number of aligned STR reads is approximately proportional to genome-wide coverage.

2.8.7 Supplemental Figure 7



Length distribution of STRs. Most periods have a median STR repeat length (thick black line) of around 40bp. The repeat region total length and length variance increase slightly with increasing repeat period.

2.9 Supplemental Tables

2.9.1 Supplemental Table 1

See the lobSTR website usage page http://lobstr.teamerlich.org/usage.html for the most up to date lobSTR usage parameters.

2.9.2 Supplemental Table 2

Genome	Autosomal Coverage	Processing time
HGDP00778	5x	1.3 hours
Male individual	36x	8.5 hours
Ajay, et al.	126x	26 hours

Table 2.4: **Processing times of Illumina genomes at various coverage levels.** Processing times as a result of running lobSTR with 25 processors (-p 25).

2.9.3 Supplemental Table 3

Sample	Coverage	STR Aligned reads	STRs $\geq$1x	STRs $\geq$3x
HGDP00456 (Mbuti Pygmy)	1.4x	70,424	50,505	3,339
HGDP00998 (Karitiana Native American)	1.3x	65,236	48,481	2,553
HGDP00665 (Sardinian)	1.5x	91,157	58,623	6,215
HGDP00491 (Bougainville Melanesian)	1.7x	97,398	61,463	6,523
HGDP00711 (Cambodian)	1.9x	104,594	66,566	7,263
HGDP01224 (Mongolian)	1.7x	93,938	61,356	5,767
HGDP00551 (Papuan)	1.6x	94,540	61,486	6,036
HGDP00521 (French)	5.9x	184,437	91,813	17,855
HGDP01029 (San)	7.7x	192,798	93,376	18,928
HGDP00542 (Papuan)	5.9x	118,232	72,654	7,065
HGDP00927 (Yoruba)	4.8x	155,136	84,828	12,774
HGDP00778 (Han)	5.0x	141,522	80,631	10,815

Table 2.5: **HGDP sample coverage and lobSTR results.**

2.9.4 Supplemental Table 4

Sample	Period	Marker	Refseq (hg18 diff)	Coverage	Converted lobSTR allele[a]	Converted HGDP allele[a]	Status[b]
HGDP01029	4	D11S2371	(TATC)11	5	0,0	0,0	2
HGDP01029	4	D12S1300	(TAGA)12	5	0,0	0,0	2
HGDP00521	4	D6S1009	(TATC)11	5	1,1	1,4	1
HGDP01029	4	D2S405	(TAGA)12	5	-2,-2	-2,0	1
HGDP00778	4	D8S1108	(TCTA)11	4	0,0	0,0	2
HGDP01029	4	D15S818	(TAGA)10	4	3,3	3,3	2
HGDP00551	4	D1S1653	(TCTA)12	4	-1,0	-1,0	2
HGDP00521	4	D5S2500	(ATAG)11	4	0,1	0,1	2
HGDP00927	4	D10S1426	(TATC)11	4	0,2	0,2	2
HGDP01029	4	D17S1308	(TGTA)11 (-1)	4	-1,-1	-1,-1	2
HGDP00521	4	D17S1308	(TGTA)11 (-1)	4	-1,0	-1,0	2
HGDP00542	4	D17S1308	(TGTA)11 (-1)	4	-1,-1	-1,-1	2
HGDP00927	2	D3S3644	(AC)16	4	-1,0	0,0	0.5
HGDP00521	2	D9S1779	(AC)14	4	0,0	-2,-2	0
HGDP00521	2	D8S503	(AC)17	4	2,4	3,6	0
HGDP00521	2	D1S2682	(CA)10	3	0,10	0,10	2
HGDP00998	4	D2S427	(GATA)9	3	0,0	0,0	2
HGDP00542	4	D8S1113	(GGAA)12	3	-6,-6	-6,-6	2
HGDP01029	4	D10S1425	(GATA)11	3	-5,0	-5,0	2
HGDP01029	4	D7S1824	(TCTA)11	3	-3,-2	-3,-2	2
HGDP00778	4	D1S3669	(TATC)10	3	1,1	1,1	2
HGDP00711	4	D3S2432	(AGAT)15 (-3)	3	-3,0	-3,0	2
HGDP00491	4	D2S405	(TAGA)12	3	0,0	0,0	2
HGDP00998	4	D16S3253	(TAGA)9	3	0,0	0,0	2
HGDP00998	4	D9S301	(GATA)15	3	-7,-1	-7,-1	2
HGDP00998	4	D19S586	(TAGA)12 (+2)	3	1,2	1,2	2
HGDP00711	2	D15S165	(AC)21	3	-6,-6	-6,-6	2
HGDP00665	2	D20S103	(AC)16	3	-1,-1	-1,-1	2
HGDP00456	3	D4S2394	(ATT)11	3	0,0	0,0	2
HGDP00711	4	D11S2371	(TATC)11	3	1,1	1,1	2
HGDP00927	4	D10S1239	(ATCT)11	3	0,1	0,1	2
HGDP00521	4	D14S1434	(GATA)10	3	0,0	0,0	2
HGDP00491	4	D19S591	(TAGA)10 (-2)	3	-1,0	-1,0	2
HGDP00521	4	D19S591	(TAGA)10 (-2)	3	-2,1	-2,1	2
HGDP00542	4	D19S591	(TAGA)10 (-2)	3	-1,0	-1,0	2
HGDP00551	4	D19S591	(TAGA)10 (-2)	3	-2,0	-2,0	2
HGDP01224	4	D19S591	(TAGA)10 (-2)	3	-2,1	-2,1	2
HGDP00927	4	D17S2196	(AGAT)9 (-2)	3	0,2	0,2	2
HGDP01029	4	D2S1391	(ATCT)14	3	-2,0	-2,0	2
HGDP00456	3	D4S2361	(TTA)13	3	-1,-1	-1,-1	2
HGDP00665	4	D1S1653	(TCTA)12	3	0,0	0,0	2
HGDP01224	4	D1S1653	(TCTA)12	3	-2,-1	-2,-1	2
HGDP00542	4	D5S2500	(ATAG)11	3	0,0	0,0	2
HGDP00778	4	D10S1426	(TATC)11	3	1,1	1,1	2
HGDP01029	4	D5S2500	(ATAG)11	3	-2,0	-2,0	2
HGDP00456	4	D17S1298	(TGAA)8	3	0,0	0,0	2
HGDP00927	4	D17S1298	(TGAA)8	3	3,3	3,3	2
HGDP00542	4	D19S254	(AGAT)13 (-6)	3	-1,1	-1,1	2
HGDP00711	2	D1S2682	(CA)10	3	0,0	0,10	1
HGDP00998	4	D5S1457	(ATAG)9	3	0,0	0,1	1
HGDP00521	2	D20S103	(AC)16	3	2,2	-1,2	1
HGDP00998	4	D20S482	(TCTA)14	3	0,0	0,1	1
HGDP01224	3	D9S910	(ATA)14	3	-7,-7	-7,-7	1
HGDP01224	4	D11S2363	(TATC)14	3	-1,-1	-1,9	1
HGDP00711	4	D19S591	(TAGA)10 (-2)	3	-1,-1	-1,0	1
HGDP00927	2	D18S1390	(TG)18	3	-1,0	-2,-1	0.5
HGDP00778	2	D8S503	(AC)17	3	2,3	3,3	0.5
HGDP00778	4	D12S1300	(TAGA)12	3	2,4	2,2	0.5
HGDP00491	2	D9S1779	(AC)14	3	0,9	-1,6	0

Table 2.6: **Comparison of lobSTR allelotype calls to the CEPH-HGDP results.** Differences between the RefSeq sequence and hg18 are indicated in parentheses. [a]Converted allelotypes given in number of repeat units different from the reference. [b]Status: 2 = both alleles called correctly, 1 = one allele of a heterozygous locus called correctly, 0.5 = one allele called correctly and one incorrectly, 0 = no correct alleles called.

2.9.5 Supplemental Table 5

Genome (platform)	Coverage	Input reads	Avg. Read length	STR Aligned reads / Mbp input	% Non-unit Reads*	STRs ≥1x	STRs ≥3x
Venter (Sanger)	7.5x	12.5M	996	24.78	7.30%	127,017	41,261
Watson (454)	7.4x	75M	183	10.41	25.00%	83,079	25,488
Moore (IonTorrent)	10.6x	860M	261	0.79	43.50%	65,758	13,413
Ajay, et al (Illumina)	126x	14B	100	4.36	8.00%	180,309	167,175

Table 2.7: **lobSTR performance on four sequencing platforms.** *Reads differing by a non-integer number of copies of the STR motif from the reference. (M = million, B = billion).

2.9.6 Supplemental Table 6

Repeat unit size	# STR loci	Percentage
2	106,457	44%
3	17,383	7%
4	70,847	30%
5	28,746	12%
6	16,626	7%
Total	240,059	100%

Table 2.8: **STR reference repeat unit size distribution.**

2.9.7 Supplemental Table 7

Repeat unit	# STR loci
AC	66,992
AT	25,661
AAAT	20,319
AG	13,778
AAAG	12,553
AAAAC	10,015
AAGG	9,862
AAAC	8,842
AGAT	7,127
AAAAT	7,115

Table 2.9: **Most frequent reference STR repeat units.**

Chapter 3

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.

3.1 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 [**?**]. To date, STR mutations have been linked to at least 40 monogenic disorders [**?**, **?**], 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 [?]. In addition to Mendelian conditions, multiple studies have suggested that STR variations contribute to an array of complex traits [?], ranging from the period of the circadian clock in Drosophila [?] to gene expression in yeast [?] and splicing in humans [?, ?].

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 [?, ?]. 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 [?]. Finally, the vibrant genetic genealogy community extensively uses these loci to develop impressive databases containing lineages for hundreds of thousands of individuals [?].

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 [?, ?] and several hundred loci involved in forensic analysis, genetic genealogy, or genetic diseases [?, ?]. 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 [?]. 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 [?] and 454 [?]. 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 [?]. 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 [?]. These events, called stutter noise, generate erroneous reads

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

In our previous studies, we created publicly available programs that specialize in STR profiling using Illumina whole-genome sequencing data [?, ?]. 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 [?], 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.

3.2 Results

3.2.1 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 [?, ?, ?]. 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 [?, ?, ?, ?, ?, ?]. 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 [?, ?]. We utilized this approach to identify STR loci in the human genome while controlling the false positive rate [?]. We first integrated the purity, composition, and length of putative STRs in the genome into a score using Tandem Repeats

Finder [TRF] [?]. 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 [?, ?, ?] [?]. 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 $\sim 1\%$ of loci that exhibited repeat variability in the permissive call set [?]. 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 [?]. 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.

3.2.2 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 3.4, [?]**). 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 [?] and RepeatSeq [?]. 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 [**?**]. 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 3.2.2a**) and 350,000 STR loci per individual (**Figure 3.2.2b**), 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 3.4, Figure 3.2.2c**). 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.

3.2.3 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 3.2.3a**). 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 electrophoresis, the gold standard for STR calling (**Methods 3.4**). We focused on datasets containing Marshfield and PowerPlex Y chromosome panel genotypes that are available for a

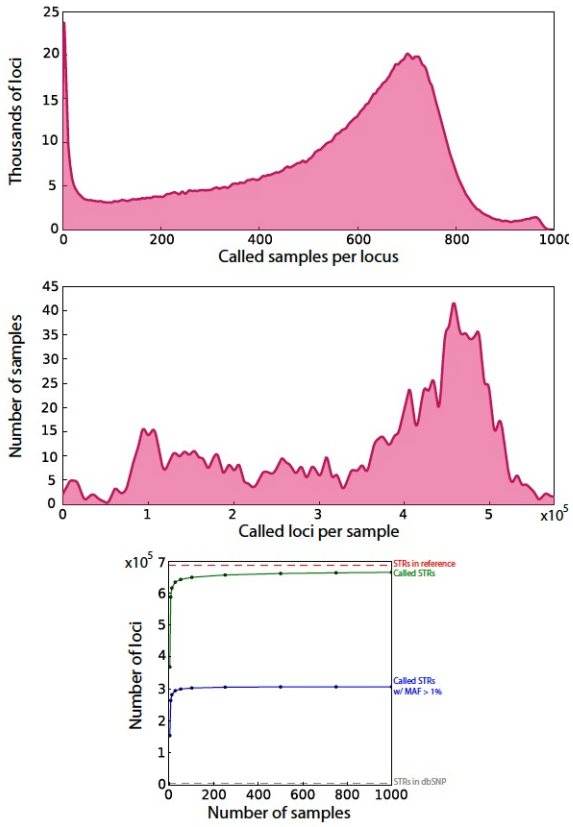


Figure 3-1: **Call set statistics (A) Distribution of the number of called samples per locus.** The average is 528 samples per STR with a standard deviation of 231 **(B) Distribution of the number of called loci per sample.** The average is 349,892 STR per sample with a standard deviation of 145,135 **(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).

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 [**?**] 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 [**?**, **?**, **?**, **?**]. 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 3.2.3b; [?]**). 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 [**?**]. For the Y chromosome, 95% of the lobSTR genotypes exactly matched the capillary genotypes for the PowerPlex Y panel [**?**].

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 4). 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 [**?**].

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 3.4**) 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 3.2.3c**). 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 [**?**]. 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 3.2.3d; [?]**). A previous study reported the spectra of these loci by genotyping ∼200 Caucasians in the United States using capillary electrophoresis [**?**]. 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 3.2.3e**). 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.

3.2.4 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

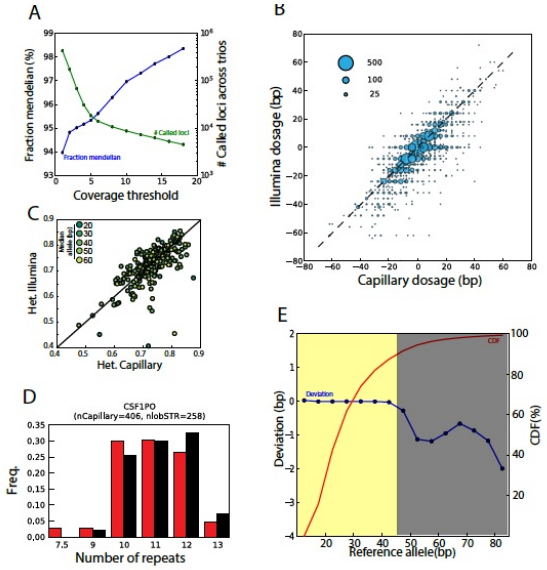


Figure 3-2: **Quality assessments of the STR catalog (A) Consistency of lobSTR calls with Mendelian inheritance.** The blue line denotes the fraction of STR loci that followed Mendelian inheritance as a function of the read coverage threshold. The green line denotes the total number of calls in the three 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).

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 [?], 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 3.2.4a; [?]**). Second, we focused on the 100 most heterozygous autosomal loci in our catalog and inspected the ability of STRUCTURE [?] 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 3.2.4b**). 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 3.2.4c**). 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.

3.2.5 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 [?] 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 supported by a much larger trio-based study of nearly 2500 STRs [?]. However, various other studies have suggested that dinucleotides have higher mutation rates [?, ?]. 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 our catalog. We found that for noncoding STRs, variability monotonically decreased with motif length (**Figure 3.2.5**). 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

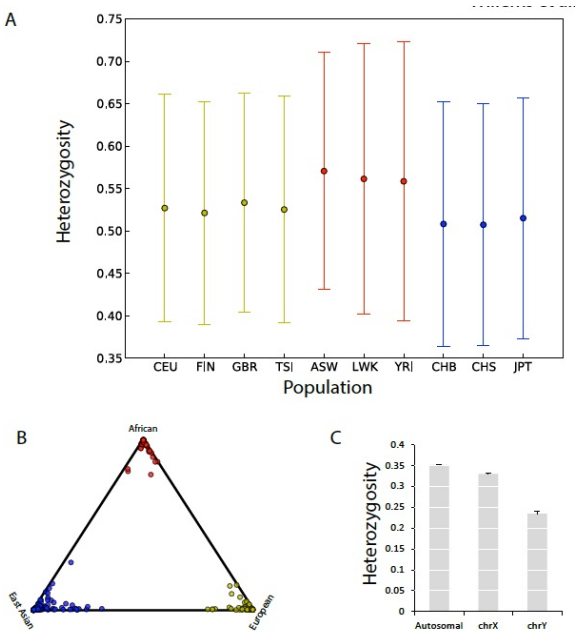


Figure 3-3: **Evaluation of the STR catalog for population genetics (A) Genetic diversity of the 10% most heterozygous autosomal loci in different populations.** Yellow: European, Red: African, Blue: East Asian. The mean heterozygosities (dot) of the African subpopulations consistently exceed those of the non-African subpopulations. The whiskers extend to Âś one standard deviation. See Supplemental Table 3 for population abbreviations **(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.

demonstrated significantly reduced heterozygosity compared to noncoding STRs for periods 2-5bp (Mann-Whitney U test; $p < 0.01$, [**?**]) 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 [**?**]. 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 [**?**]. This observation is concordant with previous

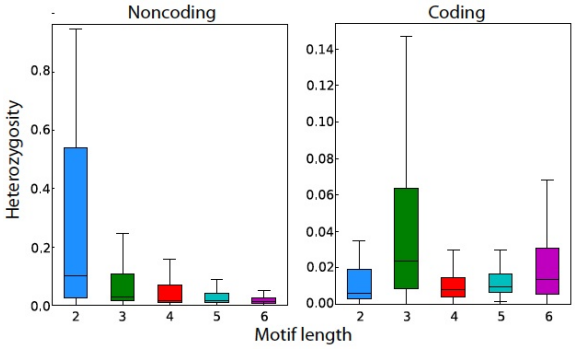


Figure 3-4: **Motif length and coding capabilities as determinants of STR variability.** STR heterozygosity monotonically decreases with motif length for noncoding loci and is generally reduced in non-coding (left) versus coding regions (right). The box extends from the lower to upper quartiles of the heterozygosity distribution and the interior line indicates the median. The whiskers extend to the most extreme points within 1.5*IQR of the quartiles.

studies [?, ?, ?].

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 [?, ?], whereas other studies claimed that AT repeats are the most variable motif [?, ?]. We repeated our analysis by stratifying the STRs based on motif sequence and major allele length [?]. The resulting per-motif variability results were remarkably similar with those generated using a comparison of orthologous STRs in humans and chimps [?]. 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 [?].

3.2.6 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 [**?**].

Next, we turned to finding the prototypal 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 [**?**]. 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} [**?**] for each STR between African, Asian, and European populations for STRs with heterozygosity above 5% [**?**]. 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 R_{st} 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 [**?**].

3.2.7 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 3.2.8a**). 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 [**?**]. 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* [**?**]. Out of these 5 genes, only *GP6* has known implications for a Mendelian condition: a mild platelet-type bleeding disorder [**?**, **?**]. 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.

3.2.8 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 [**?**]. 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.

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 3.2.8b**). 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

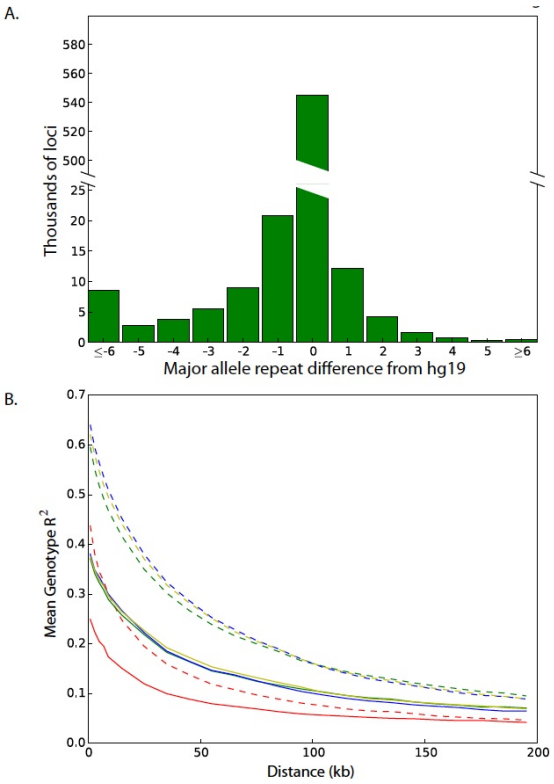


Figure 3-5: **Population-scale analyses of STR variation (A) Distribution of base-pair differences between each locus' most common allele and the NCBI reference allele (B) Patterns of linkage disequilibrium for SNPs and STRs on the X chromosome.** SNP-SNP LD (dashed lines) generally exceeds SNP-STR LD (solid lines) across a range of distances and for Africans (red), Admixed Americans (green), Europeans (yellow) and East Asians (blue).

with nearby SNPs, which likely stems from their higher mutation rates [**?**]. 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 [**?**].

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.

3.3 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 [**?**] . Similar catalogs, albeit to lesser degrees of completeness, have been produced for other types of variations, such as LINE-1 insertions [**?**] and Alu repeat variations [**?**]. 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 [**?**]. 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 [**?**]. 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 [**?**]Kirby et al. 2013). 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* [**?**], *HTT* [**?**] and *NOS3* [**?**]. 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 [?, ?]. Our hope is that this catalog can be a reference point to test this hypothesis in future studies.

3.4 Methods

3.4.1 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 [?]. 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 [?].

3.4.2 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.

3.4.3 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%.

3.4.4 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.

3.4.5 Capillary electrophoresis comparison

Capillary electrophoresis comparison Marshfield genotypes [?] were downloaded from http://www.stanford.edu/group/rosenberglab/data/rosenbergEtAl2005/combined-1048.stru.

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 R2 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.

3.4.6 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.

3.4.7 Summary statistic comparisons

The allelic spectra of the Marshfield panel were downloaded from http://research.marshfieldclinic.org/genetics/genotypingData_Statistics/markers/ and parsed using a custom Perl script (data and script available on https://github.com/erlichya/str_catalog_supplemental_scripts). 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 [?] . 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.

3.4.8 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.

3.4.9 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.

3.4.10 Sample clustering

STRUCTURE version 2.3.4 was utilized to perform the MCMC-based clustering. 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.

3.4.11 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.

3.4.12 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 $\sim$83% of our reference set of STRs.

3.4.13 R_{ST} levels

The Rst was calculated according to Slatkin [?] 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.

3.4.14 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 STRâĂŹs 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 R2 between the two SNP genotype indicator variables.

3.5 Acknowledgements

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.

Chapter 4

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 ariation in humans. *Nature Genetics*. (2015).

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.

4.1 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 [?, ?, ?]. These observations suggest that understanding the genetic architecture of the transcriptome may provide insights into the cellular-level mediators underlying complex traits [?, ?, ?]. 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 [?, ?, ?]. 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 [?, ?]. It has been speculated that such heritability gaps could indicate the involvement of repetitive elements that are not well tagged by common SNPs [?, ?].

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 [?, ?]. 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 [?, ?]. 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 [?].

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 [?, ?], change the distance between promoter elements [?, ?], alter splicing efficiency [?, ?], and induce irregular DNA structures that may modulate transcription [?]. In vivo experiments have reported specific examples of STR variations that control gene expression across a wide range of taxa, including Haemophilus influenza [?], Saccharomyces cerevisiae [?], Arabidopsis thaliana [?], and vole [?]. Recent studies reported that dinucleotide repeats are a hallmark of enhancers in Drosophila and are enriched in predicted enhancers in humans [?]. Human promoters also disproportionately harbor STRs [?] and the presence of STRs in promoters or transcribed regions greatly increases the divergence of gene expression profiles across great apes [?], 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 [?, ?, ?, ?, ?] and alternative splicing [?, ?, ?]. 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 [?]. 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.

4.2 Results

4.2.1 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 [**?**] project and whose whole genomes were sequenced by the 1000 Genomes Project [**?**]. The STR genotypes were obtained in our previous study [**?**] 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 [**?**]. 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 [**?**]. 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 4.7** and **Supplementary Fig. 4.8.1**).

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 [?, ?, ?] or disease phenotypes [?, ?]. As covariates, we included sex, population structure, and other technical parameters (**Fig. 4.2.1a** and **Supplementary Note 4.7**). 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 (**Online Methods 4.6**) and were within 100kb of the transcription start and end sites of the gene transcripts as annotated by Ensembl [?]. On average, 13 STR loci were tested for each gene (**Supplementary Fig. 4.8.2**), yielding a total of 190,016 STR×gene tests.

Our analysis identified 2,060 unique protein-coding genes with a significant eSTR (gene level FDRâĽď5%) (**Fig. 4.2.1b** and **Supplementary Data Set 1** (see *Nature Genetics* website)). The majority of these were di- and tetra-nucleotide STRs (**Supplementary Tables 4.9.1** and **4.9.2**). 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 4.9.3**). Overall, there was no bias in direction of effect (**Supplementary Table 4.9.4**). 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 (**Fig. 4.2.1b** and **Supplementary Fig. 4.8.3**). 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 (**Online Methods 4.6**). 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 4.7** and **Supplementary Fig. 4.8.4**). Three previous studies described candidate gene studies of expression STRs and involved STRs that were tested in our framework [?, ?, ?]. 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 4.9.5**).

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

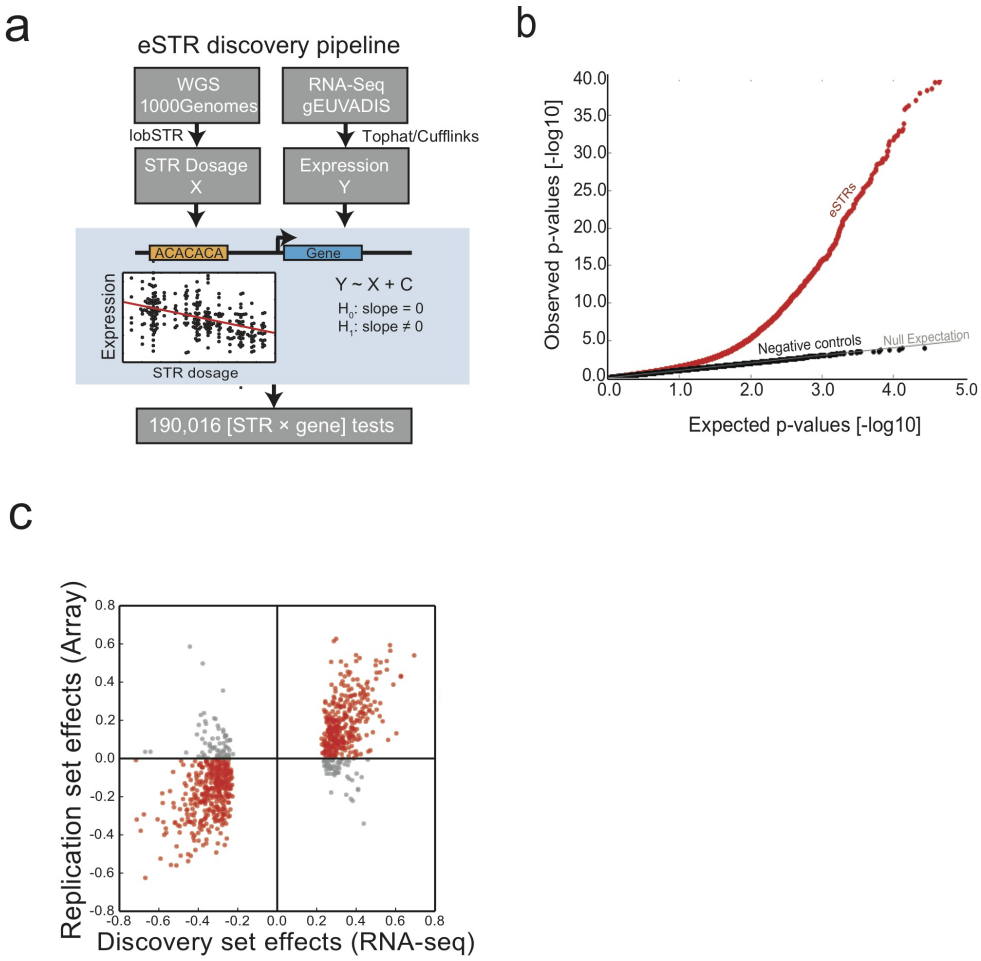


Figure 4-1: **eSTR discovery and replication. (a)** eSTR discovery pipeline. An association test using linear regression was performed between STR dosage and expression level for every STR within 100kb of a gene **(b)** Quantile-quantile plot showing results of association tests. The gray line gives 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.

LCL expression profiles were measured by Illumina expression array [?]. 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 ($R = 0.73$, $p = 1.4 \times 10^{-149}$) (**Fig. 4.2.1c**), despite only moderate reproducibility of expression profiles across platforms (**Supplementary Note 4.7** and **Supplementary Fig. 4.8.5**). 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.

4.2.2 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 [**?**, **?**] but there are known examples of STRs tagging GWAS SNPs [**?**].

To address this question, we partitioned the relative contributions of eSTRs versus all common (MAF$\geq$1%) bi-allelic SNPs, indels, and structural variants (SV) in the cis region of each gene using a linear mixed model (LMM) (**Fig. 4.2.2a**). 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 [**?**, **?**]. 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})âĄĎ(h^2_{cis}) ≈ 0 across all conditions (**Supplementary Note 4.7** and **Supplementary Fig. 4.8.6, 4.8.7**), 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 4.7** and **Supplementary Fig. 4.8.8**), 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% (**Fig. 4.2.2b**), with a median ratio of (h^2_{STR})âĄŠ(h^2_{cis}) of 12.3% ($CI_{95\%}$ 11.1%-14.2%; $n = 1,928$) (**Supplementary Table 4.9.6**). We repeated the same analysis for genes with at least moderate ($\geq 5\%$) *cis*-heritability (**Online Methods 4.6**) regardless of the presence of a significant eSTR in the discovery set. The motivation for this analysis was to avoid potential winner's curse [**?**] and to obtain a transcriptome-wide perspective on the role of STRs in gene expression (**Fig. 4.2.2c**). 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 4.9.6**). Repeating the analysis while considering STRs as a random effect showed highly similar results (**Supplementary Note 4.7**, **Supplementary Table 4.9.7**, and **Supplementary Fig. 4.8.9**). 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.

4.2.3 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 (**Fig. 4.2.3a**). 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 ($R = 0.52$; $p = 3.2 \times 10^{-130}$) (**Fig. 4.2.3b**). This further supports the additional role of eSTRs beyond traditional cis-eQTLs.

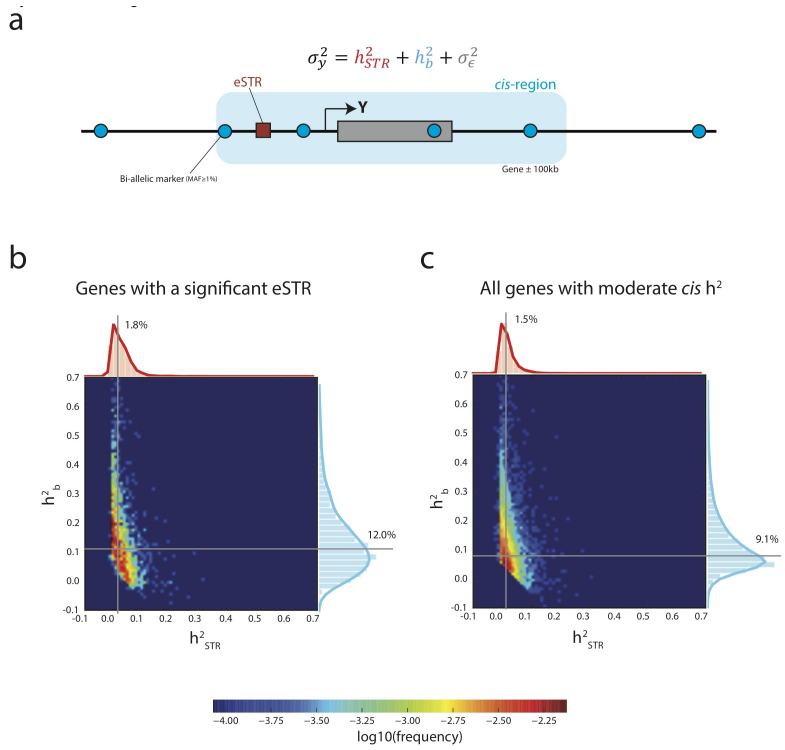


Figure 4-2: **Variance partitioning using linear mixed models (a)** The normalized variance of the expression of gene Y was modeled as the contribution of the best eSTR and all common bi-allelic markers in the cis region (Âś100kb from the gene boundaries) **(b-c)** Heatmaps show the joint distributions 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.

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%) (**Fig. 4.2.3c-e** and **Online Methods 4.6**). 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.

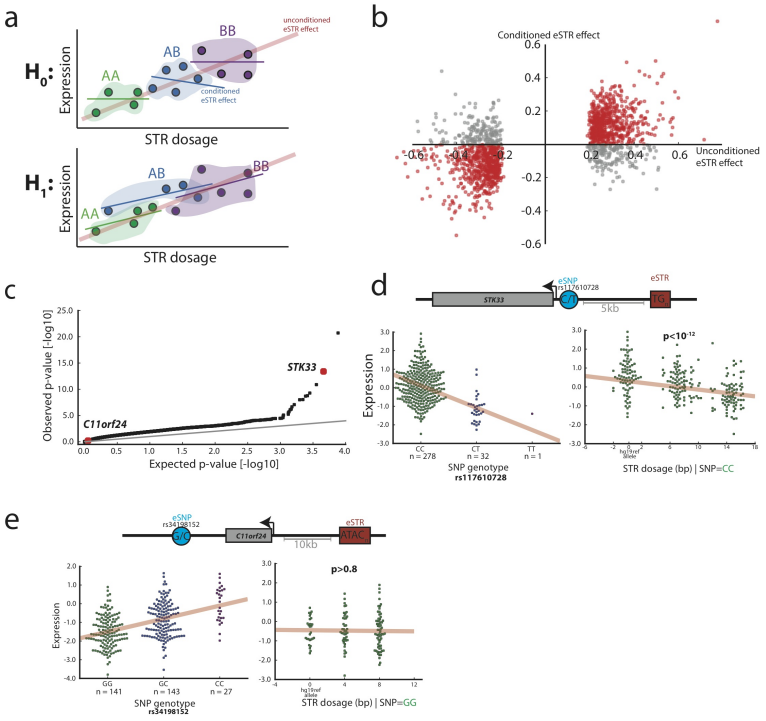


Figure 4-3: **eSTR associations in the context of eSNPs (a)** Schematic of the eSTR effect versus the effect conditioned on the lead eSNP genotype. Under the null expectation, the original association (red line) comes from mere tagging of eSNPs. Thus, the eSTR effect disappears when restricting to a 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 âĂIJCâĂİ 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 âĂIJGâĂİ eSNP allele (bottom left, green dots), the STR effect is lost (bottom right).

4.2.4 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 [**?**]. We inspected the sequence conservation [**?**] across 46 vertebrates in the sequence upstream and downstream of the eSTRs in our discovery dataset (**Fig. 4.2.4a**). 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 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 $\pm$50bp showed a more significant contrast in the conservation scores of the eSTRs versus the control STRs ($p<0.01$) (**Fig. 4.2.4a** 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 (**Fig. 4.2.4b**) and to a lesser extent in or near predicted enhancers (**Supplementary Fig. 4.8.10**). We also inspected the co-localization of eSTRs with histone modifications as annotated by the Encode Consortium [**?**] 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) (**Fig. 4.2.4b**). 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 [**?**]. 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 4.9.8**). We also performed a peak-shifting analysis using ChromHMM annotations [**?**] (**Fig. 4.2.4c**) 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 (**Online Methods 4.6** and **Supplementary Fig. 4.8.11**). We introduced different eSTR alleles to GERV [**?**], 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 hypersensi-

tivity 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.

4.2.5 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 [**?**] 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 4.2.5a** and **Supplementary Fig. 4.8.12**). 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 [**?**] 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 [**?**], some of the discovered eSTRs are shared across tissues and could play a role

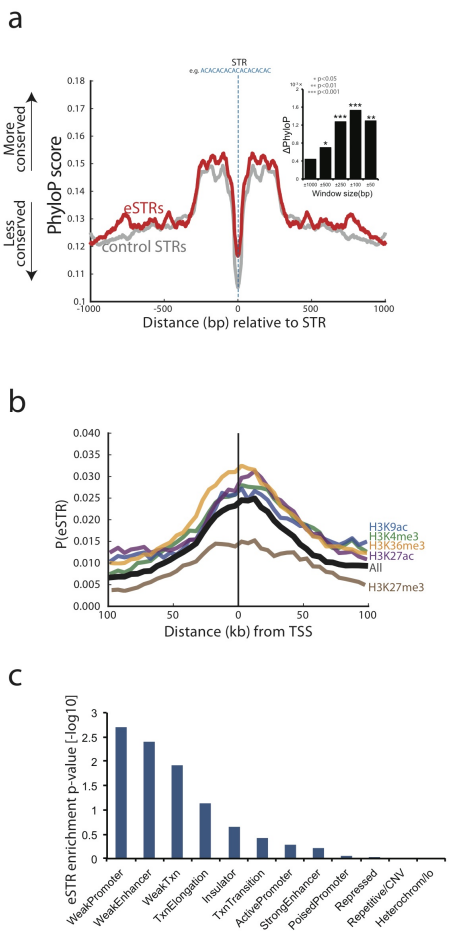


Figure 4-4: **Conservation and epigenetic analysis of eSTR loci (a)** Median PhyloP conservation score as a function of distance from the STR. Red: eSTR loci, gray: matched control STRs. Inset: the difference in the PhyloP conservation score between eSTRs and matched control STRs as a function of window 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.

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 4.2.5b** and **Supplementary Table 4.9.9**). 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 [**?**, **?**, **?**, **?**]. 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.

4.3 Discussion

Repetitive elements have often been considered as neutral with no phenotypic consequences [**?**]. 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

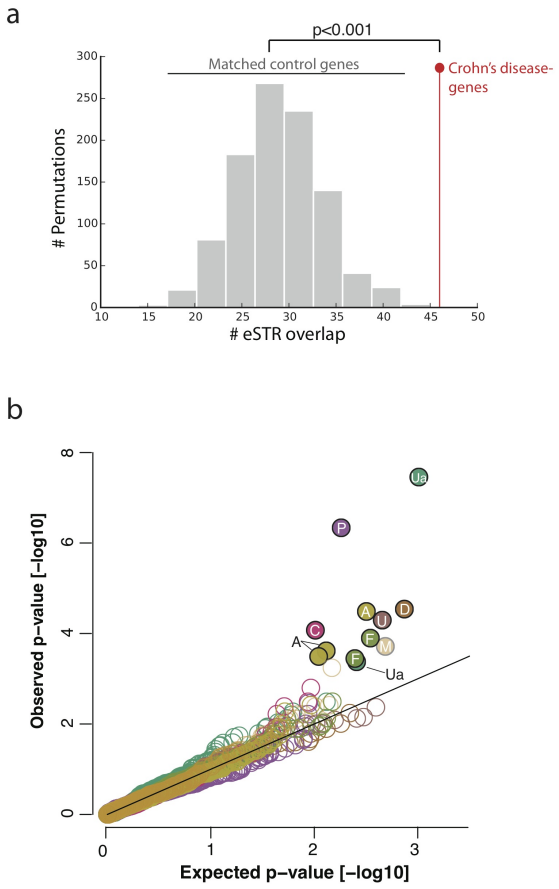


Figure 4-5: **Association of eSTRs with clinical phenotypes (a)** The overlap between eSTRs and Crohn's disease GWAS genes (red) versus random subsets of genes (gray) matched on expression and heritability profiles in LCLs **(b)** quantile-quantile plots of eSTR associations in the TwinsUK data. Only 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.

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(Willems et al. 2014). Second, based on previous findings in humans [**?**, **?**, **?**], 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 [**?**], 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 [**?**]). 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 [**?**] 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 [**?**]. 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.

4.4 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.

4.5 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.

4.6 Online Methods

4.6.1 Genotype datasets

lobSTR genotypes were generated for the phase 1 individuals from the 1000 Genomes Project as described in [**?**]. 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 [**?**] 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.

4.6.2 Targeted sequencing of promoter region STRs

We used a previously published method using capture and high-throughput sequencing [**?**] 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.

4.6.3 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 [**?**, **?**] ensGene table. The RNA-sequencing reads were mapped to the Ensembl transcriptome using Tophat v2.0.7 [**?**] with default parameters. Gene expression levels were quantified using Cufflinks v2.0.2 [**?**] 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. [**?**]. Background corrected and summarized probeset intensities (by Illumina software) contained values for 7,655 probes. Additionally, probes containing common SNPs were removed [**?**]. 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.

4.6.4 eQTL association testing

Expression values were adjusted for individual sex, individual population membership, gene expression heterogeneity, and population structure (**Supplementary Note 4.7**). 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:

$$\vec{y}_g = \alpha_g + \beta_{j,g}\vec{x}_j + \vec{\epsilon}_{j,g} \tag{4.1}$$

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).

4.6.5 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%.

4.6.6 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:

$$\vec{y}_g = \alpha_g + \beta_{j,g}\vec{x}_j + \vec{u}_g + \vec{\epsilon}_{j,g} \tag{4.2}$$

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(Yang et al. 2011) 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:

$$var(\vec{y}_g) = \beta^2_{j,g} var(\vec{x}_j) + \sigma^2_{u_g} K_g + \sigma^2_{\epsilon_{j,g}} I_n \tag{4.3}$$

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:

$$\sigma^2_p = h^2_{STR} + h^2_b + \sigma^2_\epsilon \tag{4.4}$$

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 4.7**).

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.

4.6.7 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:

$$\texttt{Model 1 (eSNP-only):} \quad \vec{y}_g = \alpha_g + \beta_{eSNP,g}\vec{x}_{eSNP,g} + \vec{\epsilon}_{j,g} \tag{4.5}$$

$$\texttt{Model 2 (joint eSNP+eSTR):} \quad \vec{y}_g = \alpha_g + \beta_{eSNP,g}\vec{x}_{eSNP,g} + \beta_{eSTR,g}\vec{x}_{eSTR,g} + \vec{\epsilon}_{j,g} \tag{4.6}$$

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 4.7**), 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.

4.6.8 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.

4.6.9 Enrichment of STRs and eSTRs in predicted enhancers

H3K27ac peaks produced by the ENCODE Project [?] 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.

4.6.10 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, naÃŕve 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.

4.6.11 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)[**?**] 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.

4.6.12 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 âĂIJReported Gene(s)âĂİ and âĂIJMapped_geneâĂİ. 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.

4.6.13 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 [**?**] 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.

4.7 Supplementary Notes

4.7.1 STR genotype error reduces power to detect eSTRs

We performed simulations to evaluate the effect of lobSTR genotype errors on our power to detect eSTR associations. We used capillary electrophoresis calls from the Marshfield panel as ground truth genotypes and lobSTR calls for the same markers in our catalog as observed genotypes. We filtered for loci with at least 25 calls for comparison. For each gene, we simulated expression values assuming a single causal STR per gene that explains h^2_{STR} percent of expression variance. We performed the analysis for h^2_{STR} equal to 0.01, 0.05, 0.1, 0.3, and 0.5. Expression values were simulated as follows:

$$Y_i = \beta X_i + \epsilon_i \tag{4.7}$$

where Y_i is the expression level for individual i, X_i is the true STR dosage for individual i, $\beta = \sqrt{h^2_{STR}}$ is the effect size of the STR, and $\epsilon_i \sim N(0, 1 - h^2_{STR})$ is the residual term for individual i.

We performed association analysis regressing $\vec{Y}$ on both $\vec{X}$ and $\vec{X}'$, where $\vec{X}'$ are the observed STR dosages, and tested whether β was significantly different than 0 in each case ($p<0.01$). We found that genotype errors limit our power to detect eSTRs (**Supplementary Fig. 1a**) and cause us to underestimate the true variance explained by STRs (**Supplementary Fig. 1b**) but do not introduce spurious eSTR signals.

4.7.2 Controlling for covariates

We controlled for a number of covariates by regressing them out of the expression dataset. The covariate-corrected expression matrix is given by:

$$Y = (1 - H)Y' \tag{4.8}$$

where Y' is an $n \times m$ matrix of normalized expression values, Y is an $n \times m$ matrix of residualized expression values, n is the number of individuals, m is the number of genes, $H = C(C^T C)^{-1} C^T$ is the hat matrix, and C is an $n \times c$ matrix of c covariates. Specifically, the columns of C consist of the following sub-matrices:

$$C = \left[\begin{array}{c|c|c|c} \vec{c_s} & C_p & C_{exp} & C_{popstruct} \end{array} \right] \tag{4.9}$$

1. **Individual sex**: this is a binary vector, $\vec{c_s} \in \{0,1\}^{n \times 1}$, where 0 denotes female and 1 male.
2. **Individual population membership**: this is a binary matrix $C_p \in \{0,1\}^{n \times pop-1}$. A "1" in position $C_p(i,j)$ denotes that individual i belongs to population j. Specifically, pop is equal to to 4 for the association tests with the gEUVADIS RNA-seq data.
3. **Gene expression heterogeneity**: Y' is a matrix that consists of all $\vec{y_g}$ as its column vectors, where $\vec{y_g}$ is a vector of expression values for gene g. To reduce variation due to experimental differences or other unidentified confounding factors across expression datasets, the top 10 principal components (PCs) corresponding to the top 10 eigenvectors of $Y'Y'^T$ were included as covariates for both the array and RNA-sequencing datasets. $C_{exp} \in \mathbb{R}^{n \times 10}$ indicates the matrix of the top 10 PCs.

4. **Population structure**: We first preprocessed the HapMap SNP dataset to include SNPs with MAF > 10%. We used Plink [**?**] for LD-pruning with a pairwise correlation threshold of 0.5, a window size of 50 SNPs, and a step size of 5 SNPs. This left 286,010 SNPs for the RNA-sequencing dataset, which we used to correct for population structure. We used the Tracy-Widom test for population stratification proposed by Patterson, et al. [**?**] to determine the number of PCs to include as covariates. Let $C_{popstruct} \in \mathbb{R}^{n \times t}$ indicate the matrix of the top t PCs removed, where t=5 for the RNA-sequencing dataset.

Residualized expression values were then used as input to the eQTL analysis.

4.7.3 Validation of promoter eSTRs

To assess the affect of low sequencing coverage on our eSTR analysis, we performed targeted sequencing of 2,472 loci in promoter regions (TSS +/-1kb) in 120 CEU and YRI individuals (see **Online Methods**), 107 of which were genotyped as part of our 1000 Genomes STR catalog [**?**]. We used lobSTR v3.0.3 to call STR genotypes from these reads. The median number of informative reads per locus was 15.

We first used this callset to assess the accuracy of STR calls from the low coverage 1000 Genomes Project dataset. To ensure a high quality callset for comparison, analysis was restricted to calls with a minimum coverage of 5x and minimum lobSTR quality score of 0.5, leaving 1,293 loci. STR dosage was highly correlated between the high coverage vs. the 1000 Genomes calls ($r^2 = 0.74$) (**Supplementary Fig 4a**). Overall, 65% of individual genotypes were concordant, with 57.8% of discordant calls due to calling only a single allele at heterozygous sites in the low coverage data. The majority of incorrect allele calls were off by one (75%) or two (15.9%) repeat units (**Supplementary Fig. 4b**), suggesting stutter noise as the primary error source for these calls.

We next used the high coverage calls to validate our eSTR associations in promoter regions. We performed eSTR analysis on these 120 samples, 29 of which overlapped samples used in our discovery dataset. We filtered the high coverage callset to contain loci where at least 30 samples had a genotype call. As for the discovery analysis, we removed any genotypes seen less than three times. After filtering 126 eSTRs could be tested using the high coverage calls.

The majority of calls showed the same direction of effect (79%, p=9.9×10^{-12}, n=126) (**Supplementary Fig. 4c**). Effect sizes were inflated in the low coverage discovery dataset

(slope=0.83), as expected due to winner's curse resulting from low power. Overall, our results suggest that the majority of associations discovered in the low coverage data are replicable using higher quality genotype calls.

4.7.4 Comparing expression across array and RNA-sequencing datasets

To determine the reproducibility of expression profiling across platforms, we compared gene expression for the 122 individuals profiled by both array and RNA-sequencing. For each platform, we obtained a 122 $\times$ 4,627 matrix Y^{Array} and Y^{RNAseq}, where $Y^{Array}_{(i,g)}$ and $Y^{RNAseq}_{(i,g)}$ give the expression of gene g in individual i on the expression array and the RNA sequencing, respectively, before quantile normalization.

We measured the reproducibility of expression profiles inside subjects by calculating the Spearman rank correlation for each pair of row vectors $Y^{Array}_{(i,.)}$ and $Y^{RNAseq}_{(i,.)}$ for $i \in \{1..122\}$ (**Supplementary Fig. 5a**). The average Spearman correlation was 0.71. A previous study by Maroni et al. [**?**] measured technical reliability of RNA-seq versus array data with independent datasets. Importantly, they reported an average Spearman correlation of 0.73 for reproducibility of expression profiles inside subjects. This result provides additional support to the technical validity of our expression analysis pipeline.

eQTL replication requires that relative differences between subjects are reproducible across experiments. We compared the order of individuals at each gene as reported by the array and the RNA-sequencing data by measuring the Spearman rank correlation of the column vectors $Y^{Array}_{(.,g)}$ and $Y^{RNAseq}_{(.,g)}$ for $g \in \{1..4,627\}$ (**Supplementary Fig. 5b**). The concordance of rank-order of individuals across platforms was moderate (average Spearman rank correlation 0.22), which implies only moderate power to replicate QTLs across the two platforms. Choy et al. performed a similar analysis with biological replicates of LCLs in two expression arrays independent from our study [**?**]. They also reported Spearman rank correlations of 0.25-0.3 for relative differences of expression between subjects, in agreement with our analysis.

4.7.5 Partitioning heritability on simulated datasets

The lead STR can often exhibit high collinearity with other *cis* variants. To rule out the possibility that the LMM could be incorrectly partitioning variance to the STR in the case of tagging another causal variant nearby, we performed simulations in which there was a single

causal SNP eQTL per gene. For each gene, we simulated expression values using the following process:

1. Choose the lead SNP from the eQTL analysis on real data as the causal variant. Let this eQTL explain σ^2 percent of expression variance.
2. Simulate expression values as $y_i = \beta x_i + \epsilon_i$ where y_i is the simulated expression value for individual i, x_i is the SNP genotype for individual i, $\beta = \sqrt{\sigma^2}$, and $\epsilon \sim N(0, 1 - \sigma^2)$.
3. Run the LMM analysis as described in the **Online Methods** to determine h^2_{STR} and h^2_b.

Notably, this procedure simulates the causal SNP based on the SNP-eQTL analysis, rendering the test more realistic. The simulation was repeated for values of σ^2 equal to 0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5 for each gene. We performed this analysis for both the cases of treating the STR as a fixed and a random effect.

We observed that in both models, h^2_b was very close to the simulated value of σ^2, as expected. Importantly, the median value for h^2_{STR} was negative for the fixed effects case and 0 for the random effects case across all simulations with $h^2_b > 1\%$. The mean values were close to 0 for most realistic values of SNP-eQTL effects and slightly biased (< 0.005) upwards in the case of very strong SNP-eQTLs (**Supplementary Fig. 6**). The median ratio of h^2_{STR} to $h^2_{STR} + h^2_b$ was exactly 0 for the fixed effects case and $< 0.1\%$ for the random effects cases when $h^2_b > 1\%$. These findings suggest that our LMM analysis reflects an accurate partitioning of variance even in the presence of strong SNP-eQTLs.

To ensure that we realistically recapitulate the architecture of eSNPs in our data, we repeated these simulations using the observed σ^2 for the lead SNP per gene as the simulated value. For the fixed effects case the median ratio of h^2_{STR} to $h^2_{STR} + h^2_b$ was 0. Restricting to genes that had a significant eSTR resulted in a slightly elevated median of 1.4%, suggesting that in some cases h^2_{STR} is mildly inflated due to tagging. However, when we repeated the simulations assuming a single causal STR, the ratio was greater than 90% for eSTRs that explain at least 5% of variation in expression. This shows that our variance partitioning analysis correctly assigns variance to STRs in the case that they are causal, but assigns little or no variance to STRs in the presence of other causal variants nearby.

In some cases a gene may be genetically controlled by multiple causal eSNPs. To account for this scenario, we repeated the simulations assuming two causal SNPs per gene. We found that h^2_b tended to be close to the sum of variance explained by the two eSNPs, whereas the median

ratio of h^2_{STR} to $h^2_{STR} + h^2_b$ was again 0 (**Supplementary Fig. 7**). This suggests that our analysis is robust across a variety of eQTL architectures.

Finally, to validate that our estimators of h^2_{STR} are not inflated, we also ran the fixed effects LMM analysis on random pairs of eSTRs and local bi-allelic variants from chromosome 2 and gene expression profiles from chromosome 1. This generated a null distribution for h^2_{STR} in the case of no association. In this negative control condition, h^2_{STR} was distributed symmetrically around 0 with mean 7×10^{-4} and median -0.002, demonstrating that the estimator is unbiased.

4.7.6 STR genotype errors result in underestimating h^2_{STR}

We performed simulations to evaluate the effect of STR genotype errors on our variance partitioning analysis. For each STR, we simulated expression of a gene based on ground truth genotypes as described in the power analysis above. We assumed a single causal STR that explains h^2_{STR} percent of expression variance, where h^2_{STR} ranged from 0 to 0.5.

We performed a linear mixed model analysis with all SNPs within 100kb of the STR as one variance component and the STR as a fixed effect using either the ground truth STR genotypes or the observed genotypes reported by lobSTR (**Supplementary Fig. 8**). While the ground truth genotypes accurately recover the simulated value of h^2_{STR}, using observed genotypes results in a strong underestimation, suggesting that our analysis is quite conservative in measuring the contribution of STRs to explaining expression variability.

4.7.7 Treating STRs as random vs. fixed effects

In our LMM analysis to partition heritability between STRs and other *cis* variants, we treated the lead STR for each gene as a fixed effect. We repeated this analysis treating the STR as a random effect to determine whether this choice significantly affects our results. We used a model of the form:

$$\vec{y}_g = \alpha_g + \vec{v}_g + \vec{u}_g + \vec{\epsilon}_{j,g} \tag{4.10}$$

where:

- $\vec{v}_g$ is a length n vector of random effects for the lead STR

- $\vec{v}_g \sim MVN_n(0, \sigma^2_{v_g} S_g)$ with $\sigma^2_{v_g}$ the percent of phenotypic variance explained by the lead STR for gene g
- S_g is a standardized IBS relatedness matrix constructed using the lead STR. It was constructed as:

$$S_g = \frac{1}{var(\vec{x})}(\vec{x} - 1_n\overline{\vec{x}})(\vec{x} - 1_n\overline{\vec{x}})^T \tag{4.11}$$

  where $\vec{x}$ is a length n vector consisting of genotypes for the lead STR.
- All other variables are as described in the Online Methods.

We used the GCTA program [**?**] to determine the REML estimates of $\sigma^2_{u_g}$ and $\sigma^2_{v_g}$. GCTA encountered numerical problems using the `-reml-no-constrain` option, likely due to the small sample size for each gene and strong correlation between the STR and bi-allelic variance components. Therefore, estimates were constrained to be between 0 and 1 and are biased to be greater than 0.

The overall phenotypic variance-covariance matrix is:

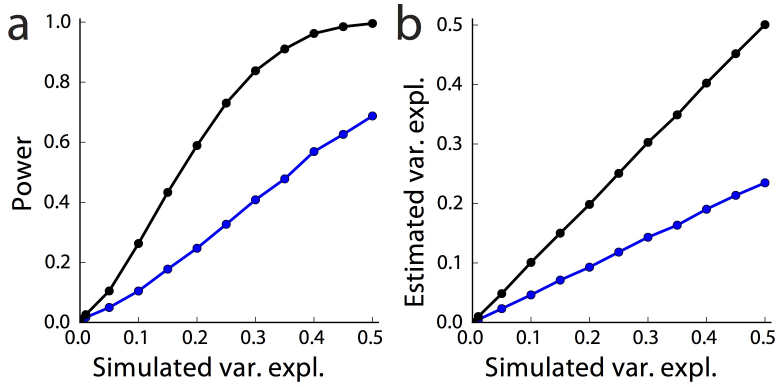
$$var(\vec{y}_g) = \sigma^2_{v_g} S_g + \sigma^2_{u_g} K_g + \sigma^2_{\epsilon_{j,g}} I_n \tag{4.12}$$

with $\sigma^2_{v_g}$ giving the percent of phenotypic variance explained by the lead STR (h^2_{STR}) and $\sigma^2_{u_g}$ giving the percent explained by other *cis* bi-allelic variants (h^2_b).

Estimates of the variance explained by STRs and by *cis* bi-allelic variants using this model are consistent with those obtained by treating STRs as a fixed effect (**Supplementary Table 6-7**). Because the random effects estimates are constrained to be between 0 and 1, the random effects model tended to partition variance all to a single variance component, but overall distributions of h^2_{STR} and h^2_b were similar to the fixed effects case (**Fig. 2b** and **Supplementary Fig. 9**).

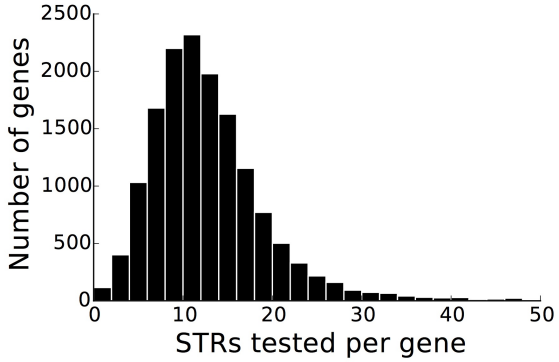
4.8 Supplementary Figures

4.8.1 Supplementary Figure 1



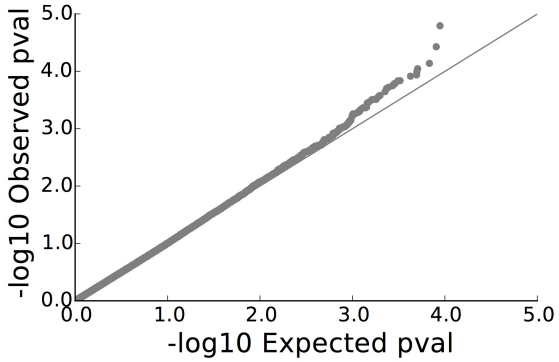
STR genotyping errors reduce power to detect eSTR associations. **a**. Power to detect associations and **b**. estimated variance explained for different simulated values of variance explained by the STR. (black: observed capillary electrophoresis genotypes, blue: lobSTR genotypes).

4.8.2 Supplementary Figure 2



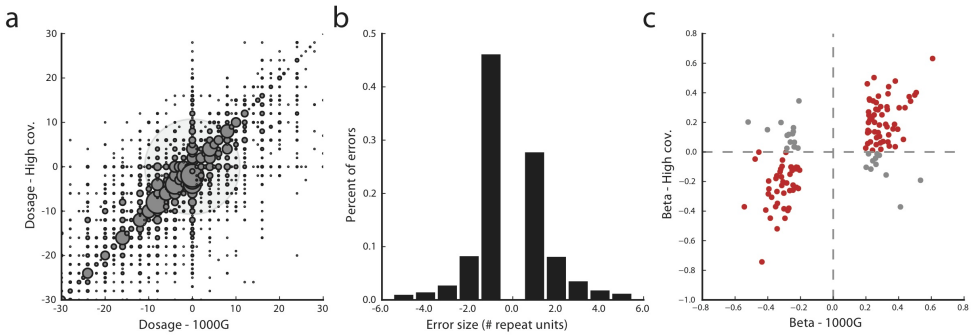
Number of STRs tested per gene. Histogram gives the number of STRs within 100kb of each gene that passed quality filters and were included in the eSTR analysis.

4.8.3 Supplementary Figure 3



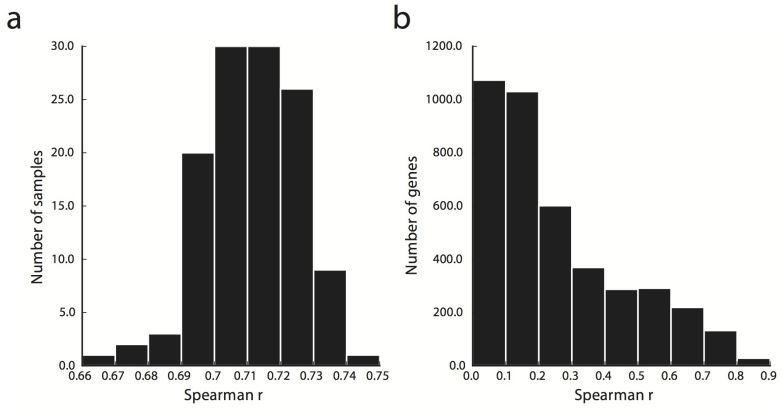
Unlinked controls follow the null. QQ plot of association tests between random unlinked STRs and genes.

4.8.4 Supplementary Figure 4



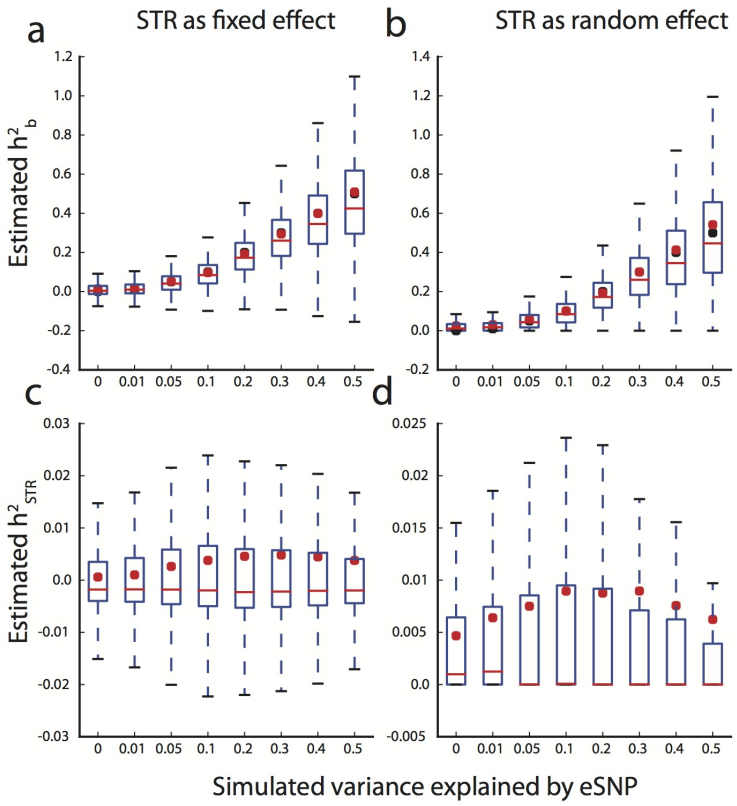
Validation of eSTR analysis using high coverage genotype calls. **a**. Comparison of STR dosage in low coverage 1000 Genomes calls vs. calls from high coverage targeted sequencing of promoter STRs. Bubble area represents the number of calls at each data point. For reference, the bubble at -20,-20 represents 176 calls. 0 denotes the reference allele. The transparent bubble in the center represents calls that are homozygous reference in both datasets. **b**. Distribution of the size of errors for discordant allele calls. The majority of errors (89.4%) are off by one or two repeat units. **c**. Comparison of eSTR effect sizes between the low and high coverage datasets. Red dots denote eSTRs with concordant effect directions.

4.8.5 Supplementary Figure 5



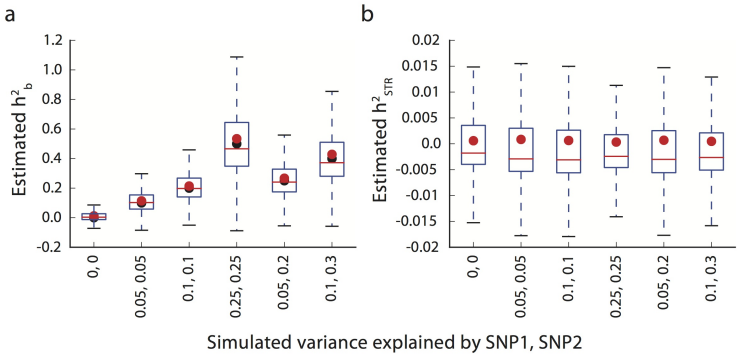
Expression values are moderately reproducible across platforms. **a**. Distribution of Spearman rank correlation coefficients between gene expression profiles of individuals measured on microarray vs. RNA-sequencing platforms. **b**. Distribution of Spearman rank correlation coefficients between the order of individuals ranked by expression levels across transcripts measured using microarray vs. RNA-sequencing platforms.

4.8.6 Supplementary Figure 6



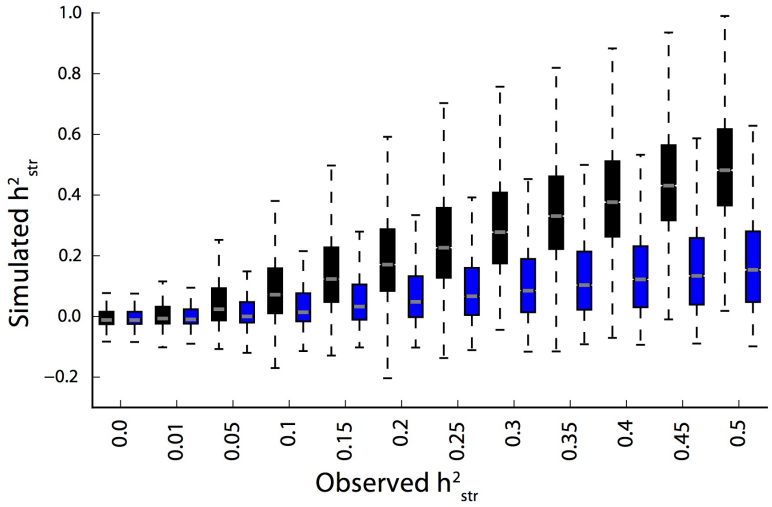
Variance partitioning simulations with a single causal SNP Plots show variance partitioning results from simulations in which each gene has a single causal eSNP. (**a**&**b**) The distributions of h_b^2 (**c**&**d**) The distributions of h_{STR}^2 (**a**&**c**) The LMM simulations with STRs as fixed effects (**b**&**d**) The LMM simulations with STRs as random effects (**a**-**d**) Black points denote the true value of the variance explained by the causal SNP. Red dots denote the average value of the estimator. Red bars denote the median value of the estimator. The figure shows that the median values of the lead STRs are largely insensitive to the presence of a strong SNP eQTL.

4.8.7 Supplementary Figure 7



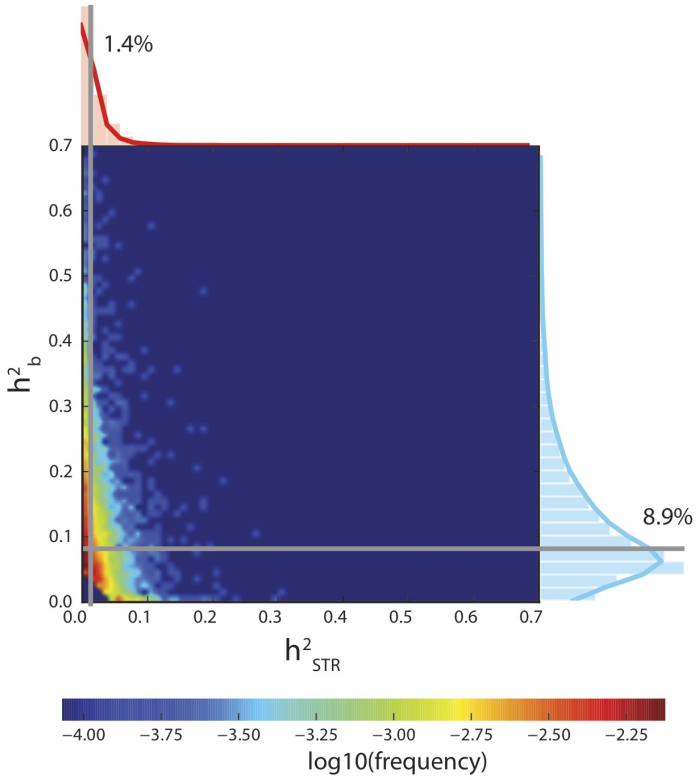
Variance partitioning simulations with two causal SNPs Plots show variance partitioning results from simulations in which each gene has two causal eSNPs. **a** The distributions of h^2_b. **b** The distributions of h^2_{STR}. Black points denote the true value of the variance explained by the causal SNPs. Red dots denote the average value of the estimator. Red bars denote the median value of the estimator.

4.8.8 Supplementary Figure 8



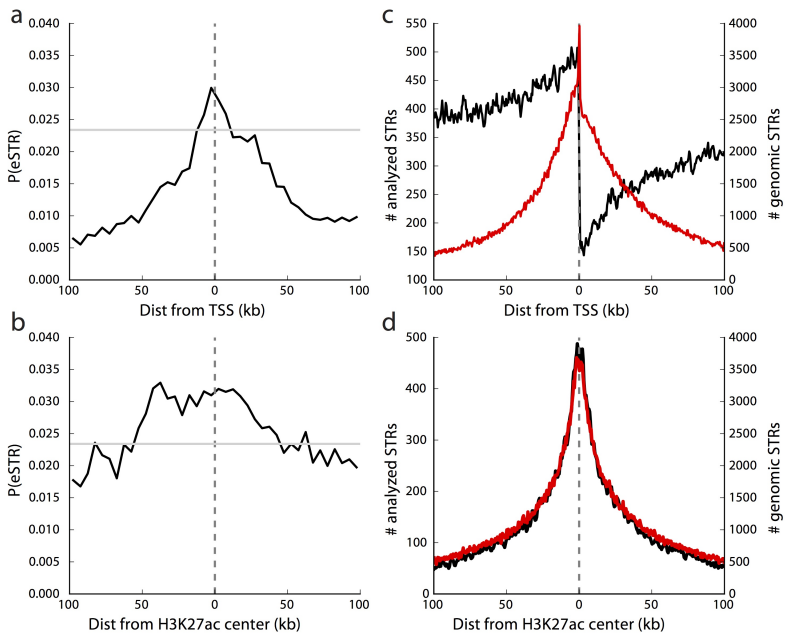
STR genotype errors cause underestimation of h^2_{STR}. The distribution of observed h^2_{STR} for each simulated value of h^2_{STR} is shown for an LMM analysis conducted using true genotypes (black) vs. observed genotypes (blue). In the presence of genotyping errors, h^2_{STR} is strongly underestimated.

4.8.9 Supplementary Figure 9



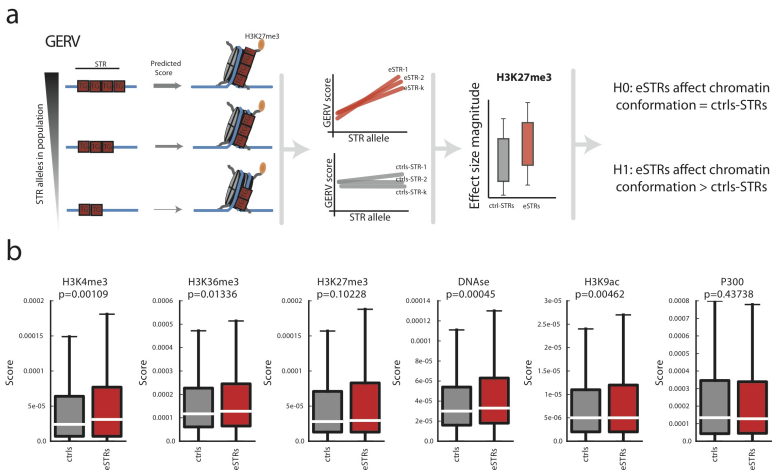
Partitioning variance when treaing the STR as a random effect. The heatmap shows the distribution of h^2_{STR} and h^2_b for each gene. Dashed gray lines give the medians of each distribution.

4.8.10 Supplementary Figure 10



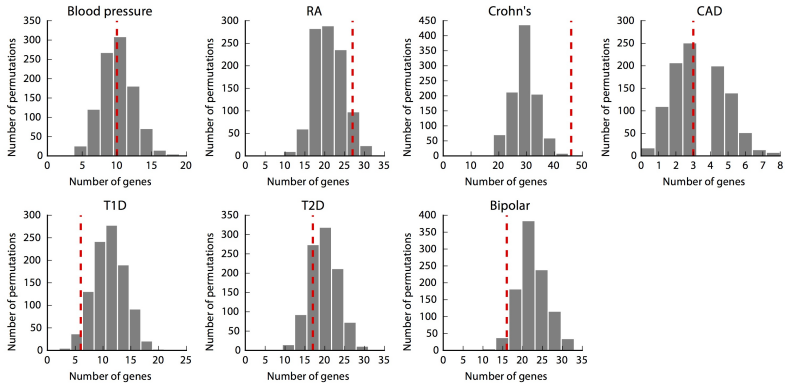
Enrichment of eSTRs at promoters and enhancers. For each distance bin around (**a**) the TSS (**b**) center of H3K27ac peaks, the plot shows the percentage of STRs that were analyzed in that bin that were called as significant eSTRs. **c** and **d** show the number of STRs in each distance bin. Black lines show the number of STRs that were included in our analysis (meaning they showed sufficient variability and are near genes). Red lines show the number of all STRs in the genome in each bin. Black lines were smoothed by averaging sliding windows of 3 consecutive data points. **a** and **b** were binned by 10kb. **c** and **d** were binned by 500bp.

4.8.11 Supplementary Figure 11



STRs modulate epigenetic signatures. **a**. Schematic of the application of GERV to predict histone modification signatures for different STR alleles. For each eSTR (red) and control STR (gray) we measured the magnitude of the slope between the STR allele and the GERV score and then tested whether the magnitudes were significantly different between the two sets. **b**. Comparison of the distribution of slope magnitudes for eSTRs (red) and controls (gray).

4.8.12 Supplementary Figure 12



Enrichment of eSTR genes in GWAS. Number of eSTR genes (red dashed line) overlapping GWAS genes for each trait. Gray bars give the distribution of the number of overlapping genes from 1000 control sets of STRs matched on expression in LCLs and on *cis* heritability. (RA=rheumatoid arthritis, CAD=coronary artery disease, T1D=type I diabetes, T2D=type 2 diabetes).

4.9 Supplementary Tables

4.9.1 Supplementary Table 1

Period	Num. eSTRs	% eSTRs	Num. all STRs	% all STRs	Enrichment	Pval
2	951	50.2%	50,184	62.0%	0.81	1.0
3	223	11.8%	7,369	9.1%	1.29	4.8×10^{-5}
4	516	27.2%	17,938	22.2%	1.23	8.2×10^{-8}
5	166	8.8%	4,466	5.5%	1.59	3.9×10^{-9}
6	39	2.1%	1,023	1.3%	1.63	2.4×10^{-3}

Distribution of motif lengths in eSTRs vs. all STRs. Distribution of motif lengths in all unique eSTR loci vs. all unique STR loci included in the analysis after applying quality filters.

4.9.2 Supplementary Table 2

Motif	Num. eSTRs	% eSTRs	Num. all STRs	% all STRs	Enrichment	Pval
AAAAAC	17	0.9%	217	0.3%	3.35	1.7×10^{-5}
AATC	10	0.5%	152	0.2%	2.81	3.2×10^{-3}
AAAAC	94	5.0%	1,822	2.2%	2.20	1.8×10^{-12}
AAC	95	5.0%	2,056	2.5%	1.97	5.0×10^{-10}
AAAC	173	9.1%	3,995	4.9%	1.85	9.0×10^{-15}
AAAG	47	2.4%	1,179	1.5%	1.70	3.6×10^{-4}
AAG	10	0.5%	285	0.4%	1.50	0.13
AAAAG	15	0.8%	449	0.6%	1.43	0.11
ATCC	16	0.8%	488	0.6%	1.40	0.11
ATC	10	0.5%	392	0.5%	1.09	0.44
AG	128	6.8%	5,174	6.3%	1.06	0.27
AAAT	198	10.4%	8,073	10.0%	1.05	0.25
AAAAT	35	1.8%	1,451	1.8%	1.03	0.45
AATG	16	0.8%	676	0.8%	1.01	0.52
AAT	74	3.9%	3,678	4.5%	0.86	0.92
AT	161	8.5%	8,775	10.8%	0.78	0.99
AC	662	34.9%	36,206	44.7%	0.78	1.0
AGAT	16	0.8%	1,561	1.9%	0.44	1.0

Distribution of motifs in eSTRs vs. all STRs. Distribution of motifs in all unique eSTR loci vs. all unique STR loci included in the analysis after applying quality filters. Only motifs for which there were at least 10 eSTRs are shown. Motifs were converted to canonical format as described in [?].

4.9.3 Supplementary Table 3

Annotation	Num. eSTRs	% eSTRs	Num. all STRs	% all STRs	Enrichment	Pval
Coding	13	0.7%	157	0.2%	3.54	9.1×10^{-5}
5' UTR	51	2.7%	897	1.1%	2.43	1.0×10^{-8}
Exon	127	6.7%	2,452	3.0%	2.21	1.5×10^{-16}
3' UTR	77	4.1%	1,569	1.9%	2.10	1.7×10^{-9}
Neargene (5')	335	17.7%	7,357	9.1%	1.95	1.5×10^{-32}
Neargene (3')	326	17.2%	7,399	9.1%	1.88	4.5×10^{-29}
Intron	1,314	69.3%	52,326	64.6%	1.07	6.1×10^{-6}
Intergenic	395	20.8%	23,373	28.9%	0.72	1.00

Distribution of genomic locations of eSTRs vs. all STRs. Annotations were compiled using Ensembl version 71. "Exon" refers to both coding and non-coding exons and untranslated regions. "Neargene" refers to regions within 5kb of a gene. "Intergenic" refers to STRs not falling into any other annotation. Note some STRs may overlap multiple annotations.

4.9.4 Supplementary Table 4

Element	Num. eSTRs	% positive effects	P-val
All eSTRs	2,060	49.5	0.708
POL24H8	75	65.3	0.011
DNAseI	41	70.7	0.012
5utr	59	64.4	0.036
Heterochrom/lo	801	47.3	0.138
NFIC	51	60.8	0.161
POL2	57	59.6	0.185
WeakPromoter	61	59.0	0.200
ELF1	36	61.1	0.243
WeakEnhancer	151	45.0	0.255
YY1	38	60.5	0.256
TCF12	29	62.1	0.265
WeakTxn	496	52.4	0.302
PAX5C20	35	60.0	0.311
ATF2	37	59.5	0.324
CREB1	26	61.5	0.327
TxnTransition	41	58.5	0.349
RUNX3	91	54.9	0.402
TxnElongation	323	47.7	0.436
intron	1,434	49.2	0.579
BCLAF	48	54.2	0.665
exon	140	52.1	0.673
neargene3	363	49.0	0.753
ActivePromoter	53	52.8	0.784
Repressed	53	52.8	0.784
neargene5	382	49.2	0.798
MTA3	28	53.6	0.851
NFATC1	32	53.1	0.860
FOXM1	38	52.6	0.871
intergenic	422	49.5	0.884
3utr	84	50.0	1.000
BCL3	37	51.4	1.000
StrongEnhancer	79	50.6	1.000
TCF3	37	48.6	1.000

Direction of effects of eSTRs. Genomic elements were annotated using Ensembl version 71 as described in the previous table. DNAseI HS sites and transcription factor binding sites were

annotated by ENCODE and were downloaded from the UCSC Table Browser for hg19. P-values are from a two-tailed binomial test for whether the percentage of positive slopes is significantly different than 50%.

4.9.5 Supplementary Table 5

Candidate gene studies					Genome-wide analysis				
Reference[1]	**Gene**	**STR location**	**Tissue**	**Direction of effect**[2]	**Position**[3]	**Effect size**	**# of samples**	**p-value**	**eSTRs**
Contente, 2002	*PIG3*	5'UTR	H1299 (non-small cell lung cancer)	Inducing	chr2: 24307212	0.37	94	0.00001	Yes
Shimajiri, 1999	*MMP9*	Promoter	TE9 (esophageal squamous cell carcinoma)	Inducing	chr20: 44637413	-0.10	272	0.06	No
Gebhardt, 1999	*EGFR*	Intron 1	*In vitro*	Repressing	chr7: 55088254	-0.13	199	0.07	No

Comparison of three candidate studies with STRs and their corresponding results in our genome-wide scan

[1] Full references are given in the main text.

[2] Inducing/repressing: length increase of the STR is associated with increase/decrease of expression.

[3] Start coordinate of the STR in hg19.

4.9.6 Supplementary Table 6

	h_b^2	h_{STR}^2	h_{STR}^2/h_{cis}^2
eSTR genes (n=1,928)	0.1203 (0.1139-0.1259)	0.0180 (0.0166-0.0199)	0.1230 (0.1106-0.1420)
Moderate cis h^2 (n=6,272)	0.0910 (0.0884-0.0938)	0.0145 (0.0137-0.0151)	0.1283 (0.1222-0.1346)
Moderate cis h^2, no eSTR (n=4,412)	0.0809 (0.0791-0.0829)	0.0136 (0.0129-0.0144)	0.1325 (0.1262-0.1397)

Heritability of gene expression explained by STRs vs. common bi-allelic variants. Values show the median and 95% confidence interval of the median across all eSTR-containing genes and genes with moderate cis heritability (âĽě5%). h_b^2 denotes the variance explained by all common cis bi-allelic variants, h_{STR}^2 denotes the variance explained by the lead STR for each gene, and $h_{cis}^2 = h_{STR}^2 + h_b^2$.

4.9.7 Supplementary Table 7

	h^2_b	h^2_{STR}	h^2_{STR}/h^2_{cis}
eSTR genes - (STR fixed)	0.1203 (0.1139-0.1259)	0.0180 (0.0166-0.0199)	0.1230 (0.1106-0.1420)
eSTR genes - (STR random)	0.1229 (0.1159-0.1295)	0.0200 (0.0178-0.0216)	0.1288 (0.1179-0.1451)
Moderate *cis* h^2_{cis} (STR fixed)	0.0910 (0.0884-0.0938)	0.0145 (0.0137-0.0151)	0.1283 (0.1222-0.1346)
Moderate *cis* h^2_{cis} (STR random)	0.0892 (0.0865-0.0918)	0.0143 (0.0137-0.0149)	0.1245 (0.1184-0.1309)

Heritability of gene expression explained by STRs vs. common bi-allelic variants in a random effects model. Values show the median and 95% confidence interval of the median across all eSTR-containing genes and genes with moderate *cis* heritability ($\geq$5%). h^2_b denotes the variance explained by all common *cis* bi-allelic markers, h^2_{STR} denotes the variance explained by the lead STR for each gene, and $h^2_{cis} = h^2_{STR} + h^2_b$.

4.9.8 Supplementary Table 8

Annotation	STR enrichment	STR p-value	SNP enrichment	SNP p-value
H3k27ac	1.18	0.001	1.91	<0.001
H3k27me3	0.87	1.000	0.81	1.00
H3k36me3	1.11	<0.001	1.20	<0.001
H3k4me1	1.18	<0.001	1.48	<0.001
H3k4me2	1.25	<0.001	1.93	<0.001
H3k4me3	1.26	<0.001	2.03	<0.001
H3k9ac	1.17	0.009	2.07	<0.001
H3k9me3	0.97	0.804	1.11	0.001
ActivePromoter	1.00	0.513	3.41	<0.001
Heterochrom/lo	0.91	1.000	0.68	1.000
Insulator	1.23	0.221	0.74	0.940
PoisedPromoter	0.56	0.899	3.14	<0.001
Repressed	0.69	0.997	0.65	1.000
StrongEnhancer	0.98	0.603	1.93	<0.001
TxnElongation	1.08	0.072	1.03	0.191
TxnTransition	1.07	0.370	1.09	0.220
WeakEnhancer	1.23	0.004	1.35	<0.001
WeakPromoter	1.48	0.002	2.02	<0.001
WeakTxn	1.09	0.012	1.04	0.057

Enrichment of eSNPs and eSTRs. Overlap of eSTRs and eSNPs with each annotation were compared to the overlap of shifted eSTR and eSNP locations. We performed 1,000 rounds of shifting eSTRs and eSNPs to generate null distributions of the percent overlap. Enrichment values give the percent of eSTRs or eSNPs overlapping each annotation divided by the average percent overlap after shifting. P-values are empirical probabilities based on comparison to the 1,000 shifted sets of locations.

4.9.9 Supplementary Table 9

Chr	Pos	Eff. size	R^2	p-val	q-score	Phenotype	eSTR gene ID	Name	Samples
chr4	9955416	0.026	0.029	3.49E-08	1.09E-05	Uric Acid	ENSG00000109667	SLC2A9	1047
chr10	27124545	-0.034	0.022	4.61E-07	1.69E-04	Phosphate	ENSG00000136754	ABI1	1139
chr1	109393265	0.151	0.012	2.89E-05	1.44E-02	Diastolic BP	ENSG00000121940	CLCC1	1475
chr6	20195837	-0.025	0.012	3.26E-05	1.62E-02	Albumin	ENSG00000172197	MBOAT1	1430
chr1	110516300	-0.052	0.012	5.07E-05	2.49E-02	Urea	ENSG00000143093	STRIP1	1372
chr15	100382014	-0.063	0.014	8.46E-05	3.73E-02	CRP	ENSG00000259363	CTD-2054N24.2	1110
chr3	58429246	-0.021	0.009	2.42E-04	5.32E-02	Albumin	ENSG00000168291	PDH8	1461
chr17	80787868	0.029	0.009	3.22E-04	5.32E-02	Albumin	ENSG00000141560	FN3KRP	1394
chr9	33502041	-0.033	0.012	1.27E-04	6.14E-02	FVC	ENSG00000165271	NOL6	1248
chr2	85624828	0.009	0.012	4.21E-04	6.57E-02	Uric Acid	ENSG00000042493	CAPG	1059
chr3	37141930	0.05	0.009	3.58E-04	8.69E-02	FVC	ENSG00000093167	LRRFIP2	1367
chr3	129174742	-0.034	0.011	1.93E-04	8.86E-02	MCV	ENSG00000172771	EFCAB12	1232

Significant associations (FDR<0.1) of eSTRs in the TwinsUK data. We considered only eSTRs for which a joint model with the lead eSNP significantly improved the explained variance of the expressed gene over a model with the lead eSNP alone. Positive effects denote STRs whose expansions are associated with increased phenotypic levels and vice versa. R^2 denotes the phenotypic variane explained by the STR. BP: blood pressure; CRP: C-reactive protein; MCV: Mean Corpuscular Haemoglobin. eSTR gene ids denote the genes whose expression levels were found to be associated with the eSTR. Samples denote the number of TwinsUK genomes that were genotyped and phenotyped for the specific eSTR/clinical phenotype association.

Chapter 5

Conclusion and future directions

Existing tools for genotyping STRs from sequencing data

When we began this work, no dedicated tool for genotyping STRs from sequencing data existed. McIver *et al.* [?] evaluated STR variation in the 1000 Genomes Project [?] samples. However their results were limited to the short variations that could be captured by aligners with poor indel sensitivity, and therefore vastly underestimated polymorphism levels. A major contribution of our work was develop the first efficient algorithm for generating accurate STR genotypes, called lobSTR [?] and described in chapter 2. Over the last several years, additional tools have arisen:

1. **STRViper** [?]: Uses insert size between paired end sequencing reads to detect STR variations.
2. **RepeatSeq** [?]: uses Bayesian model selection to genotype previously aligned STR-containing reads.
3. **STR-FM** [?]: uses a method based on lobSTR's algorithm but with a modified detection step for increased sensitivity of short repeats and may be applied to non-diploid samples.

The first two tools operate on previously existing alignments, and so are limited by the quality of the upstream aligner. The third may provide improvements to STR calling, especially at homopolymers which are extremely noisy in sequencing data. So far, these tools have not seen widepsread use in mainstream sequence analysis.

In addition to STR-specific callers, a new class of variant callers, including GATK [?] HaplotypeCaller, Platypus [?], and Scalpel [?], perform local reassembly of diploid haplotypes. These methods can theoretically genotype STRs quite accurately, albeit with greatly increased computational costs. There has so far been no systematic evaluation of their performance at STRs, but these tools may be promising for STR analysis in the future.

5.0.10 Long-read technology can capture long repetitive regions

A major limitation of analyzing STRs from high throughput sequencing data is the short read length. Only reads entirely spanning an STR are informative of the repeat length, and sufficient flanking region on either side of the STR is required for accurate alignment. While the mainstream sequencing technology from Illumina is limited to sequencing at most several hundred base pairs in a single read, alternative sequencing platforms, such as PacBio's SMRT (single molecule real time) sequencing [**?**] and the new Nanopore technology [**?**], can now produce much longer read lengths of up to several thousand base pairs.

These technologies may be used to sequence long repeats observed in expansion disorders such as Fragile X [**?**] or ataxias [**?**], and other complex regions of the genome. Chaisson *et al* [**?**] recently applied SMRT to sequence a haploid genome to 40x coverage with an average read length of 5kb. With these long reads, they were able to close 50 gaps in the human genome reference assembly, which were highly enriched for short tandem repeats and other repeats embedded in larger, more complex tandem arrays of degenerate repeats. With longer and longer read lengths, we may soon be able to analyze STRs, as well as other repetitive elements such as variable number tandem repeats (VNTRs) and retrotransposons that were previously inaccessible using sequencing studies.

Appendix A

PyBamView: a browser based application for viewing short read alignments.

Most of this chapter was first published as:

> **Gymrek M**. PyBamView: a browser based application for viewing short read alignments. *Bioinformatics*. (2014).

Abstract

Summary: Current sequence alignment browsers allow visualization of large and complex next-generation sequencing datasets. However, most of these tools provide inadequate display of insertions and can be cumbersome to use on large datasets. I implemented PyBamView, a lightweight web application for visualizing short read alignments. It provides an easy-to-use web interface for viewing alignments across multiple samples, with a focus on accurate visualization of insertions.

Availability and Implementation: PyBamView is available as a standard python package. The source code is freely available under the MIT license at https://mgymrek.github.io/pybamview.

A.1 Introduction

The rapid growth of next-generation sequencing (NGS) technologies has led to a wide variety of short read DNA datasets. Manual inspection of sequence alignments is an important aspect of quality control. While the majority of NGS analyses have focused on single nucleotide polymorphisms (SNPs), recent bioinformatics advances allow analysis of more complicated vari-

ants, such as small insertions or deletions [?], larger structural variants [?], and short tandem repeats (STRs) [?, ?]. Furthermore, widely used genome engineering techniques, such as the CRISPR-Cas9 system [?] can often produce a wide range of complex variants. In these cases, visualization of insertion and deletion events is a particularly critical analysis step.

Current genome browsers, such as UCSC [?], and IGV [?], offer visualization of alignments from BAM files across multiple samples and integration of many layers of genomics datasets. However, most existing tools have two important limitations. First, most are based on alignments to an ungapped reference sequence, which provides inadequate visualization of insertions. The BAM specification supports a padded reference, which captures multiple sequence alignment information and results in accurate insertion display by most browsers. However, most BAM files consist of pairwise alignments of short reads to a reference and do not use this feature. As a result, insertions are represented by an icon such as a vertical bar, which does not provide any visual information about the size or sequence of the inserted nucleotides. Second, the majority of alignment browsers are cumbersome to use, especially to visualize the large datasets typical of NGS experiments. They either require that the user upload large data files to a remote server or involve complicated installation and large resource requirements to run locally.

Several alignment browsers, such as Bambino [?], Consed [?], and the text-based SAMtools [?] tview, overcome these limitations: they display the sequence of insertions even when using the standard ungapped reference, and are run locally with relatively low system requirements. However, tview does not allow the user to view multiple BAM files at once, and none allow for exporting alignments as snapshots or for sharing alignments remotely through a web browser.

Here I present PyBamView, a lightweight web application for viewing alignments from BAM files. PyBamView provides alignment visualizations that accurately represent SNP, insertion, and deletion events that can easily be exported to create publication-ready figures. It runs locally from the command line with minimal resource requirements and displays alignments in a web browser. This interface allows users to quickly view alignments locally and to easily share alignments with local or remote collaborators.

A.2 Basic Usage and Features

PyBamView is a Python-based web application that is run from the command line. Users provide PyBamView with a directory containing indexed BAM files and an optional reference genome in

fasta format:

```
pybamview --bamdir DIRECTORY/WITH/BAMS --ref REF.fa
```

PyBamView will start a small webserver that can be accessed locally in a web browser. Optional arguments can serve the application over a different address for sharing the URL with remote collaborators or as a public resource. For instance, adding the options `--ip 0.0.0.0 --port 5000` will serve PyBamView over port 5000 via http. The **Supplemental Text** A.7 and program website contain a complete description of this feature.

The web browser displays a list of all samples contained in the BAM files provided. Users can select one or more samples to open in the genome-browser view. This consists of a reference track, followed by collapsible alignment tracks containing reads for each sample. While there is theoretically no limit to the number of samples analyzed, PyBamView can reasonably display five low to moderate coverage samples at once.

Users can navigate to the genomic region of interest by entering the genomic coordinate into the search bar (e.g. chr1:10000). In the default view, base pair differences from the reference genome are highlighted, allowing easy identification of SNPs and potential sequencing errors (**Figure A.2A**). A deleted base pair is indicated by a "." in the alignment, and an insertion as a "*" in the reference sequence (**Figure A.2B**). This allows easy visualization of the sequence and size of inserted bases, which is not currently possible with most alignment browsers (**Supplemental Fig. A.8.1**). Users can zoom out up to 100x to easily visualize large insertions or deletions spanning hundreds or thousands of bases. Additional features are described in the **Supplemental Text** A.7.

A.3 Example use cases

Alignment visualization is a critical step of any sequencing experiment. Here I show three examples where PyBamView provides useful visualization of sequence variants. Use cases are not limited to these examples and can theoretically include any "-seq" experiment that can be represented by a BAM file.

First, it provides accurate visualizations of different length insertions, such as different alleles of a tandem repeat (**Supplemental Fig. A.8.2A**). Furthermore, zooming out allows for visualization of large repeat expansions, such as a 60bp CAG expansion in Huntington's Disease

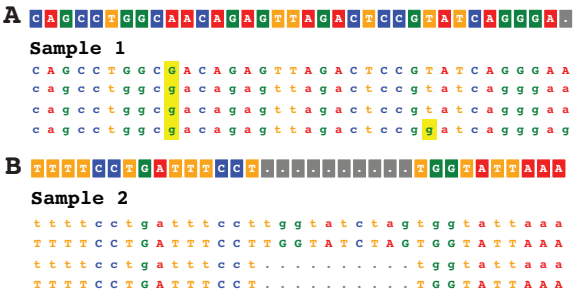


Figure A-1: **PyBamView display of sequence variants.** In each figure, the reference sequence is shown at the top followed by alignments of each read. Alignments were generated by PyBamView's PDF export feature. **(a)** Mismatches from the reference sequence, due to either SNPs or sequencing errors, are shown as highlighted bases. **(b)** Insertions are shown as gaps in the reference, which allows the length and sequence of the insertion to be easily visualized.

(**Supplemental Fig.** A.8.2**B**, simulated 250bp reads).

Second, it can be used to analyze variation across samples. This is useful in such analyses as comparing matched tumor vs. normal samples or looking for mutations in affected vs. non-affected individuals in disease genetic studies. (**Supplemental Fig.** A.8.3) shows example comparisons of individuals at a SNP, small insertion, and a large deletion spanning several kb.

Third, it can visualize complex mutations generated by genome engineering technologies such as CRISPR-Cas9 [?]. Dissecting these mutations requires adequate visualization of indels. An example alignment from a CRISPR library is shown in (**Supplemental Fig.** A.8.4).

A.4 Implementation

PyBamView is implemented as a Python-based web application using the Flask library. Alignments are processed using a Python backend, which then generates HTML, CSS, and JavaScript files that are displayed in the web browser.

PyBamView takes advantage of BAM and fasta indexing to avoid loading large files into memory. It uses the pysam and pyfasta libraries for parsing BAM and fasta files, respectively. Both libraries use efficient index data structures, which allow them to quickly fetch data from specific genomic regions of interest. Read alignments are parsed from the CIGAR scores in the BAM file and are displayed as SVG elements using Javascript. All CIGAR options reported in the

SAM specification, including the padded reference option, are supported (**Supplemental Fig. A.8.5**).

A.5 Conclusion

As the use of next generation sequencing to analyze complex genomic events grows, there is a critical need for accurate and easy-to-use visualization tools. PyBamView provides a simple yet powerful interface for alignment visualization that facilitates collaborative data analysis.

A.6 Acknowledgement

The author would like to acknowledge members of the Erlich lab, Alon Goren, and Roy Ronen for helpful feedback, and Assaf Gordon for valuable programming guidance.

This work was supported by a National Defense Science and Engineering Graduate Fellowship [32 CFR 168a].

A.7 Supplemental Text

A.7.1 Additional Features

PyBamView has several features beyond the basic alignment viewer that facilitate specific use cases.

Pre-defined lists of loci

A sequencing experiment often targets a specific set of loci. For instance, analysis of STRs may focus on a panel of markers, such as the CODIS set used in forensics or Y-STRs used for genetic genealogy. PyBamView allows the user to provide a bed file containing the coordinates and names of a set of defined loci. The browser view will then show a drop-down box, from which users can easily navigate to loci by name, rather than by genomic coordinate.

Exporting alignment figures for publication

Most alignment viewers provide attractive data visualizations, but alignment screenshots are not easily converted to publication-quality figures. PyBamView provides an "Export snapshot" option that generates a PDF image of the current alignment. The PDF file can then be easily manipulated by third-party programs to prepare publication-ready figures.

Serving PyBamView remotely

PyBamView uses the Python Flask library to launch a web application. Therefore, PyBamView can be served at a URL that is visible remotely to share data with collaborators or as a public resource. To allow viewing PyBamView from a remote computer, add the following arguments:

- `--ip 0.0.0.0`: This will allow PyBamView to be accessed on a computer other than the local computer you are running it on.
- `--port $PORT`: where $PORT is some number between 1024-65535. This must be a port that allows inbound http traffic.

For example, if you run:

```
pybamview --bamdir DIRECTORY/WITH/BAMS --ref REF.fa --ip 0.0.0.0 --port 5000
```

Then you can navigate to the URL: http://$MY_SERVER:5000

where $MY_SERVER is either the IP address or hostname of the server from which you are running PyBamView. Detailed usage instructions for serving PyBamView remotely, including how to serve PyBamView using Amazon Web Services, can be found on the usage page of the website (http://mgymrek.github.io/pybamview/usage.html).

A.7.2 Datasets for example use cases

Sequencing data for the two samples shown in (**Supplemental Fig.** A.8.2) are from whole genome sequencing of samples HGDP00521 and HGDP00533 available from the Simon's Foundation website at:
http://www.simonsfoundation.org/life-sciences/simons-human-diversity-project/.

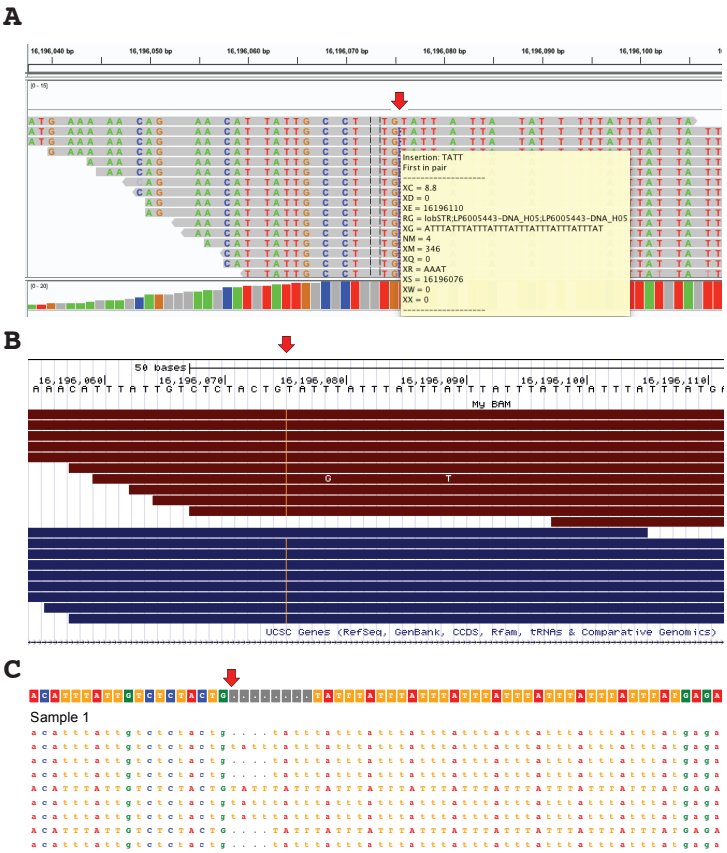
Sequencing data and variant calls for the samples shown in (**Supplemental Fig.** A.8.3) were downloaded from the 1000 Genomes Project website (ftp://ftp-trace.ncbi.nih.gov/

1000genomes/ftp/).

Example CRISPR sequencing data was taken from Hsu et al. [**?**], available as SRA experiment SRP923129 (http://www.ncbi.nlm.nih.gov/bioproject/?term=SRP023129), run SRR867203. Reads were aligned to hg19 chr2 using BWA [**?**] and down-sampled 10,000 fold using Picard's DownsampleSam tool (http://picard.sourceforge.net/command-line-overview.shtml#DownsampleSam), and were enriched for reads containing evidence of mutations for visualization purposes.

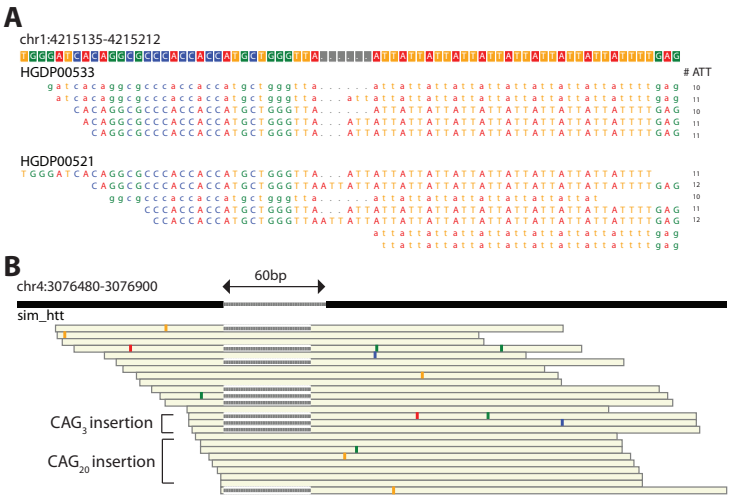
A.8 Supplemental Figures

A.8.1 Supplemental Figure 1



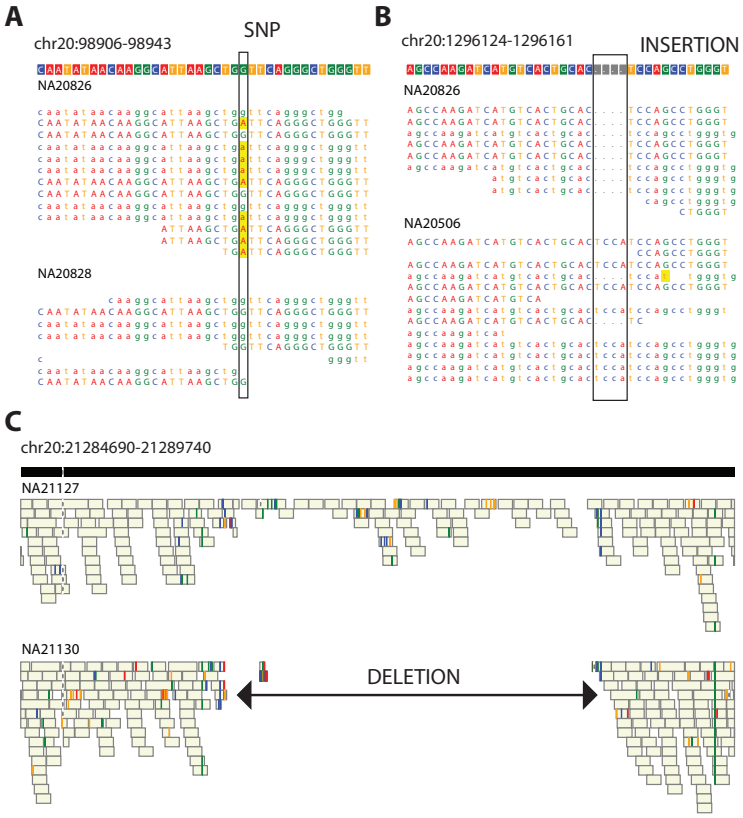
Insertion display on current browsers compared to PyBamView. The alignment shows an individual heterozygous for two alleles with insertions of different lengths from the reference sequence (shown under the red arrows). **(A)** The Integrative Genomics Viewer (IGV) displays insertions as vertical bars. The user must hover over each variant to see the sequence and length of the insertion. **(B)** Similarly, the UCSC genome browser displays insertions using an orange line, which does not distinguish between different insertion alleles. **(C)** PyBamView displays the entire sequence of each insertion, which allows the user to easily visualize the two alleles.

A.8.2 Supplemental Figure 2



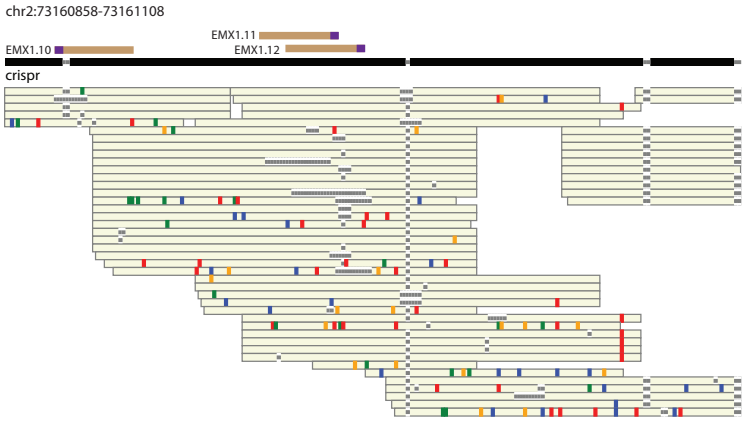
Accurate display of short tandem repeat alleles. (A) Alignments show two distinct short tandem repeat alleles (ATT11 and ATT12) that are longer than the reference allele (ATT10). The allele supported by each read is annotated to the right of the alignment. **(B)** The padded reference distinguishes between a pathogenic expansion of 60bp vs. a non-pathogenic expansion of 9bp at the CAG trinucleotide repeat implicated in Huntington's Disease. Gaps in the reference sequence and in each read are represented by gray bars. Mismatches from the reference are colored according to the nucleotide of the mismatched base. Example reads supporting each allele are annotated. Data are from simulated 250bp reads with a 0.2% sequencing error rate.

A.8.3 Supplemental Figure 3



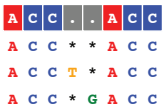
Comparing variants between 1000 Genomes Project individuals. PyBamView allows for visualization of a wide range of variant types across samples. **(A)** A SNP that is heterozygous in sample NA20826 and homozygous reference in NA20828 is highlighted in yellow. **(B)** A small insertion that is homozygous reference in NA20826 and heterozygous in NA20506. **(C)** Zooming out allows visualization of a homozygous large deletion spanning several kb in NA21130 that is heterozygous in NA21127.

A.8.4 Supplemental Figure 4



Analyzing mutations from genome engineering experiments. Example sequencing data was obtained from a CRISPR experiment by Hsu et al. [?] targeting the gene *EMX1*. Reads were down-sampled and enriched for reads containing mutations. Brown bars show the location of target sites, and purple bars show the location of PAM sequences.

A.8.5 Supplemental Figure 5

CIGAR OP.	Example seq./CIGAR	PyBamView
M (aln. match)	TGGCCCCT *8M*	
I (insertion)	TTCGGTCT *3M2I3M*	
D (deletion)	CGCCGC *5M2D1M*	
N (skipped)	TTCCAT *5M2N1M*	
S (soft clip)	CGCAGATGGCG *3S8M*	
H (hard clip)	TCCCCTTC *3H8M*	
= (seq. match)	ACGGCTTG *8=*	
X (seq. mismatch)	TACCACGG *6=2X*	
P (padding)	ACCACC *3M2P3M* ACCTACC *3M1I1P3M* ACCGACC *3M1P1I3M*	

PyBamView display of each CIGAR operation. The first column gives each CIGAR operation supported by the SAM specification. The second column gives an example sequence and corresponding CIGAR score. The third column shows how the reference (top) and read (bottom) sequences are displayed by PyBamView.

Appendix B

Identifying personal genomes by surname inference

Most of this chapter was first published as:

> **Gymrek M**, McGuire A, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science*. (2013).

Abstract: Sharing sequencing datasets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y-chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for US males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

B.1 Main Text

Surnames are paternally-inherited in most human societies, resulting in their co-segregation with Y-chromosome haplotypes [**?**, **?**, **?**, **?**, **?**]. Based on this observation, multiple genetic genealogy companies offer services to reunite distant patrilineal relatives by genotyping a few dozen highly polymorphic short tandem repeats across the Y-chromosome (Y-STRs). The association between surnames and haplotypes can be confounded by non-paternity events, mutations, and adoption of the same surname by multiple founders [**?**]. The genetic genealogy community addresses these barriers with massive databases that list the test results of Y-STR haplotypes along

with their corresponding surnames. Currently, there are at least eight databases and numerous surname project websites that collectively contain hundreds of thousands surname-haplotype records (**Supplementary Table B.5.1**).

The ability of genetic genealogy databases to breach anonymity has been demonstrated in the past. In a number of public cases, male adoptees and descendants of anonymous sperm donors used recreational genetic genealogy services to genotype their Y-chromosome haplotypes and to search the companies' databases [**?**, **?**, **?**, **?**]. The genetic matches identified distant patrilineal relatives and pointed to the potential surnames of their biological fathers. By combining other pieces of demographic information, such as date and place of birth, they fully exposed the identity of their biological fathers. Lunshof et al [**?**] was the first to speculate that this technique could expose the full identity of participants in sequencing projects. Gitschier [**?**] empirically approached this hypothesis by testing 30 Y-STR haplotypes of CEU participants in these databases and reported that potential surnames can be detected. [CEU participants are multigenerational familes of northern and western European ancestry in Utah who had originally had their samples collected by CEPH (Centre d'Etude du Polymorphisme Humain) and were later reconsented to participate in the HapMap project.] However, these surnames could match thousands of individuals and full re-identification in a single person resolution was not pursued.

Our goal was to quantitatively approach the question of how readily surname inference might be possible in a more general population, apply this approach to personal genome datasets, and demonstrate end-to-end identification of individuals using only public information. We show that full identities of personal genomes can be exposed via surname inference from recreational genetic genealogy databases followed by Internet searches. In all cases in which individuals were studied who had donated sequences, the informed consent statements they had signed stated privacy breach as a potential risk and the data usage terms did not prevent re-identification. Representatives of relevant organizations that funded the original studies were notified and confirmed the compliance of this study with their guidelines.

As a primary resource for surname inference, we focused on Ysearch (www.ysearch.org) and SMGF (www.smgf.org), the two largest public genetic genealogy databases with free-of-charge, built-in search engines. The interfaces of these engines are quite similar and allow users to insert a combination of Y-STR alleles and search for matching records based on genetic similarity. The retrieved records contain surnames typically with information about the patrilineal line, such as geographical locations, potential spelling variants, and pedigrees. In total, these databases contain 39,000 unique surname entries from approximately 135,000 records. The distribution

of records per surname is significantly correlated ($R^2 = 0.78$, $p < 1.20 \times 10^{-6}$) with surname frequencies in the US, suggesting an overall good representation of this population (**Fig. B.1A**).

To test the probability of surname inference, we challenged the two databases with an orthogonal cohort of Y-STR haplotypes consist of 34 markers (**Supplementary Table B.5.2**) from 911 individuals, primarily with Caucasian ancestry, whose surnames are known (**Supplementary Table B.5.3**). This cohort was compiled from YBase, a distinct genetic genealogy database and contains individuals with 521 surnames that segregate in the US population. In each haplotype query, our surname recovery algorithm began by retrieving the database record with the shortest Time to Most Recent Common Ancestor (TMRCA) of the input haplotype (**Supplementary Figure B.4.1**, **Supplementary Table B.5.4**). Then, it calculated a confidence score that the surname match of the retrieved record is significantly better than other matches. If the score passed a user-defined threshold, the algorithm assigned the record's surname to the input haplotype; otherwise, it categorized it as "unknown". We tested the algorithm with a range of confidence thresholds to explore the trade-off between successful versus wrong recovery of surnames. Finally, we weighted the results using a stratified sampling approach to reflect the frequency of surnames in the US population (**Supplementary Material B.3**).

Our analysis projects a success rate of approximately 12% (s.d. 2%) in recovering surnames of US Caucasian males (**Fig. B.1B**, **Supplementary Figure B.4.2**). This rate can be accomplished with a conservative threshold that would return a wrong surname in 5% of cases and label 83% of cases as unknown. Higher success rates of up to 18% can be achieved at the price of increased probability to recover an incorrect surname. Since our input cohort is based on individuals that were tested using genetic genealogy services, our results are presumably mostly relevant to socio-economic groups with high participation in these services, namely upper and middle class US Caucasians.

Combining the recovered surname with additional demographic data can narrow down the identity of the sample originator to just a handful of individuals. The analysis above indicated that most recovered surnames are quite rare with frequencies of less than 1:4,000 of the US population, corresponding to <40,000 males (**Figure B.1C**, **Supplementary Figure B.4.3**) (**Supplementary Material B.3**). We considered a scenario in which the genomic data is available with the target's year of birth and state of residency, two identifiers that are not protected by the United States Health Insurance Portability and Accountability Act (HIPAA). Searching individuals by year of birth, state, and surname combinations is supported by various online public record search engines, such as PeopleFinders.com or USA-people-search.com. Based on

extensive simulations with the US Census data, our results predict that year of birth and state alone are weak identifiers and searches based on their combination would match at least 60,000 US males in 50% of cases (**Figure B.1D**). However, when surname information is added to the search, the median list size shrinks to only 12 males, which are a few enough matches to investigate individually.

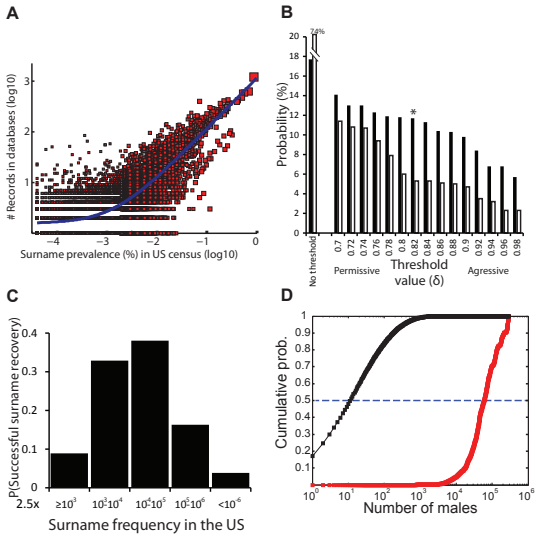


Figure B-1: **Quantitative assessment of identification via surname inference (A)** The number of Ysearch and SMGF records as a function of surname prevalence in the US population. The best fit line is shown in blue. **(B)** Expected performance of surname recovery. The probability of successful recovery (closed bars) and wrong recovery (open bars) are shown at different surname confidence thresholds. The star indicates the middle-range performance threshold that was described in the main text. **(C)** The expected distribution of recovered surnames as a function of their prevalence. Most recovered surnames are expected to have a frequency of 1:4,000 individuals or less. **(D)** The cumulative distribution function of US males with a profile that matches a specific age, state, and surname combination (black) compared to the distribution when only age and state are known (red). The median is labeled with a dashed line.

Next, we established the feasibility of Illumina sequencing to produce accurate Y-STR haplotypes. Using lobSTR, an algorithm for STR profiling from raw sequencing reads [**?**], we processed ten high coverage male genomes from the Human Genome Diversity Panel (HGDP). lobSTR produced Y-STR haplotypes with an average length of 53 out of the possible 79 genealogical markers (**Supplementary Table B.4.5**). Comparing these results to capillary electrophoresis

calls revealed 99% accuracy. We further found that even at lower sequencing coverage of 10x, informative haplotypes can be obtained by lobSTR (**Supplementary Figure B.4.4**). To test the ability to retrieve genetic genealogy records with the Illumina haplotypes, we profiled STRs from the genome of a US Caucasian male from our lab collection that was sequenced with Illumina 100bp reads to a coverage of 13x. In parallel, we submitted this sample to the genealogy service of Sorenson Genomics and created a Ysearch record based on their results. A search with the Illumina haplotype returned his Ysearch entry as a top record (**Supplementary Figure B.4.5**).

The NCBI archives host a small number of genomes from identified individuals, providing good test cases for identification via surname inference. We used lobSTR to extract Y-STR haplotypes from the genomes of John West [**?**], Michael Snyder [**?**], and Craig Venter [**?**] (**Supplementary Table B.4.6**). Searching Ysearch and SMGF with the Y-STR haplotypes of West and Snyder did not return their surnames and resulted in low matches to records with relatively ancient MRCAs 23-28 generations ago (**Supplementary Material B.3**). A search with Craig Venter's haplotype returned a clear match to a "Venter" record that was concordant at all 33 comparable markers and with an estimated TMRCA of less than 8 generations (**Figure B.1**). We further tested whether it would be feasible to trace back Craig Venter by combining his surname with demographic profiling. A query for "Surname: Venter, Year of Birth: 1946, State: California" in online public record search engines retrieved two matching records of males, one of whom was Craig Venter himself.

Surname inference from personal genomes puts the privacy of current de-identified public datasets at risk. We focused on the male genomes in the collection of Utah Residents with Northern and Western European Ancestry (CEU). The informed consent of these individuals did not definitively guarantee their privacy and stated that futuristic techniques might be able to identify them `http://hapmap.ncbi.nlm.nih.gov/downloads/elsi/CEPH_Reconsent_Form.pdf`. To test the ability to trace back the identities of these samples from personal genomes, we processed with lobSTR 32 Illumina genomes of CEU male founders that reside in public repositories of the 1000 Genomes Project [**?**] and the European Nucleotide Archive that were sequenced with read lengths of at least 76bp. Most of these genomes were sequenced to a shallow depth of less than 5x, and produced sparse Y-STR haplotypes. We selected the ten genomes that had the longest Y-STR haplotypes with a range of 34-68 markers to attempt surname recovery. Searching the genetic genealogy databases returned top-matching records with Mormon ancestry in 8 of the 10 individuals for which the top hit had at least 12 comparable markers. Moreover, for four individuals, the top match consisted of multiple records with the same surname, increasing the

confidence that the correct surname was retrieved. This potential high surname recovery rate stems from a combination of the deep interest in genetic genealogy among this population and the large family sizes, which exponentially increases the number of targeted individuals for every person that is tested.

Figure 2

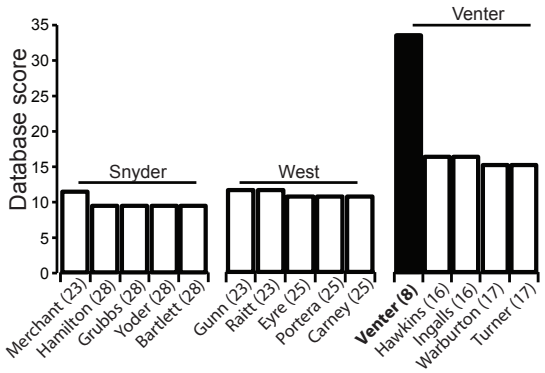


Figure B-2: **The top five records retrieved after searching Ysearch with the Y-STR haplotypes of Michael Snyder, John West, and Craig Venter.** The expected number of generations to the MRCA is given in parentheses for each record. Searching with Craig Venter returned a "Venter" record (closed bar) as the top match.

In five surname recovery cases, we fully identified the CEU individuals and their entire families with very high probabilities (**Table B.1**). These five cases belonged to three pedigrees, where in two of these pedigrees the surnames of both the paternal and maternal grandfathers were recovered. Our strategy for tracing back individuals relied on the recovered surnames as well as publicly available Internet resources such as record search engines, obituaries, and genealogical websites, and demographic metadata available in the Coriell Cell Repository website. The year of birth was inferred by subtracting the ages in Coriell from the year of collecting samples. Each search took 3 to 7 hours by a single person. The identified families matched exactly to the corresponding pedigree descriptions in the Coriell database: the number of children, the birth order of daughters and sons, and the state of residence were identical. All grandparents were alive in 1984, the year that the CEU cell line collection was established [**?**]. In the two cases of a dual surname recovery from both grandfathers, the surname of the father and the maiden name of the mother matched exactly to the grandfathers' surnames, substantially increasing the confidence of the recovery. Coriell also lists the ages during sample collection for these

two pedigrees, which agreed with the age differences of the identified family members. Using genealogical websites, we traced the patrilineal lineage that connects each identified genome through the MRCA to the record originator in the genetic genealogy database (**Fig. B.1**). This analysis revealed that two to seven meiosis events link the CEU genome to the record source. Finally, we calculated that the probability of finding random families in the Utah population with these exact demographic characteristics is less than 1 in 105-109 (**Supplementary Material B.3**). In total, surname inference breached the privacy of nearly 50 individuals from these three pedigrees.

This study shows that data release, even of a few markers, by one person can spread through deep genealogical ties and lead to the identification of another person who might have no acquaintance with the person who released his genetic data. The propagation of information through shared male lines amplifies the range of identification, allowing ∼135,000 records to potentially target several millions of US males. Another feature of this identification technique is that it entirely relies on free, publicly-available resources. It can be completed end-to-end with only computational tools and an Internet connection. The compatibility of our technique with public record search engines makes it much easier to continue identifying other datasets in the same pedigree, including female genomes, once one male target is identified. We envision that the risk of surname inference will grow in the future. Genetic genealogy enthusiasts add thousands of records to these databases every month. In addition, the advent of third-generation sequencing platforms with longer reads will enable even higher coverage of Y-STR markers, further strengthening the ability to link haplotypes and surnames.

Similar to other genetic privacy issues [?, ?, ?, ?, ?, ?, ?], preventing surname inference from public whole genome datasets might be quite challenging. Masking Y-STR markers could limit the effectiveness of the method presented in this study, but this approach is not sustainable (**Supplementary Material B.3**). Our analysis suggests that Y-STR haplotypes can be imputed back from SNPs on the Y-chromosome (Y-SNPs) when a large reference set of male genomes will be available (**Supplementary Figure B.4.6**). In addition, community efforts, such as the Y Chromosome Genome Comparison, have already started exploring the association between Y-SNPs and surnames (**Supplementary Table B.5.1**), and might allow bypassing Y-STR masking. We also posit that restricting genetic genealogy information is not practical as some of the data is already scattered in multiple end-user websites and genealogy mailing lists.

Existing policy tools, such as controlled access databases with data use agreements, may mediate the exposure of genomic information to surname inference. However, in our view, the appropriate

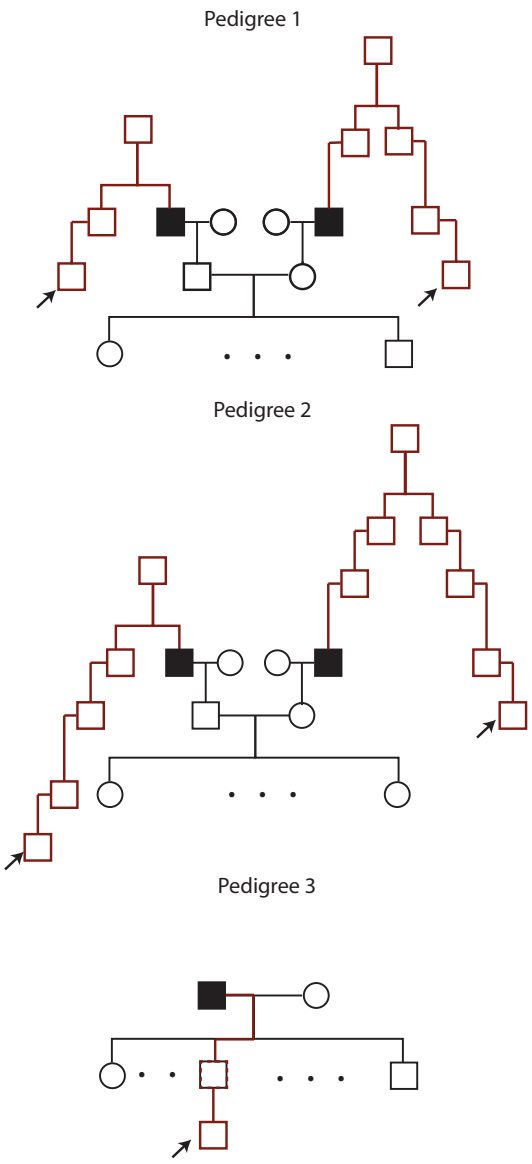


Figure B-3: **Illustrations of the three CEU pedigrees (black) showing how genetic information from distant patrilineal relatives (arrow; red - patrilineal lines) can identify individuals.** Filled rectangles represent sequenced individuals. To respect the privacy of these families, only abbreviated versions are presented. The sex of the CEU grandchildren was randomized. The numbers of grandchildren are not given.

response to genetic privacy challenges such as these is not for the public to stop donating samples or for data sharing to stop - which would be devastating reactions that could significantly hamper scientific progress. Rather, we believe that establishing clear policies for data sharing, educating participants about the benefits and risks of genetic studies [?] and the legislation of proper usage of genetic information are pivotal ingredients to support the genomic endeavor.

CEU Pedigree	Pedigree 1		Pedigree 2		Pedigree 3
Genome for surname recovery	Paternal grandfather	Maternal grandfather	Paternal grandfather	Maternal grandfather	Father
Y-STR source	Illumina WGS	Illumina WGS	Illumina WGS / Published Y-STR profiles	Illumina WGS / Published Y-STR profiles	Illumina WGS
Surname freq. in US	$\sim 10^{-6}$	$\sim 10^{-4}$	$\sim 10^{-4}$	$\sim 10^{-3}$	$\sim 10^{-6}$
Meioses between target to source	3	5	5	7	2
Relationship between target and source	Nephew	First cousin once removed	Great-great nephew	Second cousin once removed	Grandchild
Supporting evidence	State of residency, pedigree structure, age, and maiden name are the same		State of residency, pedigree structure, age, and maiden name are the same		State of residency, pedigree structure are the same (ages are not given)
P(random match)*	$<5 \times 10^{-9}$		$<5 \times 10^{-6}$		$<10^{-5}$

Figure B-4: **Comparison of CEU identification cases.**

B.2 Acknowledgements

We thank FamilyTreeDNA and SMGF for technical assistance. The authors would like also to thank Dina Esposito, Alon Goren, Gerry Fink, David Page, Wendy Kramer, and Roy Ronen, for useful discussions. Y.E. is an Andria and Paul Heafy Family Fellow. This publication was

supported by the National Defense Science and Engineering Graduate Fellowship (M.G.) and by the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University (D.G. and E.H.).

B.3 Supplementary Material

B.3.1 Evaluating the general risk of surname recovery

Downloading Ysearch data

The Ysearch website belongs to FamilyTreeDNA (FTDNA), a Texas-based genetic genealogy company. The website allows users, regardless of their testing service, to voluntarily post their Y-STR genotyping results along with their ancestral information and contact details. Based on the data posted on the website, approximately 85% of Ysearch's users were tested with FamilyTreeDNA and the other 15% were tested with other genetic genealogy services. Users from other services are advised to post their results using FamilyTreeDNA nomenclature, and the website offers a conversion table between popular genetic genealogy services and FamilyTreeDNA nomenclature.

With permission from FamilyTreeDNA, we scraped the entire Ysearch database in May 2011. Some areas are protected by reCaptcha and were accessed manually. After parsing and merging the HTML files, we obtained 95,000 surname-haplotype entries, each of which contained: Ysearch userID, surname, ancestral location, and Y-STR results.

Access to the SMGF database

The SMGF website belongs to the Sorenson Molecular Genealogy Foundation, a Utah-based non-profit genetic genealogy organization that was recently acquired by Ancestry.com. The website allows users to query the SMGF database but not to create new records, and all records are from the SMGF program. Unlike the Ysearch database, we could not download the database records to our server. With permission from SMGF, we conducted massive queries of their database using an automatic script. The webpages that contained the top 10 results based on the SMGF matching algorithm were downloaded and parsed to identify the matches.

Concordance between genealogical databases and the US population

The surname distribution in the general US population was estimated using the Census 2000 study that is based on 270 million records (http://www.census.gov/genealogy/www/data/2000surnames/index.html). The Census study lists 151,671 surnames along with their relative prevalence in the general population and ethnic composition in sorted order. To protect the privacy of the participants and due to sample size limitations, the Census data stops when the cumulative frequency of the surnames reaches 90%, and does not include surnames that are found in less than 100 individuals each.

We compared the surname distribution in Ysearch and SMGF to the distribution in the general US population in order to evaluate the completeness of the databases. We defined the census coverage probability, denoted by c, as the chance that the surname of an individual drawn at random from the US population has at least a single haplotype record in one of these databases, and found that c=68.5%. The correlation between the US population and the genealogical records was evaluated by a permutation test with 10,000 repetitions. We obtained the following statistics: $E[SSE_{permutations}] = 9.01 \times 10^6$, $\sigma(SSE_{permutations}) = 2437$. The hypothesis SSE was 1.99×10^6. The p-value was calculated using one-sided Chebyshev bound.

A mathematical model for surname leakage risk

Search method

Our database search method relied on finding a record that shares the closest Time to Most Recent Common Ancestor (TMRCA) with the queried haplotype. The rationale behind this strategy is that close patrilineal relatives have a higher probability of sharing the same surname. For instance, one can imagine that monozygotic twins have a high probability of sharing the same surname, whereas a pair of Y chromosomes whose MRCA lived before the formation of the surname system would have a low probability of sharing the same surname.

Walsh [**?**] has proposed several Bayesian models for estimating the distribution of the TMRCA in non-recombining haplotypes. We used his "infinite alleles model with differential mutation rates". Consider two Y chromosome haplotypes with n STR loci denoted by $\vec{v} = (v_1, v_2, \ldots, v_n)$ and $\vec{u} = (u_1, u_2, \ldots, u_n)$, with vector elements corresponding to the allele lengths. Let $\vec{x} = (x_1, x_2, \ldots, x_n)$ be a binary vector with $x_i = 1$ for a match at the i-th locus of $\vec{v}$ and $\vec{u}$, and $x_i = 0$ otherwise, and let $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$ be a vector whose elements denote the

probability of a mutation per meiosis in each marker. According to Walsh's model, the probability distribution function (PDF) of the TMRCA between the two haplotypes is:

$$P(t|\vec{x},\vec{\mu},N_e) = \frac{e^{-t(\frac{1}{N_e}+2\sum_{i=1}^{n}\mu_i x_i)}\Pi_{i=1}^{n}(1-e^{-2t\mu_i})^{(1-x_i)}}{I(\vec{x},\vec{\mu},N_e)} \tag{B.1}$$

where N_e is the effective male population size, and I is a normalization factor to ensure that $\sum_{t=0}^{\infty} P(t|\vec{x},\vec{\mu},N_e) = 1$. Following Thomson et al. [?], N_e was set to 10,000 males. The mutation rates were obtained from the extensive study of Ballantyne, et al [?].

The expected TMRCA is denoted by ÏĎ and is given by:

$$\tau = \sum_{t=0}^{\infty} t_i P(t_i|\vec{x},\vec{\mu},N_e) \tag{B.2}$$

The recovered surname was selected according to the record that has the minimal τ to the searched haplotype. Due to technical constraints with the web queries to SMGF and in order to reduce the amount of calculations, we did not determine τ for each of the hundreds of thousands of users in the databases. Instead, we employed the following procedure: (i) Ysearch - identify a set of candidate records that have the maximal number of matching markers to the queried haplotype (ii) SMGF âĂŞ use the native SMGF search tool to identify the top 10 candidates according to the website's proprietary algorithm (iii) Both âĂŞ calculate τ for top candidates in Ysearch and SMGF using Eq. B.1, and select the record with the minimal τ of the searched haplotype.

Retrieval confidence score

The retrieval confidence score determined the probability that the TMRCA of the retrieved record is indeed shorter than that of (i) a record with a distinct surname that has the second to shortest TMRCA and (ii) a random person from the population. Let P_1 and P_2 be the TMRCA PDFs of the best record and second best record according to Eq. B.1, and let P3 be the PDF of coalescent in a Fisher-Wright population: $P_3(t|N_e) = N_e^{-1}e^{-N_e t}$. In addition, let F_i be the cumulative probability distribution function of P_i. The retrieval confidence score, δ, is given by:

$$\delta(P_1, P_2, P_e) = \sum_{j_1=1}^{T} P_1(j_1)\Bigg(\sum_{j_2>j_1}^{T} P_2(j_2)\Big(\sum_{j_3>j_1}^{T} P_3(j_3)\Big)\Bigg) = \sum_{j=1}^{T} P_1(j)(1-F_2(j))(1-F_3(j)) \tag{B.3}$$

T is the number of generations that is practical for the patrilineal surname system and was set to 20 generations, corresponding to ∼1400 AD. P_2 was obtained by scanning records in the list that was generated in step (iii); candidate records with less than 20 markers were excluded as well as records with surnames that matched the top hit.

Surname inference

We set a threshold, δ_0, which denotes the minimal accepted quality for valid surname recovery. If the retrieval passed the confidence threshold, the algorithm inferred that the record's surname is the surname of the input haplotype. Otherwise, the algorithm rejected the inference and returned "Unknown". 1.8% of the searches returned records with an empty surname field or with strings that are not found in the surname list of the US census such as "AshkenaziJewishModal". The algorithm reported these cases as "Unknown" as well. Finally, TMRCA ties between two or more records with distinct surnames were also treated as "Unknown".

A surname inference resulted in one of the following outcomes: success âĂŞ the recovered surname is concordant with the true surname, wrong âĂŞ the recovered surname does not match the true surname, unknown âĂŞ below confidence threshold, non-valid surnames, and ties.

Following previous record linkage studies [**?**, **?**], successful recoveries included a small number of cases where the returned surname displayed a minute spelling variant from the true one, such as Abernathy and Abernethy. These cases can still direct the adversary in tracing back the target at the price of searching for a larger number of individuals. We adopted a stringent approach to detect spelling variants that required that the first letter of both surnames be identical and that the Jaro-Winkler string distance [**?**] of the surnames be at least 0.9. This relies on the observation that the suffix of a surname is more prone to mutate than the prefix [**?**]. Two percent of the queries showed spelling variants using this approach and they are summarized in the following table:

Manual inspection of the genealogical records showed that in a large number of these cases the users indicated the spelling variant as an alternative ancestral surname.

True surname	Retrieved surname	Jaro-Winkler distance
ABERNATHY	ABERNETHY	0.977
AYRES	AYERS	0.96
BAIRD	BEARD	0.933
BRALLEY	BRAWLEY	0.947
BRITTON	BRITTAIN	0.944
CHRISTIE	CHRISTISON	0.94
CLARK	CLARKE	0.967
COLLISON	CULLISON	0.964
DENNEY	DENNY	0.967
DUFF	DUFFEL	0.933
FLICKINGER	FLUCKIGER	0.93
MCMURTRY	MCMURTREY	0.984
MILLICAN	MILLIKEN	0.937
PALLETT	PARLETTE	0.919
PARLET	PARLETTE	0.956
SAYRE	SAYER	0.961
SEELYE	SEELY	0.967
WETHERINGTON	WITHERINGTON	0.961

The general risk of surname leakage from personal genomes is dictated by three factors: the prior distribution of surnames in personal genomes datasets, the distribution of haplotypes within a surname, and the ability to successfully retrieve the surname from the database using the haplotype. For simplicity, we assumed that the distribution of surnames of personal genomes is similar to the distribution of surnames in the population.

Let $I_x(h, s)$ be an indicator function that returns 1 if querying the database with the combination of haplotype h and surname s returns the outcome x, where x is either: "success", "wrong", or "unknown". Let f_s be the frequency of a surname and Îś(h,s) be the frequency of haplotype h in the surname s. Define $\beta_x(s) = \sum_{h \in H(s)} \alpha(h, s) I_x(h, s)$, where $H(s)$ is the set of haplotypes that are associated with the surname s. The probability of the surname recovery outcome x for a given population is:

$$P(x) = \frac{\sum_{s \in S} f_s \beta_x(s)}{\sum_{s \in S} f_s} \tag{B.4}$$

Where S is the set of all surnames in the population. The probability in Eq. B.4 can be assessed by sampling individuals from the population using the following estimator:

$$P(x) = \frac{\sum_{s \in S} \hat{f}_s \hat{\beta}_x(s)}{\sum_{s \in S} \hat{f}_s} c + \frac{\sum_{s \in \overline{S}} \hat{f}_s \hat{\beta}_x(s)}{\sum_{s \in \overline{S}} \hat{f}_s} (1 - c) \tag{B.5}$$

where S is the set of surnames in the sample that are known to be present in the tested databases and $\overline{S}$ is the set of surnames in the sample that are known to be absent from the tested databases. $\hat{f}_s$ is the estimated frequency of the surname based on the Census data, $\hat{\beta}_x(s) = \sum_{h \in H(s)} \hat{\alpha}(h, s) I_x(h, s)$ and $\hat{\alpha}(h, s)$ is the frequency of the haplotype-surname combination in the sample, and c is the census coverage probability that was determined above. Eq. B.5 models the outcome rates as a weighted sum of sampling individuals from two distinct strata: those whose surname is found in the databases and those who do not. The two weights mitigate potential ascertainment biases in the sample and increase the confidence that the results reflect the target population.

Estimating the risk of surname leakage by inter-database comparisons

Our input sample relied on a cohort of individuals from the YBase database. This database was maintained by DNA Heritage and was acquired by FamilyTreeDNA in April 2011. FamilyTreeDNA provided us with surname-haplotype records from the database, without other identifiers that can expose the identity of the database users. The YBase and SMGF entries are completely distinct because the SMGF database lists only SMGF users. We took the following steps to remove potential duplicate records between Ysearch and Ybase: first, we asked FamilyTreeDNA to exclude YBase entries whose email addresses appear in Ysearch as well as entries without email addresses. Second, we removed from the downloaded copy of Ysearch all ~900 users that were tested with DNA Heritage. Third, we excluded any YBase user whose haplotype did not show a combination of markers that are typical to the DNA Heritage test panel. Thus, the input cohort was tested with a different company (DNA Heritage) than the database users. This reduces the chance of ascertainment biases due to oversampling of close relatives of the database participants.

Genetic genealogy databases are subject to nomenclature heterogeneity that can confound the analysis. This is especially problematic for DNA Heritage test panels that were subject to five nomenclature changes between 2003 to 2009 (see: http://web.archive.org/web/

20100307032155/http://www.dnaheritage.com/helpfiles/DNA_Heritage_nomenclature_changes.pdf). For each input haplotype, we inspected the allelic ranges for markers that underwent significant nomenclature changes, such as DYS452, to decipher the nomenclature stratum and to standardize the haplotype according to the NIST recommended nomenclature. In addition, we set a tolerable genotype range for each marker that is equal to the marker mean value in Ysearch±3std. Entries outside of this range have a high likelihood of nomenclature differences and typos of users. This step filtered approximately 5% of YBase haplotypes. Finally, we selected only YBase haplotypes that have full genotyping results for a set of 34 STR markers (**Supplementary Table B.5.2**) and whose surnames are in the US census. At the end of this process, we retained 911 YBase records.

We used a series of Perl scripts to challenge Ysearch and SMGF with the YBase haplotypes and to compare the returned surnames to the true ones. SMGF searches were conducted with the NIST nomenclature and Ysearch searches were conducted with FamilyTreeDNA nomenclature. The standard deviation was calculated by 30 iterations of re-sampling with replacement participants from the input cohort and repeating the analysis process.

The results of the 911 queries exhibited distinct patterns between the TMRCA of records that exactly match the true surname, records with a spelling variant, and records that returned the wrong surnames (**Supplementary Figure B.4.1**). The mean TMRCA was 10.3 generations for exact matches, 15.6 generations for a spelling variant, and 24.3 generations for wrong surnames. The TMRCA distribution of exact matches appeared to follow a geometric distribution trend. The TRMCA of records with spelling variants was almost never more recent than 10 generations and was quite different from the distribution of wrong matches. This provides another support for our spelling variations detection algorithm. **Supplementary Figure B.4.2** shows the final results after processing the results according to Eq. B.5.

B.3.2 From Surnames To Individuals

The frequency distribution of leaked surnames

We determined the frequency distribution of leaked surnames from the YBase simulations using the following equation:

$$P(s \in S_i | x = success, \delta) = \frac{P(x = success | s \in S_i, \delta) P(s \in S_i)}{P(x = success | \delta)} \quad \text{(B.6)}$$

Where S_i is a subset of surnames whose frequencies fall in the i-th bin out of j possible bins. Specifically, we used the following bins:

Bin(i)	Frequency boudnaries	Example of surnames in bin
1	>1:400	Smith, Johnson
2	1:400 âĂŞ 1:4,000	Turner, Collins
3	1:4,000 âĂŞ 1:40,000	Gates, Sloan
4	1:40,000 âĂŞ 1:400,000	Bjork, Reach
5	<1:400,000	Kellog, Venter

The term $P(s \in S_i)$ in Eq. B.6 is given by the census data. The other numerator term can be approximated using a slight modification to Eq. B.5:

$$P(x = success | s \in S_i, \delta) = \frac{\sum_{s \in S} \hat{f}_s \hat{\beta}_x(s)}{\sum_{s \in S} \hat{f}_s} c_i + \frac{\sum_{s \in \overline{S}} \hat{f}_s \hat{\beta}_x(s)}{\sum_{s \in \overline{S}} \hat{f}_s} (1 - c_i) \quad \text{(B.7)}$$

Where c_i is a normalization factor that denotes the probability that a random person from the US population whose surname is in the i-th bin has at least a single entry in Ysearch and SMGF. c_i was determined by intersecting the census data with the list of Ysearch and SMGF. We used δ=0.82.

The leaked surnames are mostly found in the intermediate bin with a frequency of 1:4,000-1:40,000. Extremely rare surnames have the lowest relative risk for leakage due to the absence of records in Ysearch and SMGF. However, if these databases have even a single record for an extremely rare surname, then there is a 43% chance that the surname will be exposed (**Supplementary Figure B.4.3**). This phenomenon is potentially due to the small number of male lineages in extremely rare surnames.

Combining surnames with demographic identifiers

The joint probabilities of sex, age, and state were obtained from the US Census Population Estimates Program (www.census.gov/popest/states/asrh/files/SC-EST2009-AGESEX-RES.csv). The data is based on Census 2000 and contains a projection of residents to 2009, which

was used in the simulation. Similar to the HIPAA law, ages that are over 85 were grouped in a single category.

The simulation ran 100,000 times. In each round, a combination of state and age was selected according to their probability in the joint distribution. For instance, there are 287,000 males in California who are 25 years old and 3,500 males in Idaho who are 75 years old. Accordingly, the probability of selecting "California, 25" was 82 times higher than selecting "Idaho, 75". Next, a bin of a leaked surname was selected according to its probability in Eq. B.7 and a surname was selected according to its frequency in the bin. For instance, in the case of selecting the 1st bin ($\geq$1:400), Smith had 1.28 higher probability of being sampled than Johnson. Finally, the simulation randomly selected between the return of a spelling variant or exact match, where the former had a probability 11.11%, based on our empirical findings in the Ybase simulations. In case of no spelling variant, the surname frequency was set to the census frequency; otherwise, the surname frequency was selected to be the sum of frequencies of all surnames that can be spelling variants of the original surname according to our spelling variant definition above. The last step portrays a scenario in which the adversary first looks for the target with the returned surname and if he cannot trace the target back, he tries all spelling variants. The number of expected individuals was found by multiplying the surname frequency by the number of males with the selected age and geographical location.

We validated the results of the simulation by comparing them to real datasets of US residents from PeopleFinders (www.peoplefinders.com). These datasets are based on extensive mining of public records, such as voter and drivers license registries, and can be searched by a combination of surname, age, and state. We selected 30 random simulation rounds that passed two criteria: (a) the ages were restricted to 25-35 years to avoid potential confounding due to underrepresentation of minors in public records and conflicting records from deceased individuals (b) the expected number of individuals should be 10-100 to avoid overloading the website. In most cases the lists in PeopleFinders were smaller than expected from simulations. Although we cannot rule out incompleteness of the website, the results also suggest that any underestimation of the list size - if it exists at all - is not significant.

B.3.3 Profiling Y-STRs from sequencing data

lobSTR usage

Unless otherwise specified, lobSTR v2.0.0 was used to profile Y-STRs from raw whole-genome sequencing data [?]. In brief, lobSTR acts in three steps: detecting reads with repetitive elements that are flanked with non-repetitive regions, aligning the flanking regions to a reference, and measuring the repeat length for each STR.

Improved Y-STR reference

We modified lobSTR's standard STR reference to include the genomic locations and nomenclatures of genealogical Y-STRs. These locations were found by conducting in silico PCR on the UCSC genome browser using published Y-STR primers [?, ?, ?, ?, ?, ?, ?, ?, ?] and by searching the FamilyTreeDNA Y chromosome browser (ymap.ftdna.com). Several STR markers reside in duplicated regions of the Y chromosome. For instance, DYS385 has two distinct alleles in a single individual. Since lobSTR filters multi-mappers, we kept only one entry of these markers in the modified reference. Markers DYS448 and DYS449 consist of two STR regions separated by a non-repetitive region. For these, a separate reference entry was created for each region and the final genotype was determined by adding the alleles profiled at each of the two STR regions.

We did not include eight genealogical markers in the reference due to various technical reasons: markers GAAT1B07 and DYS724a/b (also known as CDYa/b) were excluded because their corresponding genomic coordinates could not be determined despite extensive literature searches. DYS726 was excluded because the genetic genealogy nomenclature could not be determined. DYS425 is one of the four repetitive loci of DYF371 [?], and using short reads we could not uniquely determine which locus a read originated from. DXYS156-Y was excluded because it is not specific to the Y-chromosome. Marker DYS19b was not included in because it is present in 0.2% of the population. Marker DYS640 was incorrectly annotated in our original reference and discarded from further analysis. Marker DYS464a-d was excluded because in most cases we typed fewer than four alleles and could not accurately assign typed alleles to forms a-d. In summary, our reference included 34 out of the 36 markers used by the SMGF panel and 79 out of the 87 markers in the most comprehensive test panel of FamilyTreeDNA. The genomic coordinates and conventions used for each Y-STR are given in **Supplementary Table B.5.3**). All coordinates reported in this study follow the hg19 human reference build.

Processing lobSTR calls

lobSTR returns base pair length differences from the UCSC genome reference. Genetic genealogy services use an STR nomenclature that follows the PCR product sizes according to arbitrary primers [?]. Whenever available we used the NIST nomenclature to translate lobSTR results (http://www.cstl.nist.gov/strbase/ystr_fact.htm). For searches in the Ysearch database results were converted to FamilyTreeDNA nomenclature using a conversion table available from SMGF (http://www.smgf.org/ychromosome/marker_standards.jspx).

For Y-STRs with a single genomic location, the allele with the modal number of supporting reads was used. Y-STR alleles that showed a non-integer number of repeat copies were discarded. We manually inspected a small number of calls where the modal allele was supported by less than 60% of reads aligned to the locus and enhanced the call by removing reads likely to be erroneous, such as reads that contain a high number of sequence mismatches, reads in which the STR resides towards the end of the read, or reads supporting alleles outside the normal range. Importantly, this procedure was executed completely blind to the true allele if it was known. For bi-mapper markers, such as DYS413a/b, the shortest repeat length was assigned to allele "a" and the next to allele "b".

Comparing lobSTR to the CEPH Y-STR panel

General approach

Sequence data for the CEPH panel were downloaded from the NCBI Short Read Archive from experiment SRP009145, sample SRS269343, runs SRX103805-130812. The sample included 10 HGDP individuals: HGDP00456 (Mbuti Pygmy), HGDP00665 (Sardinian), HGDP01284 (Mandenka), HGDP00542 (Papuan), HGDP00521 (French), HGDP00778 (Han Chinese), HGDP01307 (Dai), HGDP00927 (Yoruba), HGDP01029 (San), HGDP00998 (Karitiana). Autosomal coverage was calculated using the samtools [?] depth tool and gives the average depth of covered bases based on alignments using BWA [?]. lobSTR 2.0.0 with the improved Y-STR panel was used for the analysis.

Genotypes for 76 Y-STRs typed by capillary electrophoresis for the 10 HGDP samples were obtained from the CEPH website (ftp://ftp.cephb.fr/hgdp_supp9/). Forty-seven of these markers overlapped with the lobSTR reference and were used to evaluate lobSTR's ability to type Y-STRs.

lobSTR reports alleles as the length difference from the UCSC, whereas the CEPH genotypes

are reported as the number of repeat copies at each locus. To convert lobSTR output to the same format, we used for following equation: $r+l/p$, where r is the number of base pairs of the STR of the lobSTR reference, l is the reported lobSTR allele in base-pairs, and p is the period of the Y-STR. For all individuals in which lobSTR recovered a genotype for DYS385a/b, only a single allele was returned. If the returned allele matched either the "a" or "b" form reported by CEPH, it was considered as correct. This follows our search strategy with the personal genomes, where these partial calls of multi-allelic markers were used to exclude matches not containing the lobSTR call for either allele.

We noticed that the lobSTR calls for all six individuals typed for DYS481 and all three individuals typed for DYS594 are exactly one repeat away from the results in the CEPH study. There is known nomenclature heterogeneity for these markers and some test kits report them with one shorter repeat than as reported by the NIST standard . Concordantly, we converted lobSTR calls to the shorter allele nomenclature to match that reported by CEPH.

Number of markers profiled at different sequencing coverage levels

Based on our previous experience with lobSTR, we assumed that STR coverage is linearly related to autosomal coverage. For each genome, we used the Picard (http://picard.sourceforge.net) DownsampleSam tool to randomly down-sample reads from the lobSTR alignment file to simulate coverage levels corresponding to autosomal coverage ranging from 1x to 25x. For each coverage level, we repeated the lobSTR allelotyping step to call the Y-STRs. The best-fit saturation curve was found using nonlinear least squares to fit a hyperbolic curve and was extended to predict haplotype lengths for up to 50x coverage.

Further investigation of wrong Y-STR calls

In our previous studies, we found that PCR stutter noise is a major source of error in calling STR alleles. This type of noise usually adds or subtracts a single repeat unit from the true allele. We noticed that the erroneous calls in DYS490 and DYS572 are several repeats away from the true allele, reducing the probability that these errors stem from stutter noise. Further analysis found that these two markers have X chromosome homologs, and that the calling errors can be attributed to misalignment of the X chromosome STRs. We also noticed that these markers were occasionally detected in the female genomes of the CEU panel, which provides further support for this hypothesis. Future algorithm improvements can use the homolog calls from the X chromosome to detect these errors.

B.3.4 Cases of Surname Leakage from Personal Genomes

Querying genealogical databases

In all surname recovery experiments from personal genomes, database queries utilized the native search interfaces of the websites.

Ysearch was queried using the haplotype matching tool available at http://www.ysearch.org/search_search.asp?uid=&freeentry=true. Online searches were conducted with the default parameters and using the FamilyTreeDNA nomenclature. SMGF was queried using the tool at http://www.smgf.org/ychromosome/search.jspx with the options "Search by Match%) = 85%" using the NIST nomenclature.

Feasibility of record retrievals with Illumina

The US male sample from our lab collection

The sequencing experiment was approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES). To comply with the COUHES approval, we cannot share the specific Y-STR results, since this could lead to identification of the individual. As an alternative, we provide summary statistics of the length distribution of the detected Y-STR makers.

Four Catch-All buccal swabs (Epicentre, QEC89100) were used to collect the sample according to the manufacturer's protocol. Genomic DNA was obtained by QuickExtract (Epicentre), followed by phenol-chloroform purification and ethanol precipitation. Library preparation was performed according to the standard Illumina protocol. Three runs of 101bp paired-end reads were generated with a GAIIx platform, generating 740 million reads. Autosomal coverage of 13x (after removing PCR duplicates) was measured using a conventional alignment pipeline as previously described [?]. **Supplementary Figure B.4.4A** shows the overlap between the markers that were detected by Illumina versus the genealogical profile from Sorenson Genomics. textbfSupplementary Figure B.4.4B shows the number of STRs that were detected using Illumina and Sorenson as a function of their lengths.

Database retrieval

For the 10 HGDP samples and the US male sample, we supplemented our downloaded copy of the Ysearch database with capillary electrophoresis results. Retrieval of records was conducted with

the same scripts used for the Y-STR simulations described above. As an additional validation for our results, we created a Ysearch record for the US male using the Ysearch.org website that does not disclose the true surname of the sample and consists of the Y-STR makers that are shared between Sorenson Genomics and Ysearch. Again, a search with the default website interface returned our sample as the top match.

Analyzing Michael Snyder's genome

Raw reads for the blood-derived and saliva-derived DNA of Michael Snyder's genome were downloaded from the NCBI Sequence Read Archive with accessions SRX097307 and SRX097312, respectively. lobSTR 1.0.6 with the native lobSTR reference was used to process both datasets using 20 processors on a server with four 12-core AMD Opteron 6100 Series. Forty-eight Y-STR calls were generated. All Y-STR calls were concordant between the blood-derived and the saliva-derived samples.

Ysearch link to search this haplotype:

http://www.ysearch.org/search_results.asp?uid=&freeentry=true&L1=14&L2=0&L3=
16&L4=0&L5=10&L6=0&L7=0&L8=11&L9=13&L10=0&L11=0&L12=12&L13=0&L14=15&L15=0&L16=
0&L17=11&L18=11&L19=0&L20=0&L21=0&L22=0&L23=0&L24=0&L25=0&L26=0&L27=0&L28=0&
L29=0&L30=0&L31=0&L32=0&L33=0&L34=14&L35=18&L36=16&L37=19&L38=0&L39=0&L40=12&
L41=10&L54=11&L55=8&L56=0&L57=0&L58=8&L59=11&L60=10&L61=8&L62=10&L63=0&L42=0&
L64=22&L65=0&L66=0&L67=11&L68=12&L69=12&L70=0&L71=0&L49=13&L72=26&L73=0&L51=
0&L74=13&L75=11&L76=12&L77=0&L78=9&L79=12&L80=11&L43=0&L44=12&L45=12&L46=0&L47=
0&L48=13&L50=10&L52=0&L53=0&L81=9&L82=11&L83=14&L84=9&L85=15&L86=12&L87=0&L88=
0&L89=0&L90=11&L91=10&L92=11&L93=0&L94=10&L95=11&L96=0&L97=0&L98=0&L99=0&L100=
0&min_markers=8&mismatch_type=absolute&mismatches_max=0&mismatches_sliding_starting_
marker=8&recaptcha_challenge_field=03AHJ_VutYkPmq2enCrhZuu94gU9-tcPRX33GpxRzVYZGBmnUWrEecYh8
nMfhB0r8QTNBIe-_lpzJtyC3IRZ6SXlIn1Tnwb9vfGNo5ZojEQ8_8OlQgtCuVj5rTLfLLEXi4vr0-uFyo7upKwcsOFnx
recaptcha_response_field=Weighthe+resume&haplo=&region=

SMGF link to search this haplotype:

http://www.smgf.org/ychromosome/search_results.jspx?labStandard=NIST&searchType=
genetic&matchPercent=match_85&showCountries=on&showMissingData=on&showAllSurnames=
on&DYS385_a=None&DYS385_b=None&DYS426=11&DYS447=None&DYS461=None&DYS388=13&DYS437=

None&DYS448=None&DYS462=12&DYS389I=None&DYS438=10&DYS449=None&DYS463=None&DYS389B=
None&DYS439=None&DYS452=None&DYS464_a=None&DYS464_b=None&DYS390=None&DYS441=
14&DYS454=11&DYS464_c=None&DYS464_d=None&DYS391=10&DYS442=17&DYS455=11&GGAAT1B07=
None&DYS392=12&DYS444=13&DYS456=14&YCAII_a=None&YCAII_b=None&DYS393=14&DYS445=
10&DYS458=15&YGATAA10=14&DYS394=16&DYS446=None&DYS459_a=None&DYS459_b=None&YGATAC4=
None&DYS460=None&YGATAH4=None

Analyzing John West's genome

Raw reads for John West's genome were downloaded from NCBI Sequence Read Archive with accession SRA018104. lobSTR 1.0.6 with the improved Y-STR index using the same hardware settings for Michael Snyder genome.

Ysearch link to search this haplotype:

http://www.ysearch.org/search_results.asp?uid=&freeentry=true&L1=13&L2=0&L3=
14&L4=0&L5=11&L6=11&L7=14&L8=12&L9=12&L10=13&L11=0&L12=13&L13=0&L14=17&L15=0&
L16=0&L17=11&L18=10&L19=0&L20=15&L21=0&L22=0&L23=0&L24=0&L25=0&L26=0&L27=0&L28=
0&L29=0&L30=11&L31=10&L32=19&L33=23&L34=15&L35=19&L36=17&L37=17&L38=0&L39=0&
L40=12&L41=12&L54=11&L55=9&L56=0&L57=0&L58=8&L59=10&L60=10&L61=8&L62=9&L63=10&
L42=0&L64=0&L65=0&L66=16&L67=10&L68=12&L69=12&L70=15&L71=0&L49=12&L72=22&L73=
0&L51=13&L74=0&L75=11&L76=14&L77=0&L78=0&L79=0&L80=0&L43=12&L44=11&L45=14&L46=
0&L47=0&L48=13&L50=13&L52=0&L53=19&L81=9&L82=0&L83=16&L84=9&L85=16&L86=12&L87=
11&L88=13&L89=13&L90=11&L91=10&L92=12&L93=0&L94=11&L95=10&L96=0&L97=0&L98=0&
L99=0&L100=0&min_markers=8&mismatch_type=absolute&mismatches_max=0&mismatches_
sliding_starting_marker=8&recaptcha_challenge_field=03AHJ_VusNldFpOWxRw2dib-HZoXRWEvEIRysd8fl
HPkMw8AiwilCMic_tD_ntx119pL-fmM96E18ekPuaxXIu-0Dw0hIg&recaptcha_response_field=
Hcacco+and&haplo=&region=

SMGF link to search this haplotype:

http://www.smgf.org/ychromosome/search_results.jspx?labStandard=NIST&searchType=
genetic&matchPercent=match_85&showCountries=on&showMissingData=on&showAllSurnames=
on&DYS385_a=11&DYS385_b=14&DYS426=12&DYS447=None&DYS461=12&DYS388=12&DYS437=
15&DYS448=None&DYS462=11&DYS389I=None&DYS438=12&DYS449=None&DYS463=19&DYS389B=
None&DYS439=13&DYS452=None&DYS464_a=None&DYS464_b=None&DYS390=None&DYS441=14&

DYS454=11&DYS464_c=None&DYS464_d=None&DYS391=11&DYS442=17&DYS455=11&GGAAT1B07=None&DYS392=13&DYS444=12&DYS456=15&YCAII_a=19&YCAII_b=23&DYS393=13&DYS445=13&DYS458=17&YGATAA10=16&DYS394=14&DYS446=13&DYS459_a=None&DYS459_b=None&YGATAC4=None&DYS460=11&YGATAH4=11

Surname recovery using the Craig Venter dataset

Sequence reads for the Venter genome were downloaded from TraceDB (Genbank accession ABBA00000000). We trimmed the first 50bp of every read due to the high error rate at the beginning of Sanger sequence reads and discarded reads whose length after trimming was less than 100bp.

At the default settings, lobSTR 2.0.0 with the improved Y-STR index returned 40 Y-STRs after 40 minutes of runtime using the same hardware settings as described above. Markers returning a non-integer number of repeat copies were discarded. Ysearch link to search this haplotype:

http://www.ysearch.org/search_search.asp?fail=1&uid=&freeentry=true&L1=0&L2=0&L3=0&L4=0&L5=10&L6=0&L7=0&L8=12&L9=12&L10=12&L11=0&L12=13&L13=0&L14=17&L15=9&L16=0&L17=11&L18=11&L19=0&L20=0&L21=0&L22=0&L23=0&L24=0&L25=0&L26=0&L27=0&L28=0&L29=0&L30=0&L31=0&L32=19&L33=23&L34=0&L35=0&L36=0&L37=17&L38=0&L39=0&L40=12&L41=12&L54=12&L55=9&L56=15&L57=16&L58=9&L59=10&L60=10&L61=8&L62=0&L63=0&L42=0&L64=23&L65=0&L66=16&L67=10&L68=12&L69=0&L70=16&L71=8&L49=0&L72=22&L73=0&L51=0&L74=12&L75=11&L76=0&L77=0&L78=0&L79=13&L80=12&L43=12&L44=11&L45=0&L46=0&L47=0&L48=0&L50=0&L52=0&L53=0&L81=0&L82=0&L83=16&L84=9&L85=0&L86=0&L87=0&L88=0&L89=12&L90=11&L91=0&L92=0&L93=12&L94=11&L95=0&L96=25&L97=0&L98=0&L99=0&L100=0&min_markers=8&mismatch_type=absolute&mismatches_max=0&mismatches_sliding_starting_marker=8&recaptcha_challenge_field=03AHJ_VusyS2psJJIgHViP9Prgl35afzMpQdoc1uJYw3a1I3Lob-ycMFtpIo&recaptcha_response_field=tsshora+infinite&haplo=&region=

SMGF link to search this haplotype:

http://www.smgf.org/ychromosome/search_results.jspx?labStandard=NIST&searchType=genetic&matchPercent=match_85&showCountries=on&showMissingData=on&showAllSurnames=on&DYS385_a=None&DYS385_b=None&DYS426=12&DYS447=None&DYS461=12&DYS388=12&DYS437=None&DYS448=None&DYS462=11&DYS389I=None&DYS438=12&DYS449=None&DYS463=None&DYS389B=None&DYS439=12&DYS452=None&DYS464_a=None&DYS464_b=None&DYS390=None&DYS441=None&

DYS454=11&DYS464_c=None&DYS464_d=None&DYS391=10&DYS442=17&DYS455=11&GGAAT1B07=None&DYS392=13&DYS444=None&DYS456=None&YCAII_a=19&YCAII_b=23&DYS393=None&DYS445=None&DYS458=17&YGATAA10=None&DYS394=None&DYS446=None&DYS459_a=9&DYS459_b=None&YGATAC4=None&DYS460=None&YGATAH4=None

Querying Ysearch as described above returned the entry VPBT4 with surname "Venter" as the top hit. The results, including the trace numbers of supporting reads, are summarized in **Supplementary Table B.5.4** . Concordant with Craig Venter's paternal roots, the top match was the only Venter record in Ysearch with a UK ancestor. The algorithm did not return any of the other dozen Venter entries with ancestors mostly from Germany and South Africa (**Supplementary Figure B.4.5** and **Supplementary Table B.5.5**), emphasizing the specificity of the recovery.

Demographic profiling was conducted using PeopleFinders and USSearch (www.ussearch.com). Female names and users that did not exactly match year of birth=1946 were discarded.

CEU genomes

The CEU male datasets were accessed through the 1000Genomes publicly available Amazon S3 bucket and the European Nucleotide Archive. In cases of father-son pairs, we selected the father for further analysis. All datasets were first processed with lobSTR 2.0.0 with the native STR reference. We reran the 18 CEU genomes that returned the largest number of markers with the improved Y-STR panel. Overall, these genomes had longer read lengths of 76-100bp compared to 36-51bp and were therefore more amenable to STR calling. To validate calls in the low coverage genomes, Y-STRs typed using capillary electrophoresis for 16 Y-STR markers for 10 of the 17 individuals were obtained from He, et al. [**?**]. In 41/43 comparable markers the genotypes were concordant. The two incorrect cases were off by a single repeat unit and covered only by a single read. All searches were first performed using only the markers typed using lobSTR. Four genomes were supplemented with the markers from He, et al. since their searches returned a large number of poorly matching records due to low number of calls in popular markers. Autosomal coverages were measured as reported for the HGDP samples.

Determining the probability of false surname recovery

We determined the probability that at least one household would randomly match the surname and demographic characteristics of the CEU pedigrees. Let n be the number of households that hold the recovered surname in the geographical region, p the probability that a household matches additional metadata available for the sample, and f_1 and f_2 the frequencies of the recovered surname of the paternal and maternal grandfathers. If only one surname was recovered, $f_2 = 1$. The probability of at least a single random match is:

$$P(\geq 1 \texttt{ match}) = 1 - (1 - p)^n \tag{B.8}$$

In our case, n is the number of married households in Utah with the recovered surname. We approximate $n \approx \lceil n_{utah} f_1 \rceil$, where n_{utah} is the total number of married households in Utah, which according to the 2002 census matches to 443,210.

For p, we accounted for the additional metadata regarding the number of children, male/female order of the children, and knowledge of the surname of the other set of grandparents. We set p to:

$$p = f_2 p_c \frac{1}{2^k} \tag{B.9}$$

where p_c is the probability that a household has the given number of children, k is the number of children in the pedigree and $1/2^k$ is the probability that the male/female order of the children matches that in the pedigree. The upper bound of p_c is 3.5%, which corresponds to the percentage of households in Utah with 5 or more children as determined by the 2000 US Census using the search tool at factfinder2.census.gov. We used this number because data on larger households were not available. This gave the probability of finding at least one random match as:

$$P(\geq 1 \texttt{ match}) = 1 - (1 - f_2 p_c \frac{1}{2^k})^{n_{utah} f_1} \tag{B.10}$$

We note that the order in which surnames are assigned to surnames 1 and 2 does not significantly change this probability as, $1 - (1 - p)^n$ converges to np for small p, and therefore:

$$P(\geq 1 \texttt{ match}) \approx np = n_{utah} f_P f_M p_c \frac{1}{2^k} \tag{B.11}$$

which also gives the expected number of households that give random matches to the desired characteristics.

One limitation in our analysis is the $n \approx \lceil n_{utah} f_1 \rceil$ approximation that implies that the surname distribution in Utah is very close the surname distribution in the entire US. These two distributions are expected to be relatively close for highly prevalent surnames, but extremely rare surnames can be quite localized. This case was only of a concern for pedigree 3, where its surname is found in only a few hundred individuals in the US. To test the robustness of our analysis, we re-calculated the probability of a random match for this pedigree as if all individuals in the US with this surname live in Utah and each individual is a member of a distinct household. In this scenario, the probability of a random match was 0.3%, which is still significantly low. Notice that this analysis is extremely conservative. The assumption that each of the hundreds of individuals reside in a distinct household is not realistic. In addition, we did not take into account additional metadata, such as the probability to find the exact number of children and the fact that all grandparents were alive during the last year of CEU sample collection, which should further drive down the probability of a random match.

B.3.5 Y-STR masking and imputation

One potential solution to surname leakage is to mask the Y-STR loci. However, genetic masking is sensitive to imputation strategies. A striking example of this limitation was the ability to recover Jim Watson's masked ApoE status from adjacent SNPs in linkage disequilibrium [**?**], raising the possibility of also bypassing Y-STR masking.

Theoretically, it seems possible to impute genealogical Y-STR haplotypes from Y-SNPs. The rate of SNPs is 3×10^{-8} mutations per bp per generation, which translates to a rate of 0.5 *de novo* mutations in the euchromatic region of the Y chromosome per generation. On the other hand, Y-STR variations occur at a smaller rate of $\sim$0.1 mutations per haplotype of 30 markers per generation. This rate difference has been recently demonstrated by deep sequencing the Y chromosomes of two individuals that were separated by 13 meiosis events [**?**]. The two individuals had identical Y-STR haplotypes but differed at four Y-SNPs. The excess of *de novo* SNPs over STRs implies that Y-STR haplotypes can be uniquely tagged by Y-SNP haplotypes.

Y chromosome imputation has different properties imputation in autosomal regions. In the autosomes, recombination divides the chromosome into segments with distinct genealogies. The task of autosomal imputation algorithms is to detect segment transitions and match the corresponding ancestral haplotype block from the reference panel [**?**, **?**]. Y-STRs reside on one long chromosome block. The divide and conquer approach cannot work and the entire Y chromosome block must be imputed in a single step. On one hand, this drastically reduces the computation time needed for imputation. On the other hand, a necessary condition for accurate imputation is that the reference panel must include the Y-STR alleles as a single haplotype block. Accurate imputation will not work if the masked STR alleles are scattered across a collection of reference chromosomes. For instance, if the masked Y-STR haplotype is 14-15-20-11, and the reference has four chromosomes: 14-X-X-X, X-15-X-X, X-X-20-X, and X-X-X-11, where X indicates a mismatch to the masked haplotype, imputation will not return an accurate result. Given that condition, every imputed Y-STR haplotype (as opposed to alleles in the autosome) must be documented in the reference panel.

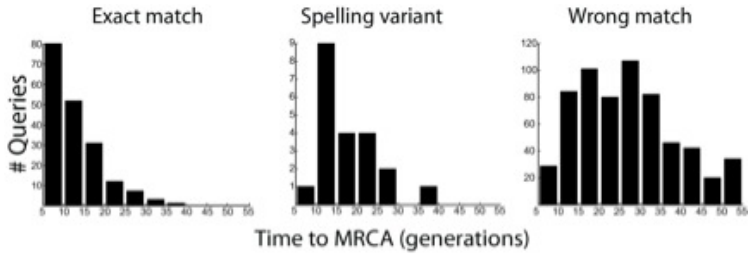
We evaluated the dependency between the reference panel size and the success rates. We focused on Ysearch since SMGF does not list the raw Y-STR haplotypes. Ysearch contains approximately 34,000 unique haplotypes of 30 popular STR markers. These haplotypes cover 34.5% of the haplotypes that segregate in the population according to the Good-Turing frequency estimation procedure (29). The reference panels were constructed by re-sampling Ysearch haplotypes using a two-stage procedure: (a) with a probability of 100%-34.5=65.5%, a mock haplotype was sampled. This denotes a haplotype in the reference panel that is not in Ysearch. Otherwise, the procedure continued to the next stage (b) a Ysearch haplotype was sampled according to its frequency in the database. This two-stage procedure was run N times, where N was the size of the reference panel. Simulating Y-SNPs was not necessary because we assumed that given the size of the haplotype block, imputation always correctly recovers the Y-STR haplotype from the Y-SNP, as long as the former is in the panel. We then conducted surname recovery experiments with YBase using the Ysearch database and the simulated reference panel. If a YBase haplotype was not part of the reference panel, then surname recovery automatically failed and was categorized under the âĂŸunknown' state.

Our results show that with large reference panels of 50,000 male genomes from the US population, the surname recovery success rate is 5% (**Supplementary Figure B.4.6**). This suggests that imputation is not an immediate threat to masking, but can be problematic as a long term solution.

In addition, we noticed that some community efforts, such as Y Chromosome Genome Comparison (daver.info/ysub), have started linking between Y-SNPs and surnames. These efforts might also enable the bypassing of Y-STR masking.

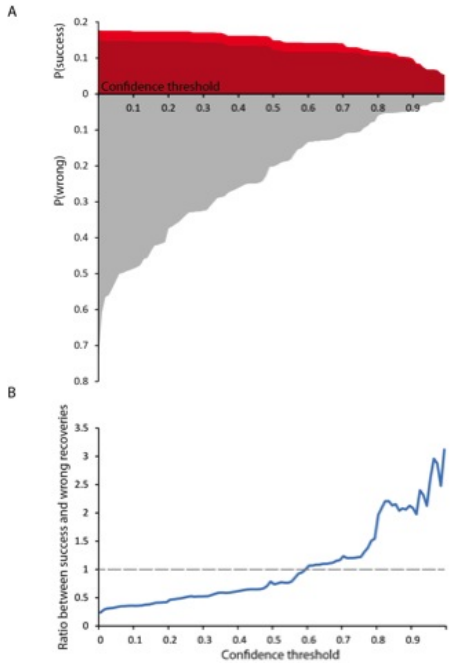
B.4 Supplemental Figures

B.4.1 Supplemental Figure 1



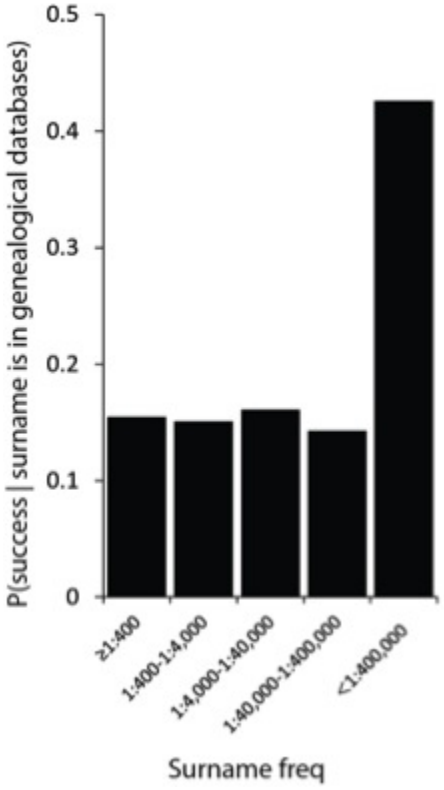
The TMRCA profiles of haplotype queries. Records that matched exactly the input surname (left) showed a geometric-like distribution. For most records with a minute spelling variant from the original surname (center) the MRCA was 10-15 generations ago. Wrong matches (right) mainly showed an ancient MRCA.

B.4.2 Supplemental Figure 2



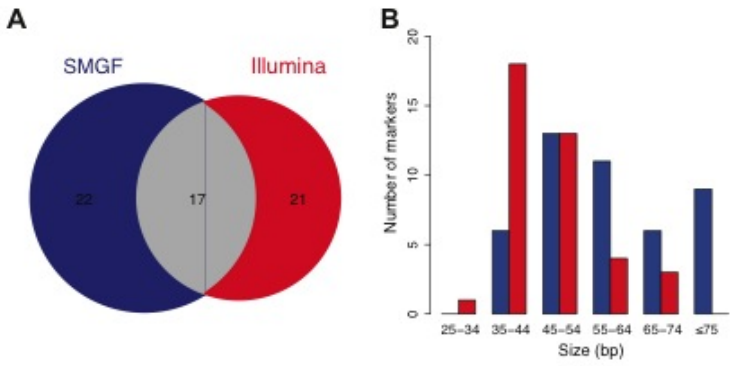
Performance of surname recovery at different confidence thresholds. (A) The rate of successful recovery with exact matches (dark red) and spelling variants (light red) versus the wrong recovery rate (gray) as a function of confidence threshold level. **(B)** The ratio between successful recoveries to wrong recoveries.

B.4.3 Supplemental Figure 3



The probability of successful recovery given that the surname has at least one record in Ysearch or SMGF as a function of the surname frequency.

B.4.4 Supplemental Figure 4

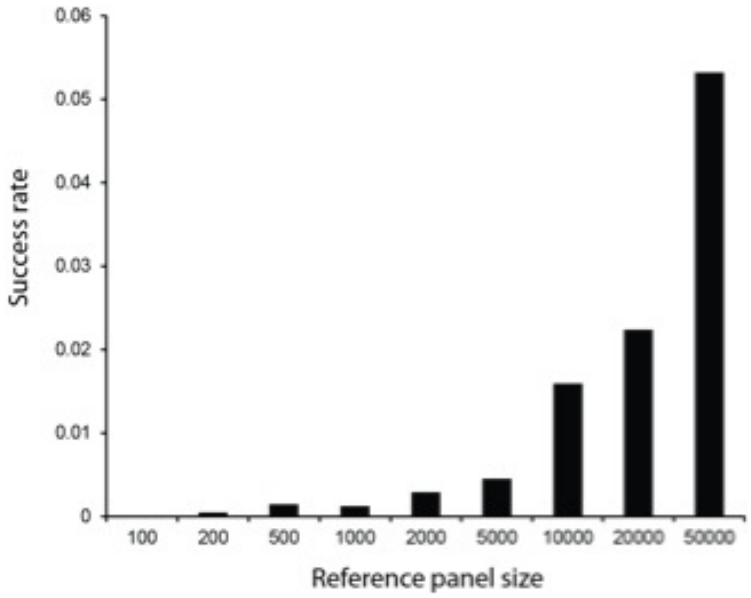


Comparison between Illumina Y-STR profiling and the Sorenson Genomics genetic genealogy service. (A) Illumina profiling returned the results of 38 Y-STR markers. The genetic genealogy service uses a panel of 49 markers, 39 of which are included in lobSTR's Y-STR reference. The results of all 17 markers that were profiled by both strategies were identical. **(B)** The distribution of total STR region lengths is shown for the markers typed by Sorenson (blue) versus markers typed by lobSTR (red).

B.4.5 Supplemental Figure 5

Ancestral origins of Venter records in Ysearch. The ancestral origin of the top match is labeled with an arrow.

B.4.6 Supplemental Figure 6



The estimated success rate for surname recovery after imputation as a function of the imputation panel size.

B.5 Supplemental Tables

B.5.1 Supplemental Table 1

	Site	Estimated number of Records	Maintained by:	Availability	Search Interface	Users
Databases	Ancestry DNA (dna.ancestry.com)	50,000	Ancestry.com	Public (fee required)	Search by STR Haplotype	Mainly Ancestry.com
	Family Tree DNA (familytreedna.com)	250,000	Family Tree DNA	Closed	Not searchable	FTDNA
	Oxford Ancestors (www.oxfordancestors.com/)	?	Oxford Ancestors	Closed	?	Oxford Ancestors
	SMGF (smgf.org/pages/ydatabase.jspx)	38,000	Sorenson Molecular Genealogy Foundation[1]	Public (free account required)	Search by surname or STR Haplotype	SMGF
	WorldFamilies (worldfamilies.net/surnames)	150,000[3]	Collection of admins of surname projects.	Public	Search by surname	Mainly FTDNA users
	Ybase	13,000	DNA Heritage[2]	Previously public. Discontinued.	Discontinued	DNA Heritage and others
	Y Chromosome Genome Comparison (daver.info/ysub)	1,000	Volunteers	Public	Download raw SNP data	23andME
	Ymatch (dna-fingerprint.com)	1,300	Family Tree DNA	Public	Search by STR and SNP haplotypes	Various companies
	Ysearch (ysearch.org)	105,000	Family Tree DNA	Public	Search by surname or STR Haplotype	Mainly FTDNA
Examples of familial sites	Brown DNA Study (http://brownsociety.org/browndna/results.htm)	800+	Brown members	Public	Table of Y-STR haplotypes	FTDNA
	Clan Donald USA (http://dna-project.clan-donald-usa.org/)	1000+	Donald clan members	Public	Table of Y-STR haplotypes	FTDNA
	McDuffie DNA Surname Project (http://www.mcduffiedna.com/)	150+	McDuffie members	Public	Table of Y-STR haplotypes	FTDNA, DNA Heritage
	SmithConnections DNA Project (http://www.smithconnections.com/)	500+	Smith members	Public	Table of Y-STR haplotypes	FTDNA
	Williams DNA Project (http://williams.genealogy.fm/)	800+	Williams members	Public	Table of Y-STR haplotypes	FTDNA

List of major genetic genealogy sites that display Y chromosome and surname information. The top section lists genetic genealogy databases. The bottom section lists examples of privately maintained familial genetic genealogy sites. [1] SMGF was recently acquired by Ancestry.com. [2] DNA Heritage was acquired by FamilyTreeDNA in 2011. [3] Includes only users whose surnames are present in the 2000 US Census.

B.5.2 Supplemental Table 2

Marker	Expected mutation rate	Mean	σ
DYS19	0.00437	14.34	0.8045
DYS385a	0.00208	12.0869	1.6522
DYS385b	0.00414	14.5464	1.449
DYS388	0.000425	12.5142	1.0753
DYS389a	0.00551	12.9668	0.6644
DYS389b	0.00383	29.326	1.0418
DYS390	0.00152	23.6032	1.0229
DYS391	0.00323	10.4858	0.6104
DYS392	0.00097	12.3413	1.1069
DYS393	0.00211	13.0752	0.6025
DYS426	0.000398	11.6459	0.5198
DYS437	0.00153	14.9094	0.6931
DYS438	0.000956	11.2206	1.0643
DYS439	0.00384	11.66	0.8567
DYS442	0.00978	17.2273	1.3301
DYS444	0.00545	12.3666	0.892
DYS445	0.00216	11.6015	0.9401
DYS446	0.00267	13.1767	1.372
DYS447	0.00212	24.6396	1.2057
DYS448	0.000394	19.3437	0.8748
DYS449	0.0122	29.5472	1.6474
DYS452	0.00402	30.1854	1.1041
DYS454	0.000475	11.0484	0.3744
DYS455	0.000426	10.648	0.9704
DYS456	0.00494	15.4571	1.1065
DYS458	0.00836	16.6389	1.2634
DYS459a	0.00013	8.753	0.5017
DYS459b	0.00013	9.601	0.5422
DYS460	0.00622	10.6976	0.639
DYS461	0.000989	11.882	0.6914
DYS462	0.00265	11.3571	0.6266
DYS464a	0.00018	13.8555	1.4488
DYS464b	0.00018	14.7374	1.0564
DYS464c	0.00018	15.8236	1.124
DYS464d	0.00018	16.5742	1.1157
DYS635	0.00385	22.6604	1.1601
GATA-A10	0.00332	15.5234	1.2242
GATA-H4	0.00322	10.7333	0.7801
GGAAT1B07	0.0024	10.2854	0.7397
YCAIIa	0.002	19.0997	0.905
YCAIIb	0.002	22.136	1.2624

Table B.1: **List of markers used to challenge Ysearch and SMGF.** Mutation rates are based on Ballantyne et al. [**?**]. YCAII was absent from this study and set to 0.002 according to Walsh [**?**]. Mean and standard deviations for marker values are calculated using Ysearch with NIST nomenclature.

B.5.3 Supplemental Table 3

Marker	Start	End	Alt location	Ref allele	Motif structure
DYS394/19	9521989	9522052		15	[TAGA]3TAGG[TAGA]n
DYS385a/b	20842518	20842573	chrY:19260956-19261212	14	[GAAA]n
DYS388	14747535	14747570		12	[ATT]n
DYS389I	14612191	14612238		12	[TCTG]m[TCTA]n
DYS389B	14612338	14612405		29	[TCTG]m[TCTA]n
DYS390	17274947	17275042		24	[TCTG]n[TCTA]m[TCTG]p[TCT
DYS391	14102795	14102838		11	[TCTA]n
DYS392	22633873	22633911		13	[TAT]n
DYS393	3131152	3131199		12	[AGAT]n
DYS406S1	23843595	23843634		10	[TATC]n
DYS413a/b	16099088	16099133	chrY:14676647-14676820	23	[TG]n
DYS426	19134850	19134885		12	[GTT]n
DYS434	14466533	14466568		9	TAAT[CTAT]n
DYS435	14496298	14496333		9	[TGGA]n
DYS436	15203862	15203897		12	[GTT]n
DYS437	14466994	14467057		16	[TCTA]n[TCTG]2[TCTA]4
DYS438	14937824	14937873		10	[TTTTC]n
DYS439	14515312	14515363		13	[GATA]n
DYS441	14981831	14981908		16	[TTCC]n
DYS442	14761103	14761168		17	[TATC]2[TGTC]3[TATC]n
DYS444	19226192	19226247		14	[TAGA]n
DYS445	22092602	22092649		12	[TTTA]n
DYS446	3131458	3131527		14	[TCTCT]n
DYS447	15278740	15278854		23	[TAATA]n[TAAAA]1[TAATA]m[T
DYS448_1	24365070	24365136		11	[AGAGAT]n
DYS448_2	24365178	24365225		8	[AGAGAT]n
DYS449_1	8218014	8218074		13	[TTTC]n

DYS449_2	8218124	8218179		14	[TTTC]n
DYS450	8126300	8126344		8	[ATTTT]n
DYS452	21620478	21620632		31	[TATAC]m[TGTAC]n[TATAC]p[C
DYS454	8224156	8224199		11	[AAAT]n
DYS455	6911569	6911612		11	[AAAT]n
DYS456	4270960	4271019		15	[AGAT]n
DYS458	7867880	7867943		16	[GAAA]n
DYS459a/b	26078851	26078890	chrY:26292857-26293004	10	[TAAA]n
DYS460	21050842	21050881		10	[ATAG]n
DYS461	21050690	21050737		12	[TAGA]n[CAGA]
DYS462	21317047	21317090		11	[TATG]n
DYS463	7643509	7643628		24	[AAAGG]m [AAGGG]n [AAGGA]
DYS472	16508484	16508507		8	[AAT]n
DYS481	8426378	8426443		22	[CTT]n
DYS485	22099634	22099681		16	[TTA]n
DYS487	8914174	8914212		13	[TTA]n
DYS490	3443765	3443800		12	[TTA]n
DYS492	17414337	17414369		12	[ATT]n
DYS494	21386168	21386197		10	[TTA]n
DYS495	15011300	15011346		15	[AAT]n
DYS505	3640831	3640878		12	[TCCT]n
DYS511	17304923	17304958		10	[GATA]n
DYS520	7730432	7730511		20	[ATAG]n[ATAC]n
DYS522	7415625	7415664		10	[GATA]n
DYS531	8466195	8466238		11	[AAAT]n
DYS533	18393226	18393273		12	[ATCT]n
DYS634	18392976	18393035		15	[CTTT]n
DYS537	19358850	19358889		10	[TCTA]n
DYS549	21520224	21520275		13	[GATA]n
DYS556	22601453	22601496		11	[AATA]n

DYS557	23234712	23234775		16	[TTTC]n
DYS565	16526732	16526775		12	[ATAA]n
DYS568	8822555	8822594		11	[AAAT]n
DYS570	6861231	6861298		17	[TTTC]n
DYS572	3679660	3679699		10	[AAAT]n
DYS575	7436257	7436296		10	[AAAT]n
DYS576	7053359	7053426		16	[AAAG]n
DYS578	22562564	22562599		9	[AAAT]n
DYS589	24485693	24485757		12	[TTTTA]n
DYS590	8555980	8556019		8	[TTTTG]n
DYS594	21656837	21656886		10	[AAATA]n
DYS607	18414382	18414457		19	[GAAG]n[GAAA][GAAG][GAAA][
DYS617	19081518	19081553		12	[TTAn]
DYS635	14379564	14379655		23	[TCTA]4[TGTA]2[TCTA]2[TGTA
DYS636	22634857	22634900		12	[ATTT]n
DYS638	17645491	17645534		11	[TTTA]n
DYS641	16134296	16134335		10	[TAAA]n
DYS643	17426012	17426066		11	[CTTTT]n
DYS714	22147731	22147865		27	[TTTCT]m[CTTCT]n[TTTCT]p
DYS717	17313245	17313324		16	[GTACT]m [GTATT]n
GATA-A10	18718879	18718938		15	[TCCA]2 [TATC]n
GATA-H4	18743553	18743600		12	[TAGA]n
YCAIIa/b	19622111	19622156	chrY:19016986-19017135	23	[CA]n
DYS395S1a/b	19739341	19739381	chrY:18899736-18899977	15	[AAC]n
DYS716	13140129	13140274		28	[ACTCGC][ACTCC]m[ATTCC]n[

Y-STR genomic locations and conventions. All coordinates are given for human genome build hg19. Conventions follow NIST guidelines whenever available. * The values for DYS448 and DYS449 were determined by adding the alleles typed at DYS448_1/DYS448_2 and DYS449_1/DYS449_2. The complete repeat structures for DYS448 and DYS449 are [AGAGAT]mN42[AGAGAT]n and [TTTC]m [N]50 [TTTC]n, respectively.

B.5.4 Supplemental Table 4

Marker	Craig Venter	Best Ysearch hit (VPBT4)
DYS388	12	12
DYS391	10	10
DYS392	13	13
DYS395S1a	15	15
DYS395S1b	16	16
DYS413a	23	23
DYS426	12	12
DYS436	12	12
DYS438	12	12
DYS439	12	12
DYS442	12	12
DYS450	8	8
DYS454	11	11
DYS455	11	11
DYS458	17	17
DYS459a	9	9
DYS461	12	
DYS462	11	
DYS472	8	8
DYS481	22	22
DYS485	16	
DYS492	13	13
DYS494	9	
DYS531	12	12
DYS534	16	16
DYS537	10	10
DYS549	12	

DYS556	11	
DYS557	16	16
DYS565	12	12
DYS568	11	11
DYS570	17	17
DYS578	9	9
DYS590	9	9
DYS594	10	10
DYS617	12	12
DYS636	12	
DYS638	11	
DYS641	10	10
DYS714	25	25
YCAIIa	19	19
YCAIIb	23	23

Craig Venter's haplotype from his personal genome versus the best Ysearch match. Only Ysearch markers with corresponding sequencing results are shown. All alleles are reported using FamilyTreeDNA nomenclature to match the Ysearch convention. For Genebank read accessions, see Supplemental Material at http://www.sciencemag.org/content/339/6117/321/suppl/DC1.

B.5.5 Supplemental Table 5

User surname	Ancestor surname	Origin
Venter	Von Dempter	Hameln, Germany
Venter	Venter	Bloemfontein, South Africa
Venter	Venter	Germany
Venter	von Dempter	Hamelin, Lower Saxony/Niedersachsen, Germany
Venter	Von Dempter	Hameln, Lower Saxony/Niedersachsen, Germany
Venter	Venter	Hamel, France
Venter	Venter	Witbank, South Africa
Venter	Von Dempter	Hameln, Germany
Venter	Venter	Hamln, Germany
Venter	Venter	Roth near Meisenheim, Palatinate/Pfalz, Germany
Venter		Lincolnshire, England
Venter	Venter	Hameln, Lower Saxony/Niedersachsen, Germany
Venter	van Deventer	Oldenzee, Netherlands

Table B.4: **Venter records in Ysearch and their ancestral origins.** In red - the top match to Craig Venter's Genome.

Appendix C

Worldwide variation in human short tandem repeats

Parts of this chapter have been submitted to *Nature* for publication:

> Mallick S, Li H, Lipson M, Mathieson I, **Gymrek M**, *et al.*. The landscape of human diveristy. Under revision. The STR work is mostly described in a supplemental chapter to this work:
>
> **Gymrek M**, Willems TF, Reich D, Erlich Y. Worldwide variation in human short tandem repeats.

Summary: We generated the most comprehensive catalog of short tandem repeat (STR) genotypes to date, based on 301 deeply sequenced human genomes. Genotypes show strong concordance with capillary electrophoresis and accurately recover population structure. We used this call set to characterize allele frequency spectra, analyze sequence determinants of STR variation, and to identify common loss of function alleles. STR genotypes are available in raw and interactive format at strcat.teamerlich.org.

C.1 Genotyping STRs

We analyzed STRs using lobSTR [**?**], a custom algorithm for genotyping short tandem repeats. We modified lobSTR's allelotyping tool to be able to call STRs directly from alignments generated by tools besides the lobSTR aligner. This greatly reduces the run time and allows rapid STR genotyping from large sequencing panels that have already been aligned using alternative indel-sensitive methods. We used raw reads aligned to GRCh37 using BWA-MEM (http://bio-bwa.sourceforge.net/) (version 0.7.10) with default parameters. These alignments were used as input to lobSTR's allelotyper (Github revision 3.0.3.24-892e). We

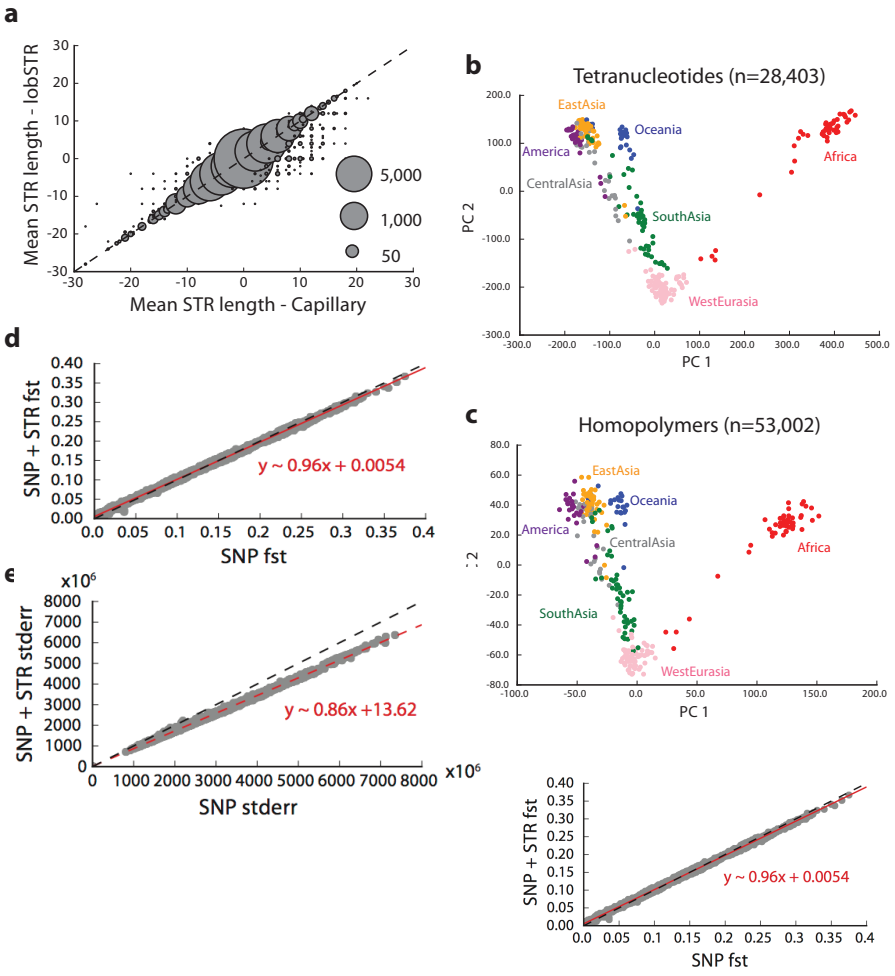


Figure C-1: **Worldwide variation in human short tandem repeats. a.** Mean STR length is reported as the average of the length difference (in bp) from the GRCh37 reference for each genotype. Bubble area scales with the number of calls compared at each point. **b.** and **c.** show the first two principal components after performing principal component analysis on tetranucleotide and homopolymer genotypes, respectively. Colors represent the region of origin of each sample. **d.** Pairwise Fst values between populations computed using only SNPs vs. using combined SNP+STR loci. **e.** Block jackknife standard errors for the SNP vs. SNP+STR F_{ST} analysis. The red dashed lines give the best-fit line, described by the formula in red. The black dashed line denotes the diagonal.

used optional parameters "--filter-mapq0 --filter-clipped --max-repeats-in-ends 3 --min-read-end-match 10" and a noise model trained on PCR-free sequencing data. We jointly genotyped samples at sites in lobSTR's reference panel: 1.6 million loci with motif lengths ranging from 1-6bp. The reference is part of the GRCh37 lobSTR resource bundle available at http://lobstr.teamerlich.org/download.html. **Table C.2** provides a summary of the reference panel.

C.2 Quality controls

lobSTR generated genotypes for an average of 1.5 million loci per sample (**Figure C.2a**) with an average of 15.3 informative reads (reads that completely span the repeat region) for each autosomal call. We removed sample S_Daur-1 from analysis because it had 6.5 standard deviations fewer calls than the mean (1.4 million calls). All other samples were within 3 standard deviations of the mean. For downstream population genetic analysis, we also removed individuals from the Bergamo and Hazara populations, as some of these individuals were outliers. Each locus had genotype calls for an average of 280 samples (95%) (**Figure C.2b**). We were not able to genotype 2% of loci in our reference. Most of these loci have allele lengths greater than 100bp that could not be spanned by Illumina reads. Genotype quality scores, which report the likelihood of the genotype call divided by the sum of likelihoods of all considered genotypes, tended to decrease for longer STRs and increase with motif length, with homopolymers showing significantly lower quality scores than other classes (**Figure C.2c**). For the majority of loci, we found no directional bias in allele length compared to the reference allele. However, as the reference track increases, calls become biased toward shorter alleles, again reflecting the limitation of calling STR genotypes using 100bp reads ((**Figure C.2d**)).

We subjected the resulting genotypes to stringent filtering to ensure high quality calls. We based our filters on coverage, call rate (percent of samples with a genotype call for a given locus), and the metrics Q and DISTENDS reported in the VCF file generated by lobSTR. Q reports the genotype quality score as described above. DISTENDS reports the mean distance between the STR boundary and the end of the read. Specifically, we calculate the difference in distance between the STR and the left and right read ends, and take the average difference across all reads for a given call. We find that high quality calls tend to have DISTENDS close to 0, meaning there is no bias towards a specific end of the read on which the STR occurs. On the other hand large positive or negative DISTENDS scores often indicate that a locus has

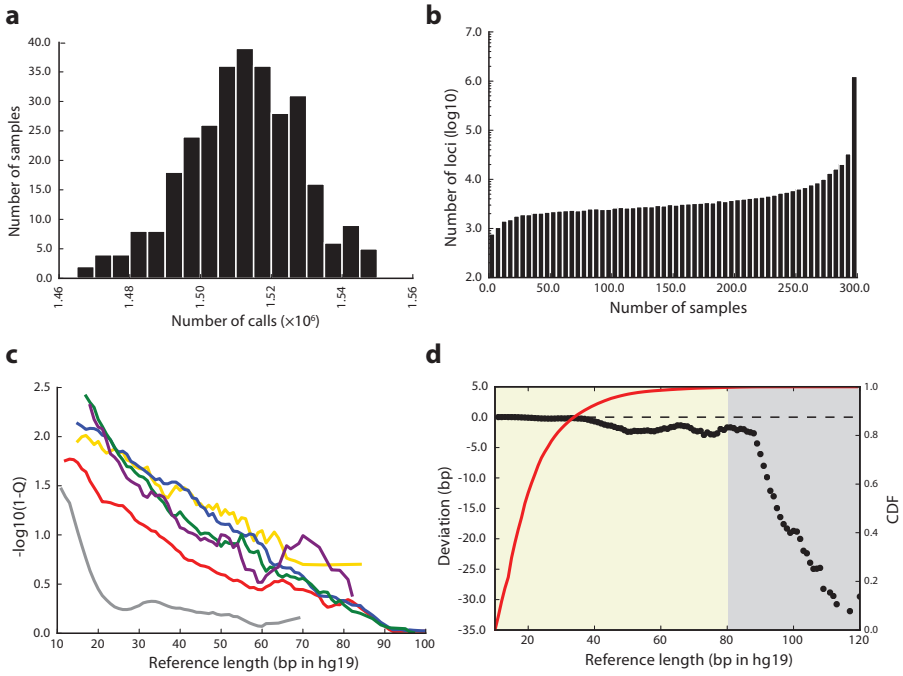


Figure C-2: **STR call set quality metrics.** **a**. Distribution of the number of STR calls per sample. **b**. Distribution of the number of samples with calls for each STR. **c**. Mean genotype quality score decreases with the length of the STR. Each line represents a different repeat motif length (gray = homopolymers, red = dinucleotides, yellow = trinucleotides, blue = tetranucleotides, green = pentanucleotides, purple = hexanucleotides). **d**. Mean length deviation from the reference allele as a function of reference length (black). As the reference track increases in length, calls tend to be biased toward alleles shorter than the reference allele (black). The red line gives the Cumulative Distribution Function (CDF) of calls vs. reference length. Gray shading: loci that were filtered from analysis. Beige: loci retained for downstream analysis.

problematic alignments.

The specific filters listed below are described in the "Best practices for using BWA-MEM alignments with lobSTR" section of the lobSTR website. We filtered loci with the following properties:

- Average coverage $5\times$
- Average $-\log_{10}(1-Q) < 0.8$
- Call rate < 0.8

Motif length	No. of loci	% in reference	Common motifs	% genotyped (post-filtering)
1	795,043	48.5	A, C	99.9 (70.3)
2	310,761	19.0	AC, AT, AG	96.2 (88.5)
3	84,869	5.2	AAT, AAC, AGG, AAG, ATC	97.6 (95.6)
4	262,179	16.0	AAAT, AAAC, AAAG, AAGG, AATG, AGAT, AGGG, ATCC, ACAT	94.3 (91.8)
5	106,481	6.5	AAAAC, AAAAT, AAAAG	97.4 (93.1)
6	79,246	4.8	AAAAAC, AAAAAT, AAAAAG	97.4 (93.3)
All	1,638,516	100.0		97.9 (81.1)

Table C.1: **Composition of GRCh37 lobSTR reference panel.** We list motifs that occur >5,000 times in the reference, in order from most to least common.

- Reference allele length >80bp

After filtering loci we additionally filtered individual calls with:

- Coverage $< 5\times$
- $-\log_{10}(1-Q) < 0.8$
- Absolute value of DISTENDS score >20

After filtering, 1.3 million loci remained for analysis.

C.3 Validation

We compared lobSTR results to genotypes for a subset of samples generated using capillary electrophoresis, the gold standard for STR genotyping. We evaluated concordance with two panels: Y chromosome STRs (mostly tetranucleotide loci), and the Marshfield set of mostly di- and tetranucleotide autosomal loci.

We obtained Y-STR genotypes for 74 samples at 39 loci that overlapped with our data from the CEPH website (ftp://ftp.cephb.fr/hgdp_supp9/genotype-supp9.txt). We calibrated capillary calls to the lobSTR format using the reference alleles annotated in Supplementary Table 5 of Gymrek et al. [**?**]. As reported there, markers DYS481 and DYS594 are off by one unit in the CEPH data, and we corrected the lobSTR calls to reflect this. We discarded marker DYS640 due to ambiguous nomenclature. We found overall concordance of 99% between the lobSTR and capillary calls.

We downloaded genotypes and additional metadata for the Marshfield markers for 1,048 samples at 627 loci from the Rosenberg lab website as reported by Pemberton et al. [**?**]. A total of 127

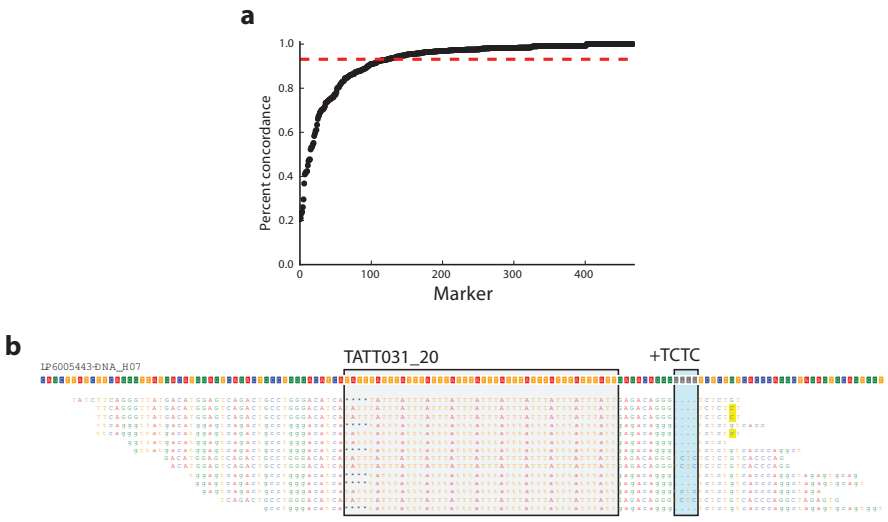


Figure C-3: **Concordance between lobSTR and capillary genotypes. a**. Concordance by marker, ordered from the marker with lowest to highest concordance. The red dashed line gives the overall concordance. **b**. Example marker with poor concordance between lobSTR and capillary data due to an annotation error. In this sample, marker TATT031_20 has a genotype of -4,0 reported by lobSTR. However, the capillary data reports -4,4, due to an extra 4bp TCTC indel (blue box) in the flanking regions that is linked with the STR allele 0. Because this indel is not included in the annotated STR sequence (gray box) lobSTR does not consider it when making a genotype call. We visualized the alignment using PyBamView [**?**].

of these samples overlapped between our sequencing and this capillary dataset, and of these, we were able to convert 468 capillary genotyped loci to loci in the lobSTR GRCh37 reference. Capillary genotypes were reported as the size of the PCR product and we converted these to lobSTR format as described on the lobSTR webpage. We rounded all genotypes to the nearest repeat unit. The overall concordance rate was 93%. We compared STR dosage, defined as the sum of lengths of the two alleles, across methods and found strong correlation ($r^2 = 0.92$) between the two datasets (**Figure C**a). In discrepant calls, lobSTR tended to underestimate the true allele length compared to the capillary data, again reflecting a bias toward detecting shorter alleles due to the read length limitation. Notably, the majority of errors originated from a small set of loci (**Figure C.3**a), with many errors potentially due to discrepancies in STR annotations between the datasets. For instance, marker TATT031_20 has a 4bp indel nearby the annotated STR sequence that is strongly linked to particular STR alleles. lobSTR only considers variation within the annotated sequence when making calls, whereas the capillary

calls consider all length variation contained in the product amplified by PCR during genotyping, resulting in discordant genotypes. Thus, both methods are correct by their own definitions, despite the apparent discrepancy. An example discordant call affected by this issue is shown in **Figure C.3b**.

We next sought to assess the accuracy of homopolymers in our data. These markers are not part of the capillary data discussed above and were excluded in previous studies of STR variation [**?**]. To this end, we tested whether the lobSTR calls from these loci could recapitulate known differences between population groups based on principal component analysis (PCA). As a positive control, we first analyzed autosomal tetranucleotides with heterozygosity greater than 10% that were called in at least 90% of samples. These loci represent a relatively high quality STR call-set. The 28,403 tetranucleotides passing the above filters were able to accurately recover known population differences in these samples (**Figure Cb**), with the first principal component separating non-African from African samples and the second primarily separating European and Asian samples. Remarkably, repeating the same analysis with 53,002 homopolymer loci, we were able to recover the majority of the structure seen by tetranucleotides (**Figure Cd**), a testament to the quality of the calls in our catalog for these difficult-to-genotype loci.

C.4 STRs improve resolution of population structure inference

Encouraged by the ability of STR calls to distinguish population structure, we sought to determine whether STRs increase the resolution of population inference beyond that which can be obtained by genome-wide SNPs. We used the smartpca tool from the EIGENSOFT [**?**] package to compute F_{ST} and block jackknife standard errors between all pairs of populations. We first computed F_{ST} and standard errors using SNPs. Genotypes were pulled down for 1,152,838 autosomal sites from a panel of SNPs known to be informative for population structure, built from a union of panels 1 and 2 of an existing dataset [**?**]. We then repeated this analysis using a dataset that combined SNP and STR genotype data. To encode STRs in bi-allelic format, we followed the convention suggested by Patterson et al. [**?**], and encoded each STR allele in the frequency range of 5-95% as a separate bi-allelic marker. This gave 357,863 STR "markers" from 160,530 unique STR loci for a total of 1.51 million markers for the combined SNP+STR analysis. Whereas the two datasets gave highly concordant F_{ST} values (slope of best fit line = 0.96, Pearson $r^2 = 0.999$) (**Figure Cd**), the combined dataset has decreased standard errors compared to SNP variation alone (slope = 0.86), documenting the added value provided by

STRs for discerning population structure (**Figure Ce**).

C.5 Patterns of STR variation

We used our catalog to examine overall trends in polymorphism at STRs. Of the 1.3 million genotyped loci, 32.2% show more than two common alleles (defined as having an allele frequency ≥ 0.01), and some loci have more than 20 common alleles. The remaining loci are either fixed across all individuals (47.6%) or have only two common alleles (20.5%). These patterns changed significantly when stratifying by motif length, with longer motif lengths showing less variability. For instance, only 23% of homopolymers are fixed compared to 70% of tetranucleotides (**Figure C.5**).

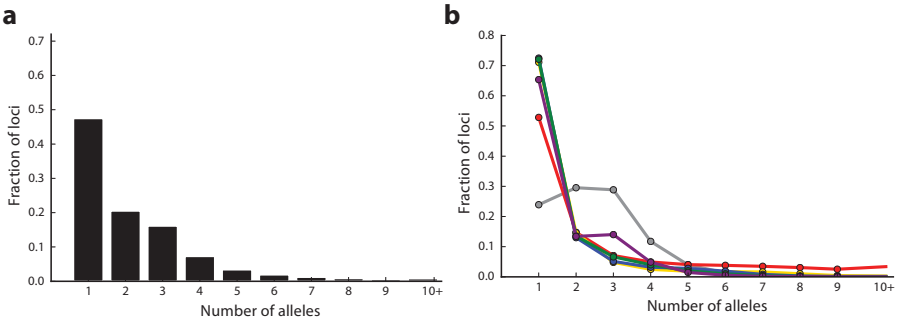


Figure C-4: **Allele frequency spectra of STRs. a**. Distribution of the number of common alleles per locus. **b**. Stratification by motif length (gray=homopolymers, red = dinucleotides, yellow = trinucleotides, blue = tetranucleotides, green = pentanucleotides, purple = hexanucleotides).

As has been previously shown, we found that STR variability depends strongly on properties of the STR itself and on local sequence features. We examined the ability of these features to explain differences in variability for all STR loci with at least two common alleles. We used heterozygosity as a metric of variation, which is defined as $1-\sum_{i=1}^{n} p_i^2$, where p_i is the frequency of allele i and n is the total number of alleles. As mentioned above, heterozygosity tends to decrease with motif length. Additionally, we found that heterozygosity is positively correlated with STR sequence purity ($r = 0.21, p < 10^{-200}$) and reference track length ($r = 0.17, p < 10^{-}200$) (**Figure C.5**). Both these observations agree with previously reported results [**?**, **?**]. We also observed a positive correlation with local recombination rate ($r = 0.028, p = 7.4 \times 10^{-209}$) (deCODE recombination maps [**?**] available on the UCSC genome browser). A joint linear model including

all of these features explained 53% of variation in heterozygosity across loci. When restricting to STRs with no sequence imperfections (interruptions in the STR), these features explained 70% of variation.

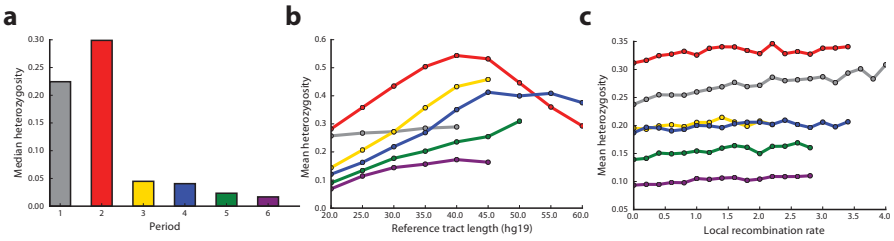


Figure C-5: **Sequence determinants of STR variation. a**. Median heterozygosity by motif length. STRs with longer motif lengths tend to be less polymorphic. **b., c**. Mean heterozygosity as a function of reference track length and local recombination rate (gray = homopolymers, red = dinucleotides, yellow = trinucleotides, blue = tetranucleotides, green = pentanucleotides, purple = hexanucleotides).

C.6 Potential loss-of-function variants at STRs

We used our catalog to identify STRs in coding regions with common loss-of-function (LoF) variants, which we identified as frameshifting variants in coding exons as defined by Refseq. We restricted to alleles found in at least 10 individuals. Seventeen loci with potential common frameshifts passed these criteria, five of which have a frameshift as the major allele (**Table C.6**). Four of the five common LoF alleles with periods 2-6 reported by [**?**]. using an independent dataset are included in our list (*TMEM254, GP6, FAM166B*, and *DCHS2*), and more than half were reported in dbSNP, suggesting that these putative LoF do not represent genotyping errors.

In 13 of the 17 cases, the potential LoF variant occurs in the last exon of the gene or toward the end of a single-exon gene, reducing its potential impact on protein function. The variants in *TMEM254* and *LFNG* occur in an internal exon. In both cases there are alternative transcript annotations that do not contain the affected exons. The putative LoF variants for *PTEN* and *RYK* occur in the first exons of these genes. On visual inspection, the CCG repeat for *RYK* occurs in a difficult-to-align GC-rich area and likely represents an alignment artifact. The variant in *PTEN* is fixed at a 1bp deletion from the reference sequence adjacent to the CGG repeat. This deletion is annotated as a 1bp intron in Refseq ((**Figure C.6**)). Notably this region is not

annotated as coding by Ensembl, Gencode, or UCSC and the frameshift allele is fixed across all samples, suggesting an error in gene annotation. In conclusion, most common STR frameshift variants are unlikely to affect protein function.

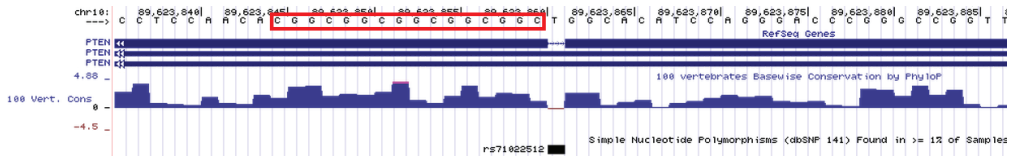


Figure C-6: **The major allele at an STR in *PTEN* is an apparent frameshift from the reference sequence.** The red box denotes the CGG repeat. The 1bp deletion at the adjacent "T" nucleotide is fixed across samples, has poor conservation compared to surrounding bases, and is not annotated as a coding region in other gene annotations, suggesting it may in fact be a misannotation and not a true frameshift variant.

STR Locus	Gene	Motif	LoF	dbSNP
chr13:51530580	RNASEH2B	A	1bp (0.030)	rs200320729 (-/A)
chr14:23528485	ACIN1^{+}	AGAGGG	-2bp (0.030)	
chr10:81841429	TMEM254*	AAAG	-4bp (0.034)	rs143538725 (-/AAAG)
chr3:133969414	RYK^{+}	CCG	1bp (0.036)	
chr15:83677271	C15orf40	A	1bp (0.078)	
chr19:55526092	GP6*	ACAG	4bp (0.093)	rs138680589 (-/CAGA)
chr5:147861098	HTR4^{+}	AAAAAG	1bp, -1bp (0.095)	
chr12:55820959	OR6C76	A	-1bp (0.218)	
chr20:3026346	GNRH2	CCCCG	5bp (0.320)	
chr16:58577316	CNOT1	A	-1bp (0.367)	
chr9:35561913	FAM166B*	ACCC	1bp, -8bp (0.402)	rs143266743 (-/CCCACCCT)
chr7:2552851	LFNG	ATCC	4bp, -4bp (0.422)	
chr6:31380147	MICA	AGC	-1bp, -4bp, 2bp, 11bp (0.810)	rs547446871 (-/G)
				rs41293539 (-/CT/CTGCTGCT/CTGCTGCTGCT)
chr4:155244402	DCHS2*	AAAC	-4bp (0.846)	rs140019361 (-/TTTG)
chr10:125780753	CHST15	C	-1bp (0.895)	rs5788645 (-/C)
chr10:89623845	PTEN	CCG	-1bp (1.000)	rs71022512 (-/A)
chr5:72743281	FOXD1	CCG	2bp (1.000)	rs587745355 (-/GC)

Table C.2: **Common loss-of-function alleles at STRs.** We give the combined allele frequencies of all frameshift alleles for each locus. dbSNP data is from versions 141 and 142. * entries are LoF alleles previously reported by Willems, et al. $^{+}$ entries are low confidence alleles likely due to alignment artifacts or stutter errors.

C.7 Conclusion

We have presented the highest quality catalog of STR variation to date, which can serve as a gold standard reference panel of STR polymorphisms across diverse populations. Additionally, our dataset provides unprecedented opportunities to study STR variation that were not possible using previous studies either due to the small number of markers or to the low quality of individual genotypes5. Importantly, it contains the first panel of previously inaccessible homopolymer genotypes and allows in-depth study of these extremely polymorphic loci for the first time. We envision that this dataset will be an invaluable resource for future studies of STR polymorphism.

Bibliography