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

Shubham Saini[1,2], Ileena Mitra[2,3], Melissa Gymrek[1,2,*]

[1] Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA USA
[2] Department of Medicine, University of California San Diego, La Jolla, CA USA
[3] Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, CA, USA.

* Correspondence should be addressed to mgymrek@ucsd.edu

Abstract

Short tandem repeats (STRs) are involved in dozens of Mendelian disorders and have been implicated in a variety of complex traits in humans. However, existing technologies have not allowed for systematic STR association studies. Genotype array data is available for hundreds of thousands of samples, but is limited to variation in common single nucleotide polymorphisms (SNPs) and does not adequately capture more complex variants like STRs. Here, we leverage next-generation sequencing from 479 families along with existing bioinformatics tools to phase STRs onto SNP haplotypes and create a genome-wide reference haplotype panel. Imputation using our panel achieved an average of 97% concordance between true and imputed STR genotypes in an external dataset and could accurately recover repeat lengths at known pathogenic loci. Imputed STRs capture on average 20% more variation in STR allele length with increased power to detect underlying STR associations compared to individual common SNPs, highlighting a limitation of standard genome-wide association studies. Our framework will enable testing for STR associations with hundreds of traits across massive sample sizes without the need to generate additional data.

Introduction

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

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

Short tandem repeats (STRs), consisting of repeated motifs of 1-6bp in tandem, comprise more than 3% of the human genome[9]. Multiple lines of evidence support a role of STRs in complex traits[10,11], particularly in neurological and psychiatric phenotypes. STRs are one of the largest sources of genetic variation in humans[12,13], and play a significant role in regulating gene expression[14,15] and splicing[16–18]. Intriguingly, more than 30 Mendelian disorders are caused by STR expansions with a range of mechanisms, including polyglutamine aggregation (Huntington's Disease, ataxias[19]), hypermethylation (Fragile X Syndrome[20]), and RNA toxicity (ALS/FTD[21]). Furthermore, causal STRs driving existing GWAS signals have already been identified[22]. Yet, STRs are often in weak LD with SNPs[12], severely limiting the power of standard GWAS to detect underlying STR associations.

Existing technologies have not allowed for systematic STR association studies. Next-generation sequencing (NGS) can be used to directly genotype short STRs, but NGS is still too expensive

to perform on sufficiently large cohorts for GWAS of most complex traits. An alternative approach is to impute STRs into existing SNP array datasets. However, statistical phasing of STRs and SNPs is challenging for several reasons: STRs and SNPs have diminished LD due to the rapid mutation rates[13,23] and high prevalence of recurrent mutations in STRs. As a result, the relationship between STR repeat number and SNP haplotype can be complicated and nonlinear, with the same STR allele present on multiple SNP haplotypes and vice versa. Finally, STRs are prone to genotyping errors induced during PCR amplification[24], further ambiguating phase information.

Sequencing related samples allows haplotype resolution by directly tracing inheritance patterns. The recent generation of deep NGS using PCR-free protocols for hundreds of nuclear families in combination with accurate tools for genotyping STRs[25] from NGS now enables applying this technique genome-wide. Here, we profiled STRs in 479 families and used pedigree information to phase STR genotypes onto SNP haplotypes to create a genome-wide reference for imputation. We used this panel to impute STRs into an external dataset of similar ethnic background with 97% concordance compared to observed STR genotypes. Notably, imputed genotypes at highly polymorphic STRs previously implicated in human disorders were highly correlated with observed genotypes across a large range of allele lengths. We show that STR imputation captures on average 20% more variation in STR allele lengths than the best tag SNP, resulting in greatly improved power over standard GWAS to detect associations due to underlying STRs.

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

Results

A catalog of STR variation in 479 families

We first generated the deepest catalog of STR variation to date in a large cohort of families included in the Simons Simplex Collection (SSC) (see **URLs**). We focused on 1,916 individuals from 479 family quads (parents and two children) with mostly European origins (**Supplementary Figure 1**) that were sequenced to an average depth of 30x using Illumina's PCR-free protocol. We used HipSTR[25] to profile autosomal STRs in each sample. To maximize the quality of

genotype calls, individuals were genotyped jointly using HipSTR's multi-sample calling mode using phased SNP genotypes and aligned reads as input (**Methods**). Multi-sample calling allows HipSTR to leverage information on haplotypes discovered across all samples in the dataset to estimate per-locus error parameters and output genotype likelihoods for each possible diploid genotype. An average of 1.14 million loci were profiled and passed HipSTR's default filtering settings in each sample (**Figure 1A**). We obtained at least one call for 97% of all loci in our reference of 1.6 million STRs with an average call rate of 90% (**Figure 1B**).

We applied additional stringent genotype quality filters to ensure accurate calls for downstream phasing and imputation analysis. Loci overlapping segmental duplications, with call rates less than 80%, or with genotype frequencies unexpected under Hardy-Weinberg Equilibrium were removed (**Methods**). We further removed loci with low heterozygosity (<0.095) to restrict analysis to polymorphic STRs. We found that these filters increased the quality of our calls, as evidenced by the average Mendelian inheritance rate of 99.8% and 97.9% at loci that passed and failed quality filters, respectively (**Figure 1C**). Notably, most known STRs implicated in expansion disorders are excluded from our dataset as they are too long to be spanned using Illumina reads and thus could not be genotyped by HipSTR. After filtering, 453,671 loci remained in our catalog.

To further assess the quality of our calls, we compared STR genotypes from the SSC to a catalog of STR variation[12] previously generated from the 1000 Genomes Project[26] data using lobSTR[27]. We found that the per-locus heterozygosities were highly concordant (r=0.96; $p<10^{-200}$; n=386,100) (**Figure 1D**), despite being generated from orthogonal datasets using distinct STR algorithms. Overall, these results show that our catalog consists of robust STR genotypes suitable for downstream phasing and imputation analysis.

A genome-wide SNP/STR haplotype reference panel

We examined the extent of linkage disequilibrium between STRs and nearby SNPs using two metrics. The first, termed "length r^2", is defined as the squared Pearson correlation between STR allele length and the SNP genotype. The second, termed "allelic r^2", treats each STR allele as a separate bi-allelic locus and is computed similar to traditional SNP-SNP LD (**Methods**). As expected, SNP-STR LD was dramatically weaker than SNP-SNP LD by both metrics (**Supplementary Figure 2**) with length r^2 generally stronger than allelic r^2. On the other hand,

nearly all STRs were in significant LD (Length r^2 $p<0.05$) with at least one nearby SNP suggesting that phasing would result in informative haplotypes.

We developed a pipeline to phase STRs onto SNP haplotypes leveraging the quad family structure (**Figure 2A**). Based on our LD analysis, we used a window size of ±50kb to phase each STR separately using Beagle[28], which was recently demonstrated to perform well in phasing STRs[29]. Beagle is able to both handle multi-allelic loci as well as incorporate pedigree information, which is not supported by competing phasing algorithms[30–32]. Resulting phased haplotypes from the parent samples were merged into a single genome-wide reference panel for downstream imputation.

We first evaluated the quality of our phased panel using a "leave-one-out" analysis in the SSC samples. For each sample, we constructed a modified reference panel with that sample's haplotypes removed and then performed genome-wide imputation. Imputed genotypes showed an average of 96.7% concordance with genotypes obtained by HipSTR (**Table 1**) with weakest performance at STRs with highest heterozygosity (**Figure 2B, Supplementary Figure 3**). As a test case, we examined per-locus imputation performance at the CODIS STRs used in forensics analysis (**Supplementary Table 1**). These markers are extremely polymorphic with an average 11.6 alleles each, and thus are representative of the most difficult loci to impute. We achieved an average concordance of 70%, with per-locus values slightly higher than those reported by a previous study by Edge, *et al*[29] likely as a result of our larger and more homogenous cohort. On the other hand, average concordance at STRs with 6 or fewer alleles was 99%, showing that even highly multi-allelic loci can be accurately imputed. We additionally computed the length r^2 and allele r^2 for each locus. As expected, length r^2 was strongest for loci with fewer alleles (**Supplementary Figure 4**) and allele r^2 was strongest for the most common alleles (**Figure 2C**). Per-locus imputation statistics are reported in **Supplementary Tables 2** and **3**).

To test our ability to impute STRs into an external dataset, we imputed STR genotypes into SNP genotypes available from the 1000 Genomes Project[26] from three different platforms: low coverage whole genome sequencing (WGS), and the Affymetrix 6.0 and Omni 2.5 genotyping arrays. We then focused on 150 samples who were also sequenced to 30x genome-wide coverage by Illumina (see **URLs**). Samples originated from multiple population backgrounds, allowing us to evaluate our panel in non-European samples. In parallel, we used HipSTR to

profile STRs from the WGS and used our panel to impute STR genotypes using the available SNP datasets. Per-locus concordance, length r^2, and allele r^2 were highly concordant between the SSC panel and 1000 Genomes samples of European origin (Pearson r=0.94, 0.63, and 0.85, respectively using genotypes imputed from WGS) (**Figure 2D, E**; **Supplementary Figure 5, Table 1**). Imputation performance did not vary when using phased genotypes obtained from WGS vs. Omni2.5 for imputation (**Supplementary Table 4**). Average concordance and length r^2 were weakest when using genotypes from Affy6.0 chips, although fewer samples were available for comparison. Concordance was noticeably weaker in African and East Asian samples, likely due to different population background compared to the SSC samples and consistent with observations from SNP imputation[26].

Imputation increases power to detect STR associations

We sought to determine whether our SNP-STR haplotype panel could increase power to detect underlying STR associations over standard GWAS. To this end, we simulated phenotypes based on a single causal STR and examined the power of the imputed STR genotypes vs. nearby SNPs to detect associations. We focused primarily on a linear additive model relating STR allele lengths to quantitative phenotypes (**Figure 3A**), since the majority of known functional STRs follow similar models (*e.g.*[14,16,33,34]). Association testing simulations were performed 100 times for each STR on chromosome 21 in our dataset (**Methods**). The strength of association for each variant as measured by the negative $\log_{10}$ p-value was linearly related with its length r^2 with the causal variant (**Figure 3B**) as has been previously demonstrated[35]. On average, the imputed STR genotypes explained 20% more variation in STR allele length compared to the best tag SNP. The advantage from STR imputation grew as a function of the number of common STR alleles (**Supplementary Figure 6**). Imputed genotypes showed a corresponding increase in power to detect associations (**Figure 3C**). Similar trends were observed for case-control traits (**Supplementary Figure 7**).

We additionally tested the ability of imputed STR genotypes to identify associations due to non-linear models relating STR genotype to phenotype. Several such models have been previously observed: for instance, STR expansion disorders follow a threshold model under which only long alleles have pathogenic effects, and several STRs acting as expression modifiers in yeast show bell-shaped associations in which moderate allele lengths are optimal[36]. We simulated quantitative phenotypes under a quadratic model where either extremely short or

long alleles conferred the highest risk (**Supplementary Figure 8A**). As expected, testing for linear association between allele length and phenotype was often underpowered compared to SNP-based tests (**Supplementary Figure 8B**). On the other hand, per-allele association tests in which each STR allele was treated as a separate bi-allelic model performed at least as well as the best SNP in all cases (**Supplementary Figure 8C**). Importantly, the underlying model relating STR length to phenotype is not known *a priori* for association studies and tests based on the true model will show maximum power. For more complex models such tests will only be possible when allele lengths are available, thus demonstrating an additional advantage of STR imputation over SNP-based tests to detect these associations.

Phasing and imputing normal alleles at known pathogenic STRs

Finally, to determine whether alleles at known pathogenic STRs could be accurately imputed, we examined results of our imputation pipeline at seven loci previously implicated in STR expansion disorders that were included in our panel (**Table 2**). Our analysis focused on alleles in the normal repeat range for each locus, since pathogenic repeat expansions are both beyond the range that can be genotyped by HipSTR and are unlikely to be present in the SSC cohort. Notably, accurate imputation of non-pathogenic allele ranges is still informative as (1) long normal or intermediate size alleles may result in mild symptoms in some expansion disorders[37,38,39] (2) longer alleles are more at risk for expansion[40] and (3) allele lengths below the pathogenic range could potentially be associated with more complex phenotypes[38].

Similar to the CODIS markers, these loci are highly polymorphic with 10 or more alleles per locus. In all cases, imputed genotypes were more strongly correlated with HipSTR genotypes compared to the best tag SNP. Visualization of SNP-STR haplotypes at the CAG repeat implicated in dentatorubral-pallidoluysian atrophy (DRPLA)[41] reveals a typical complex relationship between STR allele length and local SNP haplotype (**Figure 4A**), with the same STR allele often present on multiple SNP haplotype backgrounds. Still, for most loci there is a clear association of specific haplotypes with different allele length ranges allowing accurate imputation across a large range of allele sizes (**Figure 4B, Supplementary Figure 9**).

Resolution of SNP-STR haplotypes can be used to infer the mutation history of a specific STR locus. Notably, for many STR expansion orders it has been shown that pathogenic expansion alleles originated from a founder haplotype[42–45] associated with a long allele. We compared SNP

haplotypes at the DRPLA locus in our dataset to a previously reported founder haplotype[45]. In concordance with the hypothesis of a single founder haplotype, we found that SNP haplotypes with smaller Hamming distance to the known founder haplotype had longer CAG tracts (r=-0.79; $p<10^{-200}$). This finding demonstrates that while we were unable to directly impute pathogenic expansion alleles, STR imputation can accurately identify which individuals are at risk for carrying expansions or pre-pathogenic mutations and the inferred haplotypes can reveal the history by which such mutations arise.

Discussion

Our study combines available whole genome sequencing datasets with existing bioinformatics tools to generate the first phased SNP/STR haplotype panel allowing genome-wide imputation of STRs into SNP data. Despite their exceptionally high rates of polymorphism, we demonstrate that the majority of polymorphic STRs in the genome can be imputed to high accuracy. We additionally show that imputation greatly improves power to detect STR associations over standard SNP-based GWAS.

A widely recognized limitation of GWAS is the fact that common SNP associations still explain only a small fraction of heritability of most traits. Multiple explanations for this have been proposed, including minute effect sizes of individual variants and a potential role for high-impact rare variation[46]. However, studies in large cohorts reaching hundreds of thousands of samples[1–3], as well as deep sequencing studies to detect rare variants[47], have so far not confirmed these hypotheses. An increasingly supported idea is that complex variants not well tagged by SNPs may comprise an important component of the "missing heritability."[10,11] GWAS is essentially blind to contributions from highly polymorphic STRs and other repeats, despite their known importance to human disease and molecular phenotypes. Thus STR association studies will undoubtedly uncover additional heritability that is so far unaccounted for.

Our initial haplotype panel faces several important limitations. First, the majority of samples are of European origin, limiting imputation accuracy in other population groups. Second, imputation accuracy is mediocre for the most highly polymorphic loci, some of which will ultimately have to be directly genotyped to adequately test for associations. Notably, our work relied on existing tools originally designed for SNP imputation. In future work computational methods built specifically for imputing repeats may be able to improve performance. Importantly, long STRs

are missing from our panel due to the limitation imposed by short read lengths. However, new tools have recently been developed for genotyping expanded STRs[48,49] and longer variable number tandem repeats (VNTRs)[50] from short reads. In future work, genotypes obtained from these tools can be used to extend our panel to include additional variants. Finally, our study relied on simulated phenotypes to measure the gain in power of imputation over GWAS. Notably, while autism phenotypes are available for the SSC families, this cohort is too small to perform a GWAS and was specifically ascertained for families enriched for *de novo*, rather than inherited, pathogenic mutations. In future work our panel can be applied to impute STRs into larger cohorts for autism and other complex traits.

Overall, our STR imputation framework will enable an entire new class of variation to be interrogated by reanalyzing hundreds of thousands of existing datasets, with the potential to lead to novel genetic discoveries across a broad range of phenotypes.

URLs

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

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

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

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

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

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

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

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

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

Acknowledgements

Research reported in this publication was supported in part by the Office Of The Director, National Institutes of Health under Award Number DP5OD024577 and by a SFARI Explorer Award Number 515568. Access to SSC data was approved for this project under request id 2405.1.1. M.G. was supported in part by NIH/NIMH grant R01 MH113715. This work used the

Extreme Science and Engineering Discovery Environment (XSEDE) comet resource at the San Diego Supercomputing Center through allocations ddp268 and csd568. XSEDE is supported by National Science Foundation grant number ACI-1548562. We thank Nima Mousavi and Alon Goren for helpful comments on the manuscript. We additionally thank Vineet Bafna and Vikas Bansal for helpful discussions and providing access to compute resources. We are grateful to all of the families that participated in the Simons Simplex Collection as well as the principal investigators.

Author Contributions

M.G. conceived the study, helped design and perform analyses, and drafted the initial manuscripts. S.S. generated the reference haplotype panel, performed downstream analyses, and participated in writing the manuscript. I.M. performed simulation analyses. All authors have read and approved the final manuscript.

Competing and Financial Interests

The authors have no competing financial interests to disclose.

Figure Legends

Figure 1: A deep catalog of STR variation in the SSC cohort. A. Number of loci called per sample. Dashed line represents the mean of 1.14 million STRs per sample. **B. Call rate per locus.** Dashed line represents the mean call rate of 90%. **C. Mendelian inheritance rate at filtered vs. unfiltered loci.** The x-axis gives the posterior genotype score (Q) returned by HipSTR. The y-axis gives the average Mendelian inheritance rate for each bin across all calls on chromosome 21. Loci that were homozygous for the reference allele in all members of a family were removed. Colors represent different motif lengths. **D. Per-locus heterozygosity in SSC vs. 1000 Genomes.** Dashed red line gives the diagonal line.

Figure 2: Creating a reference SNP-STR haplotype panel. A. Schematic of phasing pipeline in the SSC cohort. To create the phased panel, STR genotypes were placed onto phased SNP haplotypes using Beagle. Any missing STR genotypes were imputed. The resulting panel was then used for downstream imputation from orthogonal SNP genotypes. Black and red denote phased and unphased variants, respectively. Positions in gray are homozygous. **B. Concordance of imputed STR genotypes vs. heterozygosity. C. Distribution of allelic r^2 of**

imputed vs. true STR genotypes as a function of allele frequency. B. and **C.** are based on the leave one out analysis in the SSC cohort. Boxes give the interquartile range, horizontal lines give medians, and black circles show outlier data points. **D. Per-locus imputation concordance in SSC vs. 1000 Genomes cohorts. E. Per-locus allelic r^2 in the SSC vs. 1000 Genomes cohorts. D, E.** are based on the subset of samples from the 1000 Genomes deep WGS cohort with European ancestry. Dashed blue lines give the best fit. Dashed black lines give the diagonal.

Figure 3: STR imputation improves power to detect STR associations. A. Example simulated quantitative phenotype based on SSC genotypes. A quantitative phenotype was simulated assuming a causal STR (red). Power to detect the association was compared between the causal STR, imputed STR genotypes, and all common SNPs (MAF>0.05) within a 50kb window of the STR. **B. Strength of association (-log10 p) is linearly related with LD with the causal variant.** For SNPs, the x-axis gives the length r^2 calculated using observed genotypes. For the imputed STR, the x-axis gives the length r^2 from leave-one-out-analysis. **C. The gain in power using imputed genotypes is linearly related to the gain in r^2 compared to the best tag SNP**. Each dot represents a single locus. Power was calculated based on the number of simulations out of 100 with nominal p-value < 0.05.

Figure 4: SNP haplotypes distinguish allele lengths at known pathogenic loci. A. Example SNP-STR haplotypes inferred in 1000 genomes European samples at the *ATN1* locus implicated in DRPLA. Each column represents a SNP from the founder haplotype reported by Veneziano, *et al.* Each row represents a single haplotype, with gray and black boxes denoting major and minor alleles, respectively. Haplotypes are grouped by the corresponding STR allele. STR alleles with inferred counts of less than 10 were excluded from the visualization. **B. Comparison of imputed vs. genotyped repeat dosages in SSC samples a the *ATN1* locus.** The x-axis gives the maximum likelihood genotype dosage returned by HipSTR. The y-axis gives the imputed dosage. Dosage is defined as the sum of the two allele lengths of each genotype relative to the hg19 reference genome. The bubble size represents the number of samples summarized by each data point. **C. Distribution of *ATN1* allele length vs. Hamming distance from the founder haplotype reported by Veneziano, *et al.*** White dots represent the median length.

Tables

	Concordance	Length r^2	Allelic r^2
SSC - LOO	96.7%	0.906	0.861
1000 Genomes - EUR	97.0%	0.921	0.871
1000 Genomes - AFR	90.6%	0.746	0.700
1000 Genomes - EAS	93.8%	0.823	0.773

Table 1: Imputation performance summary. Results indicate mean across all loci analyzed. Allele r^2 values include all alleles with minor allele frequency at least 5%.

Locus	Motif	Disorder	r^2 LOO	Conc. LOO	Best tag SNP	$r^2_{bestSNP}$	# alleles
3:63898361	CAG	SCA7[a]	0.75	0.92	rs58676857	0.57	10
5:146258291	CAG	SCA12[a]	0.88	0.94	rs2082405	0.64	14
12:7045880	CAG	DRPLA[b]	0.79	0.80	rs34199021	0.68	19
12:112036754	CAG	SCA2[a]	0.30	0.94	rs141683900	0.27	13
14:92537353	CAG	SCA3[a]	0.86	0.87	rs2402108	0.72	20
16:87637889	CAG	HDL2[c]	0.55	0.88	rs2434850	0.34	15
19:46273457	CAG	DM1[d]	0.88	0.85	rs2070737	0.59	25

Table 2: Imputation performance at known pathogenic repeats. [a]SCA=spinocerebellar ataxia. [b]DRPLA=Dentatorubral-pallidoluysian Atrophy. [c]Huntingon's Disease-Like 2. [d]Myotonic Dystrophy Type 1.

Online Methods

Phasing SNPs in the SSC

SNP genotypes were called from gVCF files obtained through SFARI base (see **URLs**) using the GATK version 3 joint calling pipeline[51]. Variants overlapping sites reported in the 1000

Genomes Project[26] phase 3 data were retained for downstream analysis. SNP genotypes were phased using SHAPEIT[30] version 2.r837 with 1000 Genomes Phase 3 genotypes as a reference panel and ignoring pedigree information. SHAPEIT's duoHMM[52] version 0.1.7 method was used to refine phased haplotypes using pedigree structure and correcting for Mendelian errors.

Genome-wide multi-sample STR genotyping

Aligned BAM files for whole genome sequencing data of individuals from the SSC Phase I collection were obtained through SFARI base (see **URLs**) and processed using Amazon Web Services (AWS). STRs were jointly genotyped on the AWS EC2 platform in batches of 500 loci. We streamed the corresponding region of each BAM file and of the phased SNP VCF files to a local EBS volume attached to each EC2 instance using samtools[53] version 1.4 and tabix[54] version 1.2, respectively. HipSTR[25] version v0.5 was called individually per locus with default parameters. Phased SNPs were provided as input to allow HipSTR to perform physical phasing when possible. Resulting VCF files from each batch were merged to create a genome-wide callset in VCF format.

Filtering STR genotype calls

STR calls were filtered using the filter_vcf.py script in the HipSTR package with suggested parameters (--min-call-qual 0.9 --max-call-flank-indel 0.15 --max-call-stutter 0.15). We used the following criteria to remove problematic loci from the callset: *(i)* STR loci overlapping segmental duplications (UCSC Table Browser[55] hg19.genomicSuperDups table) were removed from the callset using intersectBed[56] v2.25.0; *(ii)* Pentanucleotides and hexanucleotides containing homopolymer runs of at least 5 or 6 nucleotides, respectively, in the hg19 reference genome were removed as they were found to contain an excess of indels in the homopolymer regions; *(iii)* loci with call rate <80%; *(iv)* loci with heterozygosity <0.095, corresponding to a minor allele frequency of 5% for biallelic markers, were removed to restrict to polymorphic STRs; *(v)* loci with significantly more or fewer heterozygous genotypes compared to the expectation under Hardy-Weinberg equilibrium ($p<0.01$) as described previously[57,58].

Comparison to 1000G catalog

STRs for 1000 Genomes samples as described in Willems *et al.*[12] were downloaded from the strcat site (see **URLs**). Heterozygosity was computed using the PyVCF package (see **URLs**) for the 1000 Genomes calls and using a custom script for the SSC data to collapse alleles of

identical length into a single allele. Loci passing all filters described above except the heterozygosity filter were included in the comparison. Analysis was restricted to loci with at least 500 calls in the 1000 Genomes dataset.

Phasing STRs

Beagle version 4.0[28] was used to phase each STR separately using phased SNP genotypes, pedigree information, and unphased STR genotypes as input. In order to leverage the HipSTR genotype likelihoods (GL field), Beagle requires all samples to have GL information. To accommodate this, phasing was performed in two steps. First, samples with missing data were removed and the remaining samples were phased using the "-gl" Beagle flag. Next, missing samples were added back to the VCF and all samples were jointly phased in a second Beagle round using default parameters. In this step Beagle additionally imputed any calls with missing genotypes. Phased STRs and SNPs for only the unrelated parent samples from each locus were then merged into a single genome-wide reference panel in VCF format.

Imputation performance metrics

Let $X = \{x_1, x_2, \ldots x_n\}$ be the true STR genotypes for samples $1..n$ and $Y = \{y_1, y_2, \ldots y_n\}$ be the imputed STR genotypes. Each genotype x_i is defined as $x_i = (x_{i1}, x_{i2})$ where x_{i1} and x_{i2} give the (unordered) lengths of the two STR alleles for a diploid sample and similarly for Y. We then define the following metrics:

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

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

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

For all concordance metrics, outlier genotypes containing alleles seen less than 3 times in the entire cohort were removed from the analysis.

Evaluation in the 1000 Genomes data

STRs were jointly genotyped in 150 high-coverage WGS datasets that were also profiled by the 1000 Genomes Project (see **URLs**) using HipSTR version 0.6 followed by the filtering steps described above for the SSC cohort. Separately, STRs were imputed into SNP data downloaded from the 1000 Genomes Project site from three sources (WGS, phased SNPs from Affy6.0 array, and phased SNPs from Omni2.5 array, see **URLs**) with Beagle using the SSC SNP-STR haplotype panel.

Simulations for power analysis

We analyzed parental genotypes for 5,838 STRs across chromosome 21 that passed filtering and quality control as described above. For each STR, we simulated quantitative phenotype datasets under the model: $P = \beta G + E$, where P is a vector of standard normalized phenotypes, β gives the effect size, E gives the error term drawn from a normal distribution $N(0, 1-\beta)$, and G is a vector of the sum of genotype lengths for each individual scaled to have mean 0 and variance 1. For each simulated phenotype dataset, we tested the causal STR, the imputed STR genotypes, and the best tag SNP (strongest length r^2) within 50kb of the STR for association. Association tests were performed using the Python statsmodels library OLS method (see **URLs**).

We performed additional simulations under a case control model shown in **Supplementary Figure 7**. Phenotypes (0=control, 1=case) were drawn for each sample according to the model $logit(p_i) = \beta X_i$ where p_i is the probability that sample i is a case and X_i is the scaled genotype for individual i as described above. Association tests were performed using the Python statsmodels Logit method.

For the non-additive phenotype example, we performed simulations under a quadratic model: $P = \beta G^2 + E$ where G is a vector of the squared sum of allele lengths scaled by the mean allele length, and P, β, E are as described above. Two sets of association tests were performed: the first tested for association between STR length and phenotype (**Supplementary Figure 8B**) and the second set performed a separate association test for each STR allele treating the allele as a bi-allelic locus (**Supplementary Figure 8C**).

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

Comparison to DRPLA founder haplotypes

The founder haplotype for the expansion allele in *ATN1* implicated in DRPLA was taken from Table 1 of Veneziano *et al.*[45] and consists of rs4963516, rs1007924, rs7310941, rs7303722, rs2239167, rs34199021, rs2071075, rs2071076, and rs2159887 with hg19 alleles G, A, G, T, A, A, T, C, and C respectively. Distance from the founder haplotype was calculated as the number of mismatches.

Data Availability

Phased SNP-STR haplotypes for 1000 Genomes Project phase 3 samples and example commands for imputation are available at https://gymreklab.github.io/2018/03/05/snpstr_imputation.html. Upon acceptance for publication STR genotypes and phased SNP-STR haplotypes for the SSC samples will be made available at https://base.sfari.org/.

Code Availability

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

References

1. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).
2. Turcot, V. *et al.* Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. *Nat. Genet.* **50,** 26–41 (2018).
3. Scott, R. A. *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* **66,** 2888–2902 (2017).
4. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).
5. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547,** 173–178 (2017).

6. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530,** 177–183 (2016).
7. Boettger, L. M. *et al.* Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. *Nat. Genet.* **48,** 359–366 (2016).
8. Leffler, E. M. *et al.* Resistance to malaria through structural variation of red blood cell invasion receptors. *Science* **356,** (2017).
9. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
10. Press, M. O., Carlson, K. D. & Queitsch, C. The overdue promise of short tandem repeat variation for heritability. *Trends Genet.* **30,** 504–512 (2014).
11. Hannan, A. J. Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* (2018). doi:10.1038/nrg.2017.115
12. Willems, T. *et al.* The landscape of human STR variation. *Genome Res.* **24,** 1894–1904 (2014).
13. Gymrek, M., Willems, T., Reich, D. & Erlich, Y. Interpreting short tandem repeat variations in humans using mutational constraint. *Nat. Genet.* **49,** 1495–1501 (2017).
14. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48,** 22–29 (2016).
15. Quilez, J. *et al.* Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res.* **44,** 3750–3762 (2016).
16. Hefferon, T. W., Groman, J. D., Yurk, C. E. & Cutting, G. R. A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proceedings of the National Academy of Sciences* **101,** 3504–3509

(2004).

17. Hui, J. *et al.* Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. *EMBO J.* **24,** 1988–1998 (2005).
18. Tollervey, J. R. *et al.* Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat. Neurosci.* **14,** 452–458 (2011).
19. Mirkin, S. M. Expandable DNA repeats and human disease. *Nature* **447,** 932–940 (2007).
20. Sutcliffe, J. S. *et al.* DNA methylation represses FMR-1 transcription in fragile X syndrome. *Hum. Mol. Genet.* **1,** 397–400 (1992).
21. van Blitterswijk, M., DeJesus-Hernandez, M. & Rademakers, R. How do C9ORF72 repeat expansions cause amyotrophic lateral sclerosis and frontotemporal dementia: can we learn from other noncoding repeat expansion disorders? *Curr. Opin. Neurol.* **25,** 689–700 (2012).
22. Grünewald, T. G. P. *et al.* Chimeric EWSR1-FLI1 regulates the Ewing sarcoma susceptibility gene EGR2 via a GGAA microsatellite. *Nat. Genet.* **47,** 1073–1078 (2015).
23. Sun, J. X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44,** 1161–1165 (2012).
24. Gymrek, M. A genomic view of short tandem repeats. *Curr. Opin. Genet. Dev.* **44,** 9–16 (2017).
25. Willems, T. *et al.* Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* (2017). doi:10.1038/nmeth.4267
26. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
27. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.* **22,** 1154–1162 (2012).
28. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data

inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).

29. Edge, M. D., Algee-Hewitt, B. F. B., Pemberton, T. J., Li, J. Z. & Rosenberg, N. A. Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 5671–5676 (2017).
30. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9,** 179–181 (2011).
31. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44,** 955–959 (2012).
32. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. (2015). doi:10.1101/028282
33. Contente, A., Dittmer, A., Koch, M. C., Roth, J. & Dobbelstein, M. A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nat. Genet.* **30,** 315–320 (2002).
34. Shimajiri, S. *et al.* Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS Lett.* **455,** 70–74 (1999).
35. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2014).
36. Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M. & Verstrepen, K. J. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324,** 1213–1216 (2009).
37. Wheeler, A. C. *et al.* Associated features in females with an FMR1 premutation. *J. Neurodev. Disord.* **6,** 30 (2014).
38. Ha, A. D., Beck, C. A. & Jankovic, J. Intermediate CAG Repeats in Huntington's Disease:

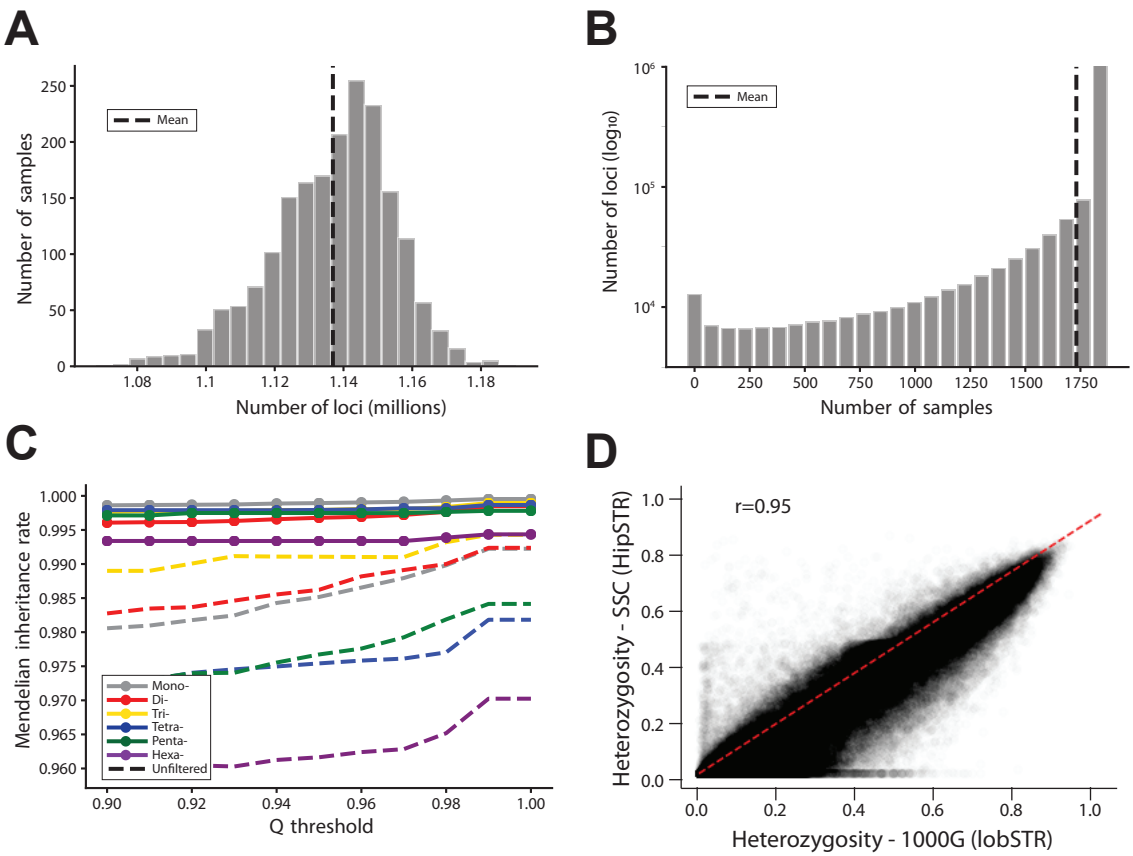
Analysis of COHORT. *Tremor Other Hyperkinet. Mov.* **2,** (2012).

39. Brenman, L. M. Spinocerebellar Ataxia Type 6 (SCA6) Phenotype in a Patient with an Intermediate Mutation Range CACNA1A Allele. *J. Neurol. Neurophysiol.* **04,** (2013).
40. Lee, D.-Y. & McMurray, C. T. Trinucleotide expansion in disease: why is there a length threshold? *Curr. Opin. Genet. Dev.* **26,** 131–140 (2014).
41. Koide, R. *et al.* Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). *Nat. Genet.* **6,** 9–13 (1994).
42. Gan, S.-R., Ni, W., Dong, Y., Wang, N. & Wu, Z.-Y. Population genetics and new insight into range of CAG repeats of spinocerebellar ataxia type 3 in the Han Chinese population. *PLoS One* **10,** e0134405 (2015).
43. Paradisi, I., Ikonomu, V. & Arias, S. Huntington disease-like 2 (HDL2) in Venezuela: frequency and ethnic origin. *J. Hum. Genet.* **58,** 3–6 (2013).
44. Laffita-Mesa, J. M. *et al.* De novo mutations in ataxin-2 gene and ALS risk. *PLoS One* **8,** e70560 (2013).
45. Veneziano, L. *et al.* A shared haplotype for dentatorubropallidoluysian atrophy (DRPLA) in Italian families testifies of the recent introduction of the mutation. *J. Hum. Genet.* **59,** 153–157 (2014).
46. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13,** 135–145 (2012).
47. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536,** 41–47 (2016).
48. Tang, H. *et al.* Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *Am. J. Hum. Genet.* **101,** 700–715 (2017).
49. Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome

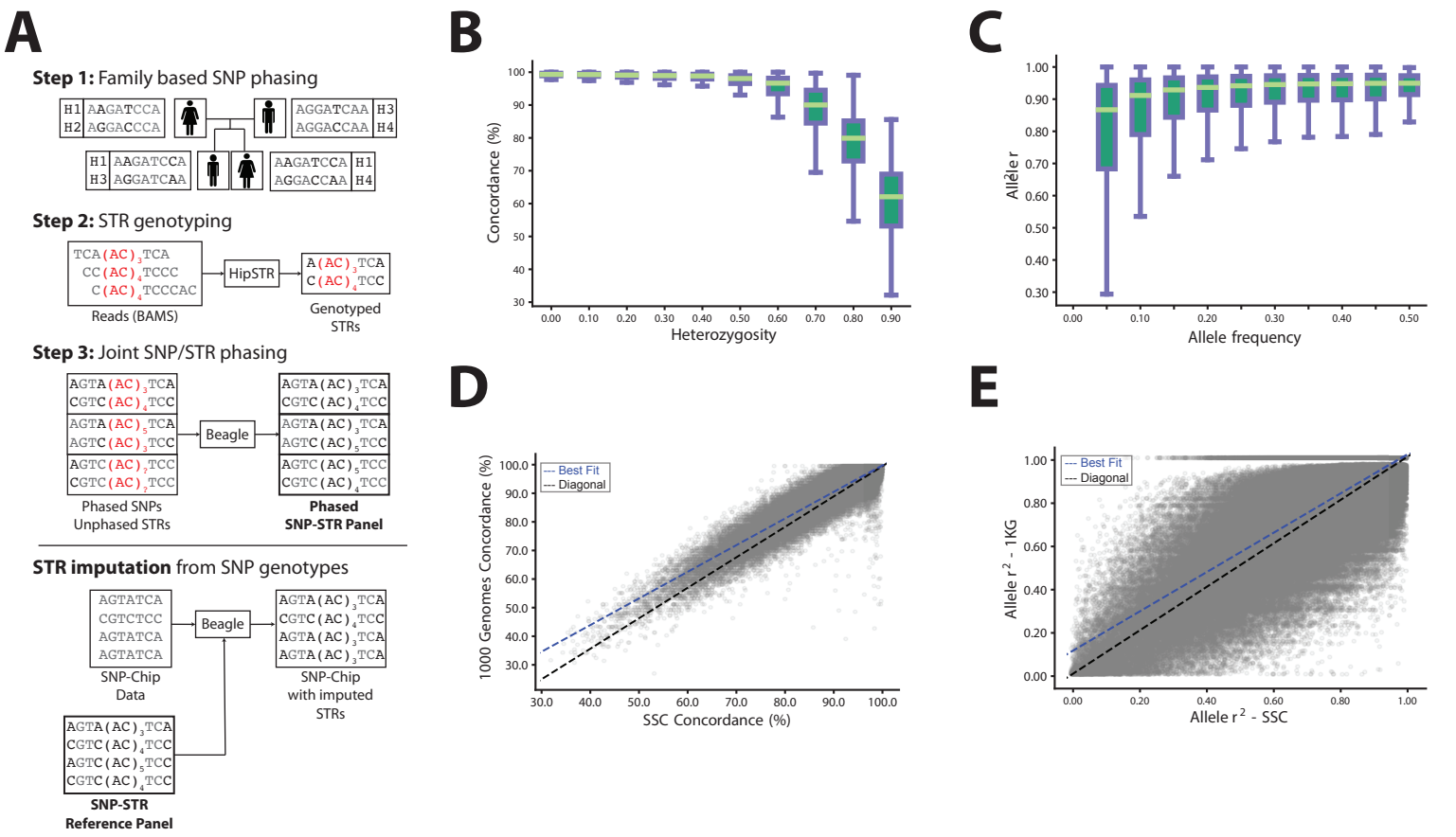
sequence data. *Genome Res.* **27,** 1895–1903 (2017).

50. Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V. & Bafna, V. Targeted Genotyping of Variable Number Tandem Repeats with adVNTR. (2017). doi:10.1101/221754
51. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
52. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10,** e1004234 (2014).
53. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
54. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27,** 718–719 (2011).
55. Karolchik, D. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32,** 493D–496 (2004).
56. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
57. Chakraborty, R., De Andrade, M., Daiger, S. P. & Budowle, B. Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. *Ann. Hum. Genet.* **56,** 45–57 (1992).
58. Fisher, S. A., Lewis, C. M. & Wise, L. H. Detecting population outliers and null alleles in linkage data: application to GAW12 asthma studies. *Genet. Epidemiol.* **21 Suppl 1,** S18–23 (2001).

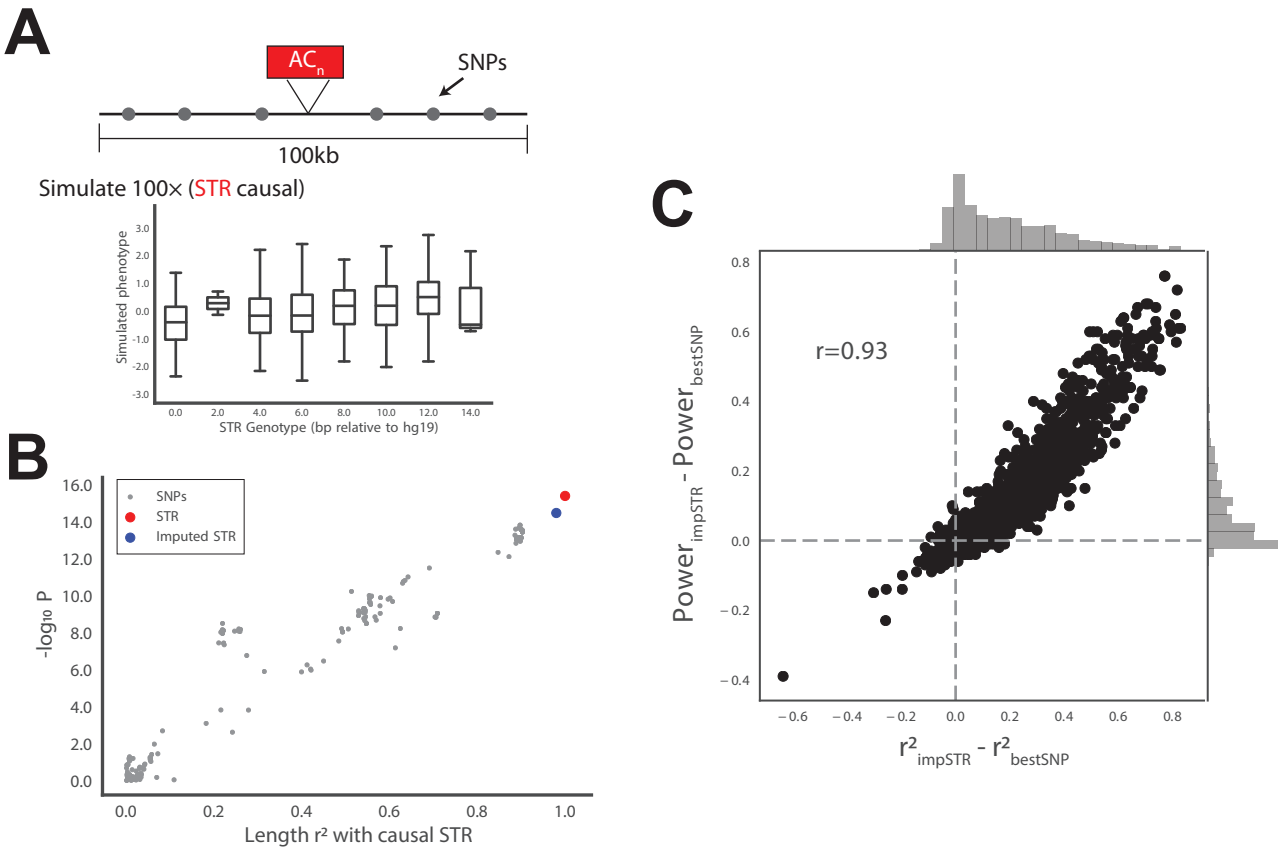
Saini, et al. Figure 1



Saini, et al. Figure 2



Saini, et al. Figure 3



Saini, et al. Figure 4