

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Methods for Exploring Ancestry-specific Genome Diversity and its Impact on Human Traits

Permalink

<https://escholarship.org/uc/item/149695cq>

ISBN

9798241632241

Author

Ziaei Jam, Helyaneh

Publication Date

2026-03-31

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Methods for Exploring Ancestry-specific Genome Diversity and its Impact on Human Traits

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

in

Computer Science

by

Helyaneh Ziaei Jam

Committee in charge:

Professor Melissa Gymrek, Chair

Professor Tiffany Amariuta-Bartell

Professor Vineet Bafna

Professor Taylor Berg-Kirkpatrick

2026

The dissertation of Helyaneh Ziaei Jam is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2026

Dedication

To my parents, who put their children before themselves.

To Mohammad, my bedrock.

Table of Contents

DISSERTATION APPROVAL PAGE	III
DEDICATION	IV
TABLE OF CONTENTS	V
LIST OF FIGURES	VIII
LIST OF TABLES	X
ACKNOWLEDGMENTS	XI
VITA	XIV
ABSTRACT OF THE DISSERTATION	XVI
INTRODUCTION	1
GENETIC VARIATION AND UNDERREPRESENTATION OF COMPLEX VARIANTS	1
AN OVERVIEW OF TANDEM REPEATS	2
ANCESTRAL BIAS IN GENOMIC RESEARCH	4
GENOME-WIDE ASSOCIATION STUDIES	5
POLYGENIC RISK SCORE	7
CONTRIBUTED RESEARCH	9
1 CHAPTER 1: A DEEP POPULATION REFERENCE PANEL OF TANDEM REPEAT VARIATION . 21	
1.1 INTRODUCTION	21
1.2 RESULTS	23
1.2.1 A genome-wide catalog of TR variation	23
1.2.2 Characterizing population-specific TR variation	27
1.2.3 Sequence determinants of TR heterozygosity	31
1.2.4 Detecting TRs associated with gene expression	33
1.2.5 Phased reference panel allows accurate imputation of TR variants	34
1.3 DISCUSSION	36
1.4 METHODS	39
1.4.1 Dataset description	39
1.4.2 TR genotyping with published tools	40
1.4.3 Filtering initial TR genotypes	42
1.4.4 Merging all populations	43
1.4.5 Ensemble genotyping	44
1.4.6 Experimental validation of TR genotypes	47
1.4.7 Published allele frequencies for forensics and disease-associated TRs	48
1.4.8 Mendelian inheritance analysis	49
1.4.9 Characterization of population-specific TR variation	49
1.4.10 Inspecting TRs not matching hg38	50
1.4.11 Detecting population-specific expansions	51
1.4.12 Visualizing repeat composition with TRviz	53
1.4.13 Finding sequence determinants of TR heterozygosity	53
1.4.14 eQTL analysis in the GEUVADIS cohort	54
1.4.15 Phasing and imputation	56

1.5	DATA AVAILABILITY	58
1.6	CODE AVAILABILITY.....	59
1.7	ACKNOWLEDGMENTS	59
1.8	APPENDIX	59
1.9	REFERENCES.....	86
2	CHAPTER 2: LONGTR: GENOME-WIDE PROFILING OF GENETIC VARIATION AT TANDEM REPEATS FROM LONG READS.....	95
2.1	INTRODUCTION	95
2.2	RESULTS AND DISCUSSION.....	95
2.3	CONCLUSIONS.....	100
2.4	METHODS	101
2.4.1	Overview of the LongTR method.....	101
2.4.2	Evaluating LongTR.....	105
2.4.3	Benchmarking against TRGT	106
2.4.4	Benchmarking against Straglr.....	106
2.4.5	Benchmarking against Tandem-genotypes.....	106
2.4.6	Benchmarking against adVNTR	106
2.4.7	Evaluation on ONT reads.....	107
2.4.8	Comparison to short read STR calls	107
2.5	DECLARATIONS.....	107
2.6	AVAILABILITY OF DATA AND MATERIALS:	108
2.7	ACKNOWLEDGEMENTS	109
2.8	APPENDIX	111
2.9	REFERENCES.....	122
3	CHAPTER 3: IMPROVING POLYGENIC RISK SCORE PREDICTION IN MULTI-ANCESTRY COHORTS.....	126
3.1	INTRODUCTION	127
3.2	RESULTS	129
3.2.1	Overview of PRS analysis.....	129
3.2.2	Overview of imPRS	130
3.2.3	Simulations.....	130
3.2.4	Real Data analysis.....	132
3.3	METHODS	133
3.3.1	Overview of imPRS	133
3.3.2	Overview of Existing PRS Methods	136
3.3.3	Simulation setting	137
3.3.4	Building PRS models.....	138
3.3.5	Real Data Analysis	138
3.4	DISCUSSION	139
3.5	CODE AVAILABILITY.....	141
3.6	AUTHOR CONTRIBUTIONS	141
3.7	ACKNOWLEDGEMENTS	141
3.8	APPENDIX	141
3.9	REFERENCES.....	153

List of Abbreviations

bp – base pairs (as a unit of length)
eQTL – expression quantitative trait locus
GTEx – Genotype-Tissue Expression Project
GWAS – genome-wide association study
Indel – insertion or deletion
kb – kilobases (thousands of bases, as a unit of length)
LD – linkage disequilibrium
mb – megabases (millions of bases, as a unit of length)
PC – (genetic) principal component
PRS – polygenic risk score
QC – quality control
SNP/SNV – single nucleotide polymorphism/variant
STR – short tandem repeat
AoU – All of Us cohort
VNTR – variable number tandem repeat

List of Figures

Figure 1.1: A deep catalog of TR variation across human populations	62
Figure 1.2: Per-locus comparison between WGS-based TR genotypes and capillary electrophoresis results	63
Figure 1.3: Comparison of genotypes called by capillary electrophoresis vs. EnsembleTR.....	64
Figure 1.4: Comparison of population-specific allele length distributions called by EnsembleTR vs. published frequencies at the CODIS forensic loci	65
Figure 1.5: Comparison of population-specific allele length distributions called by EnsembleTR vs. published frequencies at the disease-associated loci	66
Figure 1.6: Distribution of the fraction of non-reference alleles in individuals by population	67
Figure 1.7: Cumulative number of TR alleles with each additional sample.....	68
Figure 1.8: TR variants capture known population structure. a. Principal components analysis of all samples.....	69
Figure 1.9: Comparison of TR patterns in coding regions vs. genome-wide. a. Comparison of TR repeat unit length distributions.....	70
Figure 1.10: TRs exhibit a range of polymorphism levels across loci.....	71
Figure 1.11: Comparison of TR heterozygosity values across populations.....	72
Figure 1.12: Comparison of TR heterozygosity values across populations.....	73
Figure 1.13: Characterizing population-specific TR variation. a-b. Distribution of variant allele sizes.....	74
Figure 1.14: Example TRs for which the majority of observed alleles show dramatic deviations from the hg38 reference	75
Figure 1.15: Experimental validation of genotypes at the CAG repeat in an intron of CA10.....	76
Figure 1.16: Allele sequence visualization with TRviz	77
Figure 1.17: Sequence determinants of TR polymorphism. a. Heterozygosity is correlated with total TR length	78
Figure 1.18: Heterozygosity is correlated with total TR length.....	80
Figure 1.19: Significant HOMER results not marked as likely false positives	81
Figure 1.20: Integrated Gradients attribution scores for most confidently predicted true positive (correctly predicted polymorphic) and true negative (correctly predicted stable) STRs by the CNN model	82
Figure 1.21: TRs associated with gene expression in LCLs. a. Schematic overview of eTR detection.....	83
Figure 1.22: Imputation accuracy decreases with the number of common alleles	84
Figure 1.23: Phasing and imputation at TRs. a. Imputation accuracy decreases with heterozygosity	85

Figure 2.1: Base pair error distribution for PacBio HiFi and Illumina reads. Bars show the.....	111
Figure 2.2: LongTR workflow	112
Figure 2.3: Benchmarking LongTR against TRGT	113
Figure 2.4: Mendelian Inheritance for an Ashkenazi Trio at homopolymer TRs as a function of LongTR quality score.	114
Figure 2.5: IGV screenshots of examples where LongTR matches the assembly but TRGT does not.	115
Figure 2.6: IGV screenshots comparing HG002 assembly alleles with PacBio HiFi reads aligned to GRCh38 at loci where TRGT and LongTR agree but do not match the assembly.....	116
Figure 2.7: Waterfall plots for reads from samples with an HTT expansion genotyped by LongTR.....	117
Figure 2.8: ONT vs PacBio.....	118
Figure 2.9: Comparison between STR calls by HipSTR using Illumina short reads vs. LongTR on PacBio HiFi reads from sample HG002	119
Figure 3.1: Overview of PRS construction	142
Figure 3.2: Benchmarking PRS prediction accuracy across ancestral architectures.	143
Figure 3.3: Benchmarking PRS prediction accuracy across 6 traits in UKB samples	144
Figure 3.4: Benchmarking PRS prediction accuracy across ancestral architectures.	145
Figure 3.5: Benchmarking PRS prediction accuracy in EAS, AFR, and EUR populations	146
Figure 3.6: GWAS for platelet count.	147
Figure 3.7: GWAS for triglycerides.....	148
Figure 3.8: GWAS for red blood cell count.	149
Figure 3.9: GWAS for HDL cholesterol.	150
Figure 3.10: GWAS for white blood cell count.	151
Figure 3.11: GWAS for glycated haemoglobin.....	152

List of Tables

Table 2-1: Inferred error model parameters at homopolymer TRs in PacBio HiFi reads from HG002.....	120
Table 2-2: Benchmarking LongTR against TRGT in patients with pathogenic expansions.....	121

Acknowledgments

I extend my deepest gratitude to my thesis advisor, Professor Melissa Gymrek, for her exceptional mentorship and support throughout the course of my PhD. Her scientific rigor, generosity, authenticity, and contagious curiosity have shaped my development not only as a researcher, but as a human being. Throughout my PhD, she gave me enough freedom to explore different topics, and she spent time convincing me of the necessity of every analysis we did, rather than doing it by force and the power that comes with her position. I enjoyed every conversation we had in our weekly meetings and other occasions, where I could witness her attention to details, her modesty and positivity. It has been the greatest privilege to learn from her example—both in science and in life. I am also sincerely grateful to my thesis committee members for their thoughtful feedback and support over the years. Each of them has provided invaluable guidance at key moments in my journey, and I am deeply thankful for their time, insight, and encouragement.

Throughout my PhD, I have been very fortunate to be surrounded by an extraordinary community of family, friends, collaborators, and lab mates who have made these years not only intellectually fulfilling, but also full of joy and unforgettable experiences. I'd like to thank Nima Mousavi, whose introduction led me to the Gymrek Lab and whose final work laid the foundation for my first major project. I am also incredibly grateful to Yang Li, Nichole Ma, and Ross DeVito for their tremendous efforts on the EnsembleTR project—this work would not have come together without their dedication and collaboration. I would also like to thank Sara Javadzadeh and Jonghun Park from the Bafna Lab for their valuable contributions to the EnsembleTR and LongTR projects. I am grateful to the members of H3Africa cohort for their invaluable input for the EnsembleTR project and the amazing opportunities and experiences that followed.

I'd also like to thank Tara Mirmira and Jonathan Margoliash —thank you for answering a million questions with patience and generosity. Your input accelerated my final project in countless ways. I'm so lucky to have shared a lab with so many amazing people. I also want to acknowledge the fellow members of my graduate cohort, who made this journey all the more rewarding, and the alumni who offered practical advice and modeled how to move through this path with resilience and grace.

I'd also like to thank the many people I've had the chance to collaborate with, who further helped shape my research, expand my knowledge, and fuel my curiosity. I'm especially grateful to Dr. Justin Zook from the Genome in a Bottle consortium—I had the honor of working with him on two projects and really appreciated his commitment and vision. I also want to thank Professor Mark Chaisson; I've truly enjoyed our collaboration and our conversations at different conferences. I'm grateful as well to Professor Abraham Palmer and his PhD student Denghui Chen for giving me the opportunity to go beyond the human genome through our work on rat repeats. Finally, I'd like to thank Milad Mortazavi and Professor Jonathan Sebat for their valuable feedback, which helped us make LongTR a more useful tool.

I'm deeply grateful to everyone who made the last five years an incredible and joyful chapter of my life, to name a few, I'm grateful to Mohammad, Hasti, Ashkan, Setareh, Hosein B, Hosein Y, Nazli, Farshid, Ali KB, Yasi, Samira, Ali RJ, Sahand, Rojin, Ronak, Misagh, Nasim, Milad, Mehrnaz, Raman, Nicole, Vahid, Rosa, Aria, Jenni, Sina M, Shima, Shayan, Kiana, Alma, Pegah, Fathol, Siavash, Soheil, Mahsa, Amirhosein, Shayesteh, Peyman, Keyvan, Sogol, Shahab and many others who I've been fortunate enough to call friends and family.

Chapter 1, in full, is a reprint of the material “A deep population reference panel of tandem repeat variation” by Helyaneh Ziaei Jam, Yang Li, Ross DeVito, Nima Mousavi, Nichole Ma, Ibra

Lujumba, Yagoub Adam, Mikhail Maksimov, Bonnie Huang, Egor Dolzhenko, Yunjiang Qiu, Fredrick Elishama Kakembo, Habi Joseph, Blessing Onyido, Jumoke Adeyemi, Mehrdad Bakhtiari, Jonghun Park, Sara Javadzadeh, Daudi Jjingo, Ezekiel Adebisi, Vineet Bafna & Melissa Gymrek as it appears in Nature Communications 2023. The dissertation author is the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material in “LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads” by Helyaneh Ziaei Jam, Justin M. Zook, Sara Javadzadeh, Jonghun Park, Aarushi Sehgal and Melissa Gymrek as it appeared in Genome Biology 2024. The dissertation author is the primary investigator and author of this paper.

Chapter 3, is unpublished material by Helyaneh Ziaei Jam, Tara Mirmira, and Melissa Gymrek. The dissertation author was the primary investigator and author of this paper.

Vita

2015-2020	Bachelor of Science in Computer Engineering, Sharif University of Technology
2021-2025	Research Assistant, University of California San Diego
2025	Master of Science in Computer Science, University of California San Diego
2026	Doctor of Philosophy in Computer Science, University of California San Diego

PUBLICATIONS

1. H. Ziaei Jam, Y. Li, R. DeVito, N. Mousavi, N. Ma, I. Lujumba, Y. Adam, M. Maksimov, B. Huang, E. Dolzhenko, Y. Qiu, F. E. Kakembo, H. Joseph, B. Onyido, J. Adeyemi, M. Bakhtiari, J. Park, S. Javadzadeh, D. Jjingo, E. Adebisi, V. Bafna, M. Gymrek, A deep population reference panel of tandem repeat variation. *Nat. Commun.* **14**, 6711 (2023).
2. H. Ziaei Jam, J. M. Zook, S. Javadzadeh, J. Park, A. Sehgal, M. Gymrek, LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads. *Genome Biol.* **25**, 176 (2024).
3. A. C. English, E. Dolzhenko, H. Ziaei Jam, S. K. McKenzie, N. D. Olson, W. De Coster, J. Park, B. Gu, J. Wagner, M. A. Eberle, M. Gymrek, M. J. P. Chaisson, J. M. Zook, F. J. Sedlazeck, Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nat. Biotechnol.*, doi: 10.1038/s41587-024-02225-z (2024).
4. O. S. Lundström, M. Adriaan Verbiest, F. Xia, H. Z. Jam, I. Zlobec, M. Anisimova, M. Gymrek, WebSTR: A Population-wide Database of Short Tandem Repeat Variation in Humans. *J. Mol. Biol.* **435**, 168260 (2023).
5. A. Sehgal, H. Ziaei Jam, A. Shen, M. Gymrek, Genome-wide detection of somatic mosaicism at short tandem repeats. *Bioinformatics* **40** (2024).
6. M. Mortazavi, J. Guevara, J. Diaz, S. Tran, H. Ziaei Jam, S. Batalov, M. Bainbridge, A. D. Besterman, M. Gymrek, A. A. Palmer, J. Sebat, Long read genome sequencing elucidates diverse functional consequences of structural and repeat variation in autism, *medRxiv* (2025). <https://doi.org/10.1101/2025.07.20.25331880>.

7. B. Weisburd, E. Dolzhenko, M. F. Bennett, M. C. Danzi, A. English, L. Hiatt, H. Tanudisastro, N. E. Kurtas, H. Z. Jam, H. Brand, F. J. Sedlazeck, M. Gymrek, H. Dashnow, M. A. Eberle, H. L. Rehm, Defining a tandem repeat catalog and variation clusters for genome-wide analyses and population databases, *bioRxiv* (2024).
<https://doi.org/10.1101/2024.10.04.615514>.
8. I. Lujumba, Y. Adam, H. Ziaei Jam, I. Isewon, N. Monnakgotla, Y. Li, B. Onyido, K. Fredrick, F. Adegoke, J. Emmanuel, J. Adeyemi, O. Ibitoye, S. Owusu-Ansah, M. B. Akanle, H. Joseph, M. Nsubuga, R. Galiwango, M. Okitwi, N. Magdalene, O. Walter, Z. Mngadi, M. Adebiyi, J. Oyelade, M. Nel, D. Jjingo, M. Gymrek, E. Adebiyi, A practical guide to identifying associations between tandem repeats and complex human traits using consensus genotypes from multiple tools. *Nat. Protoc.*, doi: 10.1038/s41596-025-01231-y (2025).
9. B. Huang, A. Durvasula, N. Mousavi, H. Ziaei-Jam, M. Maksimov, K. E. Lohmueller, M. Gymrek, Genome-wide selection inference at short tandem repeats. *PLoS Genet.* **21**, e1011959 (2025).

ABSTRACT OF THE DISSERTATION

Methods for Exploring Ancestry-specific Genome Diversity and its Impact on Human Traits

by

Helyaneh Ziaei Jam

Doctor of Philosophy in Computer Science

University of California San Diego, 2026

Professor Melissa Gymrek, Chair

Most genetic studies have historically centered on single-nucleotide polymorphisms (SNPs) in individuals of European ancestry, overlooking both complex genomic variants and global diversity. This thesis focuses on improving our understanding and analysis of tandem repeats (TRs)—a major, yet underexplored, source of genomic variation—through a multi-

ancestry lens. First, we developed EnsembleTR, a genotyping framework that harmonizes TR calls across methods, and used it to construct a reference panel of 1.7 million TRs from 3,550 individuals in the 1000 Genomes Project and H3Africa cohorts. This panel enabled various analyses, such as identifying ancestry-specific repeat expansions, expression quantitative trait loci (eQTL) mapping, and investigating sequence determinants of TR stability. To further improve our understanding of TRs, we built LongTR, a new tool for accurate TR quantification from both PacBio and Oxford Nanopore data, demonstrating superior accuracy and scalability over existing tools. Finally, we investigated two novel approaches for constructing multi-ancestry polygenic risk scores (PRS). First, we developed imPRS, that jointly models posterior effect sizes across populations and introduces a correlation parameter to enforce similar effect sizes for the shared variants among population. Second, we explored building the PRS model without reliance on ancestry labels in genome-wide association studies (GWAS), linkage disequilibrium (LD) estimation, and PRS calculation, by pooling all samples from every ancestry into one single cohort. Together, these contributions advance the inclusion of structurally complex variants and diverse populations in genetic analyses, with implications for improved variant discovery and equitable risk prediction.

Introduction

Genetic variation and underrepresentation of complex variants

Genetic variation underlies the diversity of human traits and influences susceptibility to a wide range of diseases. With the advent of whole genome sequencing (WGS), researchers can now systematically identify and study these variations across the entire genome, enabling a deeper understanding of both rare and common forms of genetic diversity. The functional consequences of genetic variants depend heavily on their genomic context.¹ Variants in protein-coding genes can lead to loss of function, gain of function, or altered protein products. Variants located in non-coding regions—such as promoters, enhancers, and other regulatory elements—can modulate gene expression by influencing transcription factor binding or chromatin state.²

Genetic variation occurs at multiple scales and takes on many forms, each with the potential to affect biological function in distinct ways. Single nucleotide polymorphisms (SNPs), which involve a change at a single base pair, are the most common form of genetic variation. Though many individual SNPs have modest effects, they collectively contribute to variation in traits such as disease risk, physical characteristics, and drug response³ (reviewed in⁴). Short insertions and deletions (indels), involving the gain or loss of a few nucleotides, can have more substantial consequences, especially when they disrupt protein-coding regions or alter regulatory sequences. Beyond these small-scale variants, the genome also harbors larger and more complex forms of variation—including structural variants such as deletions, duplications, inversions, and translocations, often span hundreds to thousands of base pairs.

While complex variants are less frequent than SNPs, their functional effects can be more dramatic. Many of these more complex variants have historically been difficult to detect and interpret due to sequencing and computational limitations, but recent improvements have begun to reveal their contribution to human phenotypes. Emerging evidence suggests that complex variants may account for portions of the missing heritability not explained by SNPs or simple indels alone, highlighting their potential importance in shaping a wide range of traits and diseases (reviewed in⁵).

An Overview of Tandem Repeats

Tandem repeats (TRs) are another class of more complex genetic variants, characterized by the presence of DNA sequences repeated adjacently in the genome. These repetitive elements are broadly categorized into short tandem repeats (STRs), which have repeat units ranging from 1 to 6 base pairs, and variable number tandem repeats (VNTRs), which typically consist of longer repeat units⁶. Together, TRs constitute approximately 3% of human genome and represent one of its most mutable components, with mutation rate estimated around 5.5×10^{-6} (compared to $\sim 1.2 \times 10^{-8}$ for SNPs) due to high rates of replication slippage and other mutational processes^{7,8}.

TRs are increasingly being recognized for their contribution to both rare and common traits. Expansions of specific STRs are known to cause dozens of Mendelian disorders, including Huntington's disease⁹, Fragile X syndrome¹⁰, and other (reviewed in¹¹). Beyond rare diseases, recent work has also implicated TR variation in complex traits such as height¹², blood cell phenotypes¹³, and neuropsychiatric conditions⁷. Additionally, TRs can influence gene expression¹⁴ and alternative splicing^{15,16}, often through changes in local chromatin accessibility, transcription factor binding, or RNA processing¹⁷⁻¹⁹.

Despite their prevalence and functional potential, TRs have historically been underrepresented in large-scale genomic studies. This is due, in part, to their complexity, which makes them difficult to genotype using standard variant calling pipelines optimized for SNPs and short indels. Since their emergence, short-read sequencing technologies have provided broad and affordable coverage of the human genome and serve as the backbone of most large-scale population sequencing efforts. Over the past decade, multiple tools have been developed to enable characterization of tandem repeats from short-read sequencing data^{20–26}. However, accurately genotyping TRs from short reads is challenging due to their repetitive nature, sequencing errors within repeats, and alignment ambiguities, especially for longer or imperfect repeat regions (reviewed in²⁷). These limitations have constrained the scope and resolution of TR analyses in many cohorts. Long-read sequencing technologies provide a promising path forward. By generating reads that span entire repeat regions, they enable more direct and accurate characterization of tandem repeat length and structure. Several tools have been developed to leverage long-read data for tandem repeat genotyping and have demonstrated the ability to resolve longer and more complex repeats than is possible with short-read sequencing^{28–32}. However, long-read technologies still lag behind short-read sequencing in throughput, accuracy, and cost-efficiency, and large cohorts with long-read data are not yet widely available (reviewed in²⁷).

As tools and technologies for TR genotyping continue to evolve, integrating TR variation into broader studies of human genetics will be crucial. Doing so has the potential to uncover novel biological insights and to help explain portions of heritability that remain unresolved by more tractable forms of variation.

Ancestral Bias in Genomic Research

Over the past decade, most of the large-scale sequencing and genotyping efforts have disproportionately focused on individuals of European ancestry^{33–35}. As a result, much of our current understanding of human genetic variation—and its relationship to traits and disease—has been shaped by a relatively narrow slice of global diversity. This imbalance has important scientific and ethical consequences. From a scientific perspective, limiting studies to a single ancestry group reduces our ability to uncover the full spectrum of human genetic variation. Many genetic variants that are common in other populations may be rare or absent in Europeans, meaning key contributors to phenotypic variation and disease risk remain undiscovered or poorly understood. This becomes a significant barrier for use of genetic findings in clinical settings³⁶. Furthermore, while emerging evidence suggests that for most traits the overall genetic architecture is broadly similar across populations, important exceptions exist. In some cases, key variants or regulatory mechanisms appear to differ, potentially due to population-specific selection, demographic history, or environmental exposures^{37,38}. Accounting for this diversity is essential—not only for improving the accuracy and equity of genetic research, but also for uncovering novel biology that may otherwise remain hidden.

In recent years, several initiatives have sought to address this gap by building more diverse genomic resources. The 1000 Genomes Project (1KG), for example, included individuals from multiple global populations and provided a foundational reference panel for human genetic variation³⁹. While early phases of the project used low-coverage sequencing, more recent phases have produced deep WGS data, offering improved resolution across variant types. Another major effort is the All of Us (AoU) Research Program, which aims to recruit over one million participants with a strong emphasis on diversity, including racial and ethnic minorities historically

underrepresented in biomedical research⁴⁰. These efforts represent important steps toward building a more equitable and comprehensive picture of human genomics.

Complementing these large short-read datasets, long-read sequencing technologies are now being applied to individuals from diverse backgrounds. Although these long-read datasets are still smaller in size, they provide a crucial layer of information, enabling the characterization of complex and previously intractable variants across different populations (reviewed in⁴¹).

Together, these resources are expanding our ability to study population-specific patterns of variation, such as differences in allele frequencies, mutation rates, and linkage disequilibrium structures. They also enable more robust inferences about shared and unique features of human biology.

Genome-wide association studies

Genome-wide association studies (GWAS) have become a foundational tool in human genetics for identifying genetic variants associated with complex traits and diseases. In a typical GWAS, researchers test millions of genetic variants across the genome—most commonly single nucleotide polymorphisms (SNPs)—for statistical association with a phenotype of interest in a cohort of individuals. This is generally done using linear or logistic regression models depending on whether the trait is quantitative or binary.⁴²

To ensure that observed associations reflect true genetic effects rather than artifacts of population structure or other confounders, GWAS models include various covariates. Among the most common are principal components (PCs), which are derived from genetic data using principal component analysis. PCs capture broad patterns of population structure and ancestry, helping to

adjust for differences among individuals due to their genetic background rather than the phenotype.⁴³ Additional covariates often include age, sex, and technical variables such as batch or sequencing platform.

Despite these precautions, interpreting GWAS results is not straightforward. One of the primary challenges is the presence of linkage disequilibrium (LD)—a phenomenon where nearby genetic variants tend to be inherited together⁴⁴. As a result, a significant GWAS signal at a given locus often implicates multiple correlated variants, making it difficult to pinpoint the true causal variant⁴⁵. Further, LD patterns differ across populations due to their unique demographic histories, which adds another layer of complexity.

An ongoing debate in the GWAS field concerns how best to handle genetic diversity in the analysis. Current common approach is to stratify analyses by ancestry groups, performing separate GWAS within each population. This can reduce false positives from population structure but limits the sample size and may miss shared genetic effects. Alternatively, meta-analysis combines results from population-stratified GWAS while accounting for differences in allele frequency and effect sizes^{46,47}. A third option is to pool all individuals together into a single GWAS, using covariates like PCs to control for structure⁴⁸.

However, these approaches are not without limitations. For example, the strategy of assigning individuals to discrete populations is often biologically simplistic. Human genetic variation exists along a continuum, and many individuals—particularly those from admixed populations like African Americans or Latino individuals—do not fit neatly into one ancestry group⁴⁹. For these individuals, it is not always clear whether they should be analyzed within a specific population group, in pooled GWAS, or excluded altogether. This raises concerns about equity and

generalizability in genetic research, as admixed individuals are frequently underrepresented or analytically sidelined.

Another key consideration is statistical power: larger sample sizes increase the ability to detect small-effect variants, motivating pooled analyses. Pooling samples together might have the potential to improve prediction for populations with limited sample size, by borrowing information from well represented populations. Yet this should be weighed against contamination of ancestry-specific signals, specific to the smaller cohort ancestries, dominated by the other cohorts.

Polygenic risk score

Many human traits and common diseases are polygenic—influenced by the cumulative effect of many genetic variants, each contributing a small amount to overall risk or variation. Traits like height, body mass index (BMI), and educational attainment, as well as diseases such as type 2 diabetes, coronary artery disease, and schizophrenia, all exhibit this highly polygenic architecture^{50–55}.

Polygenic risk scores (PRS) aim to estimate an individual's genetic liability to such traits by aggregating the effects of many loci across the genome. Typically, PRS construction begins with summary statistics obtained from GWAS. To build a predictive model, various methods are used to select and weight SNPs—ranging from simple p-value thresholding⁵⁶ to more complex approaches like LASSO-based regression⁵⁷ and Bayesian shrinkage models⁵⁸ that account for LD and effect size distributions. These models attempt to balance including true signals while minimizing noise and overfitting.

While early PRS methods were designed for single populations, recent developments have focused on multi-ancestry PRS^{59,60}. These approaches aim to improve accuracy and generalizability by sharing information across populations, often through statistical structures that model correlations in effect sizes^{61,62}. Importantly, because GWAS sample sizes for non-European populations are typically much smaller, building ancestry-specific PRS in these groups has limited power. Multi-ancestry models, by leveraging information from larger studies, can partially address this limitation and improve performance in underrepresented populations. While multi-ancestry PRS models are generally more robust than single-ancestry ones, their performance still lags significantly in non-European individuals.

This performance gap is partly due to the differences in LD patterns and allele frequencies across populations, and the way it influences how effectively SNPs tag causal variants. Although specialized fine-mapping tools exist to identify likely causal variants within associated regions, using approaches such as sum of single effects⁶³, Bayesian inference⁴⁵, and functional annotation integration⁶⁴, PRS generally do not aim to pinpoint causality. As a result, if a causal variant is missing, poorly tagged, or represents a complex variant not captured by SNP-based GWAS in a target population, the predictive accuracy and transferability of the PRS decline^{36,65}.

As mentioned in previous section, this challenge is further compounded by population structure. Individuals from admixed or intermediate backgrounds do not fit neatly into discrete categories. Because multi-ancestry PRS models typically assume well-defined population labels, this makes application to such individuals more difficult^{60,62}.

Contributed research

Chapters 1 and 2 present full reprints of peer-reviewed papers that form the core of my doctoral research, in order of publication. Chapter 1 details my work leading the construction of a reference panel of 1.7 million high-quality tandem repeat genotypes across 3,550 samples from diverse populations, which we further used to study TRs in various contexts, including their effects on gene expression. Chapter 2 describes the development of a tool for TR genotyping from long-read sequencing data, including benchmarking against both short- and long-read methods and discussion of implications for the future of TR analysis. Chapter 3 describes two approaches to build PRS in diverse ancestries.

1. M. Claussnitzer, J. H. Cho, R. Collins, N. J. Cox, E. T. Dermitzakis, M. E. Hurles, S. Kathiresan, E. E. Kenny, C. M. Lindgren, D. G. MacArthur, K. N. North, S. E. Plon, H. L. Rehm, N. Risch, C. N. Rotimi, J. Shendure, N. Soranzo, M. I. McCarthy, A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
2. M. J. Li, B. Yan, P. C. Sham, J. Wang, Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. *Brief. Bioinform.* **16**, 393–412 (2015).
3. A. Muhammad, I. T. Aka, K. A. Birdwell, A. S. Gordon, D. M. Roden, W.-Q. Wei, J. D. Mosley, S. L. Van Driest, Genome-wide approach to measure variant-based heritability of drug outcome phenotypes. *Clin. Pharmacol. Ther.* **110**, 714–722 (2021).
4. P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, J. Yang, 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
5. R. L. Collins, M. E. Talkowski, Diversity and consequences of structural variation in the human genome. *Nat. Rev. Genet.* **26**, 443–462 (2025).
6. H. Ziaei Jam, Y. Li, R. DeVito, N. Mousavi, N. Ma, I. Lujumba, Y. Adam, M. Maksimov, B. Huang, E. Dolzhenko, Y. Qiu, F. E. Kakembo, H. Joseph, B. Onyido, J. Adeyemi, M. Bakhtiari, J. Park, S. Javadzadeh, D. Jjingo, E. Adebiyi, V. Bafna, M. Gymrek, A deep population reference panel of tandem repeat variation. *Nat. Commun.* **14**, 6711 (2023).

7. I. Mitra, B. Huang, N. Mousavi, N. Ma, M. Lamkin, R. Yanicky, S. Shleizer-Burko, K. E. Lohmueller, M. Gymrek, Patterns of de novo tandem repeat mutations and their role in autism. *Nature* **589**, 246–250 (2021).
8. D. Porubsky, H. Dashnow, T. A. Sasani, G. A. Logsdon, P. Hallast, M. D. Noyes, Z. N. Kronenberg, T. Mokveld, N. Koundinya, C. Nolan, C. J. Steely, A. Guaracino, E. Dolzhenko, W. T. Harvey, W. J. Rowell, K. Grigorev, T. J. Nicholas, M. E. Goldberg, K. K. Oshima, J. Lin, P. Ebert, W. S. Watkins, T. Y. Leung, V. C. T. Hanlon, S. McGee, B. S. Pedersen, H. C. Happ, H. Jeong, K. M. Munson, K. Hoekzema, D. D. Chan, Y. Wang, J. Knuth, G. H. Garcia, C. Fanslow, C. Lambert, C. Lee, J. D. Smith, S. Levy, C. E. Mason, E. Garrison, P. M. Lansdorp, D. W. Neklason, L. B. Jorde, A. R. Quinlan, M. A. Eberle, E. E. Eichler, Human de novo mutation rates from a four-generation pedigree reference. *Nature* **643**, 427–436 (2025).
9. T. Pringsheim, K. Wiltshire, L. Day, J. Dykeman, T. Steeves, N. Jette, The incidence and prevalence of Huntington’s disease: a systematic review and meta-analysis. *Mov. Disord.* **27**, 1083–1091 (2012).
10. J. Hunter, O. Rivero-Arias, A. Angelov, E. Kim, I. Fotheringham, J. Leal, Epidemiology of fragile X syndrome: a systematic review and meta-analysis. *Am. J. Med. Genet. A* **164A**, 1648–1658 (2014).
11. S. M. Mirkin, Expandable DNA repeats and human disease. *Nature* **447**, 932–940 (2007).
12. R. E. Mukamel, R. E. Handsaker, M. A. Sherman, A. R. Barton, Y. Zheng, S. A. McCarroll, P.-R. Loh, Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science* **373**, 1499–1505 (2021).
13. J. Margoliash, S. Fuchs, Y. Li, A. Massarat, A. Goren, M. Gymrek, Polymorphic short tandem repeats make widespread contributions to blood and serum traits. <https://doi.org/10.1101/2022.08.01.502370>.
14. S. F. Fotsing, J. Margoliash, C. Wang, S. Saini, R. Yanicky, S. Shleizer-Burko, A. Goren, M. Gymrek, The impact of short tandem repeat variation on gene expression. *Nat. Genet.* **51**, 1652–1659 (2019).
15. Y. Cui, F. J. Arnold, J. S. Li, J. Wu, D. Wang, J. Philippe, M. R. Colwin, S. Michels, C. Chen, T. Sallam, L. M. Thompson, A. R. La Spada, W. Li, Multi-omic quantitative trait loci link tandem repeat size variation to gene regulation in human brain. *Nat. Genet.* **57**, 369–378 (2025).
16. K. Hamanaka, D. Yamauchi, E. Koshimizu, K. Watase, K. Mogushi, K. Ishikawa, H. Mizusawa, N. Tsuchida, Y. Uchiyama, A. Fujita, K. Misawa, T. Mizuguchi, S. Miyatake, N. Matsumoto, Genome-wide identification of tandem repeats associated with splicing variation across 49 tissues in humans. *Genome Res.* **33**, 435–447 (2023).
17. X. Zhang, L. Zhang, E. Wang, S. Benton, E. Modolo, M. Maksimov, S. Shleizer-Burko, Q. Gong, C. Wang, M. Lamkin, E. Mendenhall, M. Gymrek, A. Goren, Systematic evaluation

- of the impact of promoter proximal short tandem repeats on expression, *bioRxiv*org (2025). <https://doi.org/10.1101/2025.09.14.676153>.
18. I. Sela, D. B. Lukatsky, DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. *Biophys. J.* **101**, 160–166 (2011).
 19. P. Garg, A. Martin-Trujillo, O. L. Rodriguez, S. J. Gies, E. Hadelia, B. Jadhav, M. Jain, B. Paten, A. J. Sharp, Pervasive cis effects of variation in copy number of large tandem repeats on local DNA methylation and gene expression. *Am. J. Hum. Genet.* **108**, 809–824 (2021).
 20. N. Mousavi, S. Shleizer-Burko, R. Yanicky, M. Gymrek, Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 (2019).
 21. T. Willems, D. Zielinski, J. Yuan, A. Gordon, M. Gymrek, Y. Erlich, Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods*, doi: 10.1038/nmeth.4267 (2017).
 22. M. Gymrek, D. Golan, S. Rosset, Y. Erlich, lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.* **22**, 1154–1162 (2012).
 23. G. Highnam, C. Franck, A. Martin, C. Stephens, A. Puthige, D. Mittelman, Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res.* **41**, e32 (2013).
 24. E. Dolzhenko, J. J. F. A. van Vugt, R. J. Shaw, M. A. Bekritsky, M. van Blitterswijk, G. Narzisi, S. S. Ajay, V. Rajan, B. R. Lajoie, N. H. Johnson, Z. Kingsbury, S. J. Humphray, R. D. Schellevis, W. J. Brands, M. Baker, R. Rademakers, M. Kooyman, G. H. P. Tazelaar, M. A. van Es, R. McLaughlin, W. Sproviero, A. Shatunov, A. Jones, A. Al Khleifat, A. Pittman, S. Morgan, O. Hardiman, A. Al-Chalabi, C. Shaw, B. Smith, E. J. Neo, K. Morrison, P. J. Shaw, C. Reeves, L. Winterkorn, N. S. Wexler, US–Venezuela Collaborative Research Group, D. E. Housman, C. W. Ng, A. L. Li, R. J. Taft, L. H. van den Berg, D. R. Bentley, J. H. Veldink, M. A. Eberle, Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).
 25. H. Dashnow, B. S. Pedersen, L. Hiatt, J. Brown, S. J. Beecroft, G. Ravenscroft, A. J. LaCroix, P. Lamont, R. H. Roxburgh, M. J. Rodrigues, M. Davis, H. C. Mefford, N. G. Laing, A. R. Quinlan, STRling: a k-mer counting approach that detects short tandem repeat expansions at known and novel loci. *Genome Biol.* **23**, 257 (2022).
 26. E. Dolzhenko, M. F. Bennett, P. A. Richmond, B. Trost, S. Chen, J. J. F. A. van Vugt, C. Nguyen, G. Narzisi, V. G. Gainullin, A. M. Gross, B. R. Lajoie, R. J. Taft, W. W. Wasserman, S. W. Scherer, J. H. Veldink, D. R. Bentley, R. K. C. Yuen, M. Bahlo, M. A. Eberle, ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol.* **21**, 102 (2020).
 27. H. A. Tanudisastro, I. W. Deveson, H. Dashnow, D. G. MacArthur, Sequencing and characterizing short tandem repeats in the human genome. *Nat. Rev. Genet.* **25**, 460–475 (2024).

28. H. Ziaei Jam, J. M. Zook, S. Javadzadeh, J. Park, A. Sehgal, M. Gymrek, LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads. *Genome Biol.* **25**, 176 (2024).
29. E. Dolzhenko, A. English, H. Dashnow, G. De Sena Brandine, T. Mokveld, W. J. Rowell, C. Karniski, Z. Kronenberg, M. C. Danzi, W. A. Cheung, C. Bi, E. Farrow, A. Wenger, K. P. Chua, V. Martínez-Cerdeño, T. D. Bartley, P. Jin, D. L. Nelson, S. Zuchner, T. Pastinen, A. R. Quinlan, F. J. Sedlazeck, M. A. Eberle, Characterization and visualization of tandem repeats at genome scale. *Nat. Biotechnol.*, doi: 10.1038/s41587-023-02057-3 (2024).
30. S. Mitsuhashi, M. C. Frith, T. Mizuguchi, S. Miyatake, T. Toyota, H. Adachi, Y. Oma, Y. Kino, H. Mitsuhashi, N. Matsumoto, Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol.* **20**, 58 (2019).
31. C. Readman, R.-B. Indhu-Shree, F. Jan M, B. Inanc, Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol.* **22**, 224 (2021).
32. D. Bolognini, A. Magi, V. Benes, J. O. Korbel, T. Rausch, TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data. *Gigascience* **9** (2020).
33. A. C. Need, D. B. Goldstein, Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* **25**, 489–494 (2009).
34. A. B. Popejoy, S. M. Fullerton, Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
35. N. A. Rosenberg, L. Huang, E. M. Jewett, Z. A. Szpiech, I. Jankovic, M. Boehnke, Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356–366 (2010).
36. A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, M. J. Daly, Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
37. T. Lappalainen, Y. I. Li, S. Ramachandran, A. Gusev, Genetic and molecular architecture of complex traits. *Cell* **187**, 1059–1075 (2024).
38. A. R. Bentley, S. L. Callier, C. N. Rotimi, Evaluating the promise of inclusion of African ancestry populations in genomics. *NPJ Genom. Med.* **5**, 5 (2020).
39. 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
40. *Precision Medicine Initiative (PMI) Working Group. The Precision Medicine Initiative Cohort Program.*

41. G. A. Logsdon, M. R. Vollger, E. E. Eichler, Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
42. J. Hardy, A. Singleton, Genomewide association studies and human disease. *N. Engl. J. Med.* **360**, 1759–1768 (2009).
43. A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
44. J. K. Pritchard, M. Przeworski, Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
45. F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, E. Eskin, Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
46. C. J. Willer, Y. Li, G. R. Abecasis, METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
47. B. M. L. Baselmans, R. Jansen, H. F. Ip, J. van Dongen, A. Abdellaoui, M. P. van de Weijer, Y. Bao, M. Smart, M. Kumari, G. Willemsen, J.-J. Hottenga, BIOS consortium, Social Science Genetic Association Consortium, D. I. Boomsma, E. J. C. de Geus, M. G. Nivard, M. Bartels, Multivariate genome-wide analyses of the well-being spectrum. *Nat. Genet.* **51**, 445–451 (2019).
48. J.-A. Dias, T. Chen, H. Xing, X. Wang, A. A. Rodriguez, R. K. Madduri, P. Kraft, H. Zhang, Evaluating multi-ancestry genome-wide association methods: Statistical power, population structure, and practical implications. *Am. J. Hum. Genet.* **113**, 399 (2026).
49. J. Morales, D. Welter, E. H. Bowler, M. Cerezo, L. W. Harris, A. C. McMahon, P. Hall, H. A. Junkins, A. Milano, E. Hastings, C. Malangone, A. Buniello, T. Burdett, P. Flicek, H. Parkinson, F. Cunningham, L. A. Hindorff, J. A. L. MacArthur, A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* **19**, 21 (2018).
50. A. R. Wood, T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson, A. Y. Chu, K. Estrada, J. Luan, Z. Kutalik, N. Amin, M. L. Buchkovich, D. C. Croteau-Chonka, F. R. Day, Y. Duan, T. Fall, R. Fehrmann, T. Ferreira, A. U. Jackson, J. Karjalainen, K. S. Lo, A. E. Locke, R. Mägi, E. Mihailov, E. Porcu, J. C. Randall, A. Scherag, A. A. E. Vinkhuyzen, H.-J. Westra, T. W. Winkler, T. Workalemahu, J. H. Zhao, D. Absher, E. Albrecht, D. Anderson, J. Baron, M. Beekman, A. Demirkan, G. B. Ehret, B. Feenstra, M. F. Feitosa, K. Fischer, R. M. Fraser, A. Goel, J. Gong, A. E. Justice, S. Kanoni, M. E. Kleber, K. Kristiansson, U. Lim, V. Lotay, J. C. Lui, M. Mangino, I. Mateo Leach, C. Medina-Gomez, M. A. Nalls, D. R. Nyholt, C. D. Palmer, D. Pasko, S. Pechlivanis, I. Prokopenko, J. S. Ried, S. Ripke, D. Shungin, A. Stancáková, R. J. Strawbridge, Y. J. Sung, T. Tanaka, A. Teumer, S. Trompet, S. W. van der Laan, J. van Setten, J. V. Van Vliet-Ostaptchouk, Z. Wang, L. Yengo, W. Zhang, U. Afzal, J. Arnlöv, G. M. Arscott, S. Bandinelli, A. Barrett, C. Bellis, A. J. Bennett, C. Berne, M. Blüher, J. L. Bolton, Y. Böttcher, H. A. Boyd, M. Bruinenberg, B. M. Buckley, S.

Buyske, I. H. Caspersen, P. S. Chines, R. Clarke, S. Claudi-Boehm, M. Cooper, E. W. Daw, P. A. De Jong, J. Deelen, G. Delgado, J. C. Denny, R. Dhonukshe-Rutten, M. Dimitriou, A. S. F. Doney, M. Dörr, N. Eklund, E. Eury, L. Folkersen, M. E. Garcia, F. Geller, V. Giedraitis, A. S. Go, H. Grallert, T. B. Grammer, J. Gräßler, H. Grönberg, L. C. P. G. M. de Groot, C. J. Groves, J. Haessler, P. Hall, T. Haller, G. Hallmans, A. Hannemann, C. A. Hartman, M. Hassinen, C. Hayward, N. L. Heard-Costa, Q. Helmer, G. Hemani, A. K. Henders, H. L. Hillege, M. A. Hlatky, W. Hoffmann, P. Hoffmann, O. Holmen, J. J. Houwing-Duistermaat, T. Illig, A. Isaacs, A. L. James, J. Jeff, B. Johansen, Å. Johansson, J. Jolley, T. Juliusdottir, J. Juntila, A. N. Kho, L. Kinnunen, N. Klopp, T. Kocher, W. Kratzer, P. Lichtner, L. Lind, J. Lindström, S. Lobbens, M. Lorentzon, Y. Lu, V. Lyssenko, P. K. E. Magnusson, A. Mahajan, M. Maillard, W. L. McArdle, C. A. McKenzie, S. McLachlan, P. J. McLaren, C. Menni, S. Merger, L. Milani, A. Moayyeri, K. L. Monda, M. A. Morken, G. Müller, M. Müller-Nurasyid, A. W. Musk, N. Narisu, M. Nauck, I. M. Nolte, M. M. Nöthen, L. Oozageer, S. Pilz, N. W. Rayner, F. Renstrom, N. R. Robertson, L. M. Rose, R. Roussel, S. Sanna, H. Scharnagl, S. Scholtens, F. R. Schumacher, H. Schunkert, R. A. Scott, J. Sehmi, T. Seufferlein, J. Shi, K. Silventoinen, J. H. Smit, A. V. Smith, J. Smolonska, A. V. Stanton, K. Stirrups, D. J. Stott, H. M. Stringham, J. Sundström, M. A. Swertz, A.-C. Syvänen, B. O. Tayo, G. Thorleifsson, J. P. Tyrer, S. van Dijk, N. M. van Schoor, N. van der Velde, D. van Heemst, F. V. A. van Oort, S. H. Vermeulen, N. Verweij, J. M. Vonk, L. L. Waite, M. Waldenberger, R. Wennauer, L. R. Wilkens, C. Willenborg, T. Wilsgaard, M. K. Wojczynski, A. Wong, A. F. Wright, Q. Zhang, D. Arveiler, S. J. L. Bakker, J. Beilby, R. N. Bergman, S. Bergmann, R. Biffar, J. Blangero, D. I. Boomsma, S. R. Bornstein, P. Bovet, P. Brambilla, M. J. Brown, H. Campbell, M. J. Caulfield, A. Chakravarti, R. Collins, F. S. Collins, D. C. Crawford, L. A. Cupples, J. Danesh, U. de Faire, H. M. den Ruijter, R. Erbel, J. Erdmann, J. G. Eriksson, M. Farrall, E. Ferrannini, J. Ferrières, I. Ford, N. G. Forouhi, T. Forrester, R. T. Gansevoort, P. V. Gejman, C. Gieger, A. Golay, O. Gottesman, V. Gudnason, U. Gyllenstein, D. W. Haas, A. S. Hall, T. B. Harris, A. T. Hattersley, A. C. Heath, C. Hengstenberg, A. A. Hicks, L. A. Hindorff, A. D. Hingorani, A. Hofman, G. K. Hovingh, S. E. Humphries, S. C. Hunt, E. Hyppönen, K. B. Jacobs, M.-R. Jarvelin, P. Jousilahti, A. M. Jula, J. Kaprio, J. J. P. Kastelein, M. Kayser, F. Kee, S. M. Keinänen-Kiukaanniemi, L. A. Kiemeny, J. S. Kooner, C. Kooperberg, S. Koskinen, P. Kovacs, A. T. Kraja, M. Kumari, J. Kuusisto, T. A. Lakka, C. Langenberg, L. Le Marchand, T. Lehtimäki, S. Lupoli, P. A. F. Madden, S. Männistö, P. Manunta, A. Marette, T. C. Matise, B. McKnight, T. Meitinger, F. L. Moll, G. W. Montgomery, A. D. Morris, A. P. Morris, J. C. Murray, M. Nelis, C. Ohlsson, A. J. Oldehinkel, K. K. Ong, W. H. Ouwehand, G. Pasterkamp, A. Peters, P. P. Pramstaller, J. F. Price, L. Qi, O. T. Raitakari, T. Rankinen, D. C. Rao, T. K. Rice, M. Ritchie, I. Rudan, V. Salomaa, N. J. Samani, J. Saramies, M. A. Sarzynski, P. E. H. Schwarz, S. Sebert, P. Sever, A. R. Shuldiner, J. Sinisalo, V. Steinthorsdottir, R. P. Stolk, J.-C. Tardif, A. Tönjes, A. Tremblay, E. Tremoli, J. Virtamo, M.-C. Vohl, Electronic Medical Records and Genomics (eMEMERGE) Consortium, MIGen Consortium, PAGEGE Consortium, LifeLines Cohort Study, P. Amouyel, F. W. Asselbergs, T. L. Assimes, M. Bochud, B. O. Boehm, E. Boerwinkle, E. P. Bottinger, C. Bouchard, S. Cauchi, J. C. Chambers, S. J. Chanock, R. S. Cooper, P. I. W. de Bakker, G. Dedoussis, L. Ferrucci, P. W. Franks, P. Froguel, L. C. Groop, C. A. Haiman, A. Hamsten, M. G. Hayes, J. Hui, D. J. Hunter, K. Hveem, J. W. Jukema, R. C. Kaplan, M. Kivimäki, D. Kuh, M. Laakso, Y. Liu, N. G. Martin, W. März, M. Melbye, S. Moebus, P. B. Munroe, I. Njølstad, B. A. Oostra, C. N. A. Palmer, N. L. Pedersen, M.

- Perola, L. Pérusse, U. Peters, J. E. Powell, C. Power, T. Quertermous, R. Rauramaa, E. Reinmaa, P. M. Ridker, F. Rivadeneira, J. I. Rotter, T. E. Saaristo, D. Saleheen, D. Schlessinger, P. E. Slagboom, H. Snieder, T. D. Spector, K. Strauch, M. Stumvoll, J. Tuomilehto, M. Uusitupa, P. van der Harst, H. Völzke, M. Walker, N. J. Wareham, H. Watkins, H.-E. Wichmann, J. F. Wilson, P. Zanen, P. Deloukas, I. M. Heid, C. M. Lindgren, K. L. Mohlke, E. K. Speliotes, U. Thorsteinsdottir, I. Barroso, C. S. Fox, K. E. North, D. P. Strachan, J. S. Beckmann, S. I. Berndt, M. Boehnke, I. B. Borecki, M. I. McCarthy, A. Metspalu, K. Stefansson, A. G. Uitterlinden, C. M. van Duijn, L. Franke, C. J. Willer, A. L. Price, G. Lettre, R. J. F. Loos, M. N. Weedon, E. Ingelsson, J. R. O'Connell, G. R. Abecasis, D. I. Chasman, M. E. Goddard, P. M. Visscher, J. N. Hirschhorn, T. M. Frayling, Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
51. A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, C. Powell, S. Vedantam, M. L. Buchkovich, J. Yang, D. C. Croteau-Chonka, T. Esko, T. Fall, T. Ferreira, S. Gustafsson, Z. Kutalik, J. Luan, R. Mägi, J. C. Randall, T. W. Winkler, A. R. Wood, T. Workalemahu, J. D. Faul, J. A. Smith, J. H. Zhao, W. Zhao, J. Chen, R. Fehrmann, Å. K. Hedman, J. Karjalainen, E. M. Schmidt, D. Absher, N. Amin, D. Anderson, M. Beekman, J. L. Bolton, J. L. Bragg-Gresham, S. Buyske, A. Demirkan, G. Deng, G. B. Ehret, B. Feenstra, M. F. Feitosa, K. Fischer, A. Goel, J. Gong, A. U. Jackson, S. Kanoni, M. E. Kleber, K. Kristiansson, U. Lim, V. Lotay, M. Mangino, I. M. Leach, C. Medina-Gomez, S. E. Medland, M. A. Nalls, C. D. Palmer, D. Pasko, S. Pechlivanis, M. J. Peters, I. Prokopenko, D. Shungin, A. Stančáková, R. J. Strawbridge, Y. J. Sung, T. Tanaka, A. Teumer, S. Trompet, S. W. van der Laan, J. van Setten, J. V. Van Vliet-Ostaptchouk, Z. Wang, L. Yengo, W. Zhang, A. Isaacs, E. Albrecht, J. Ärnlöv, G. M. Arscott, A. P. Attwood, S. Bandinelli, A. Barrett, I. N. Bas, C. Bellis, A. J. Bennett, C. Berne, R. Blagieva, M. Blüher, S. Böhringer, L. L. Bonnycastle, Y. Böttcher, H. A. Boyd, M. Bruinenberg, I. H. Caspersen, Y.-D. I. Chen, R. Clarke, E. W. Daw, A. J. M. de Craen, G. Delgado, M. Dimitriou, A. S. F. Doney, N. Eklund, K. Estrada, E. Eury, L. Folkersen, R. M. Fraser, M. E. Garcia, F. Geller, V. Giedraitis, B. Gigante, A. S. Go, A. Golay, A. H. Goodall, S. D. Gordon, M. Gorski, H.-J. Grabe, H. Grallert, T. B. Grammer, J. Gräßler, H. Grönberg, C. J. Groves, G. Gusto, J. Haessler, P. Hall, T. Haller, G. Hallmans, C. A. Hartman, M. Hassinen, C. Hayward, N. L. Heard-Costa, Q. Helmer, C. Hengstenberg, O. Holmen, J.-J. Hottenga, A. L. James, J. M. Jeff, Å. Johansson, J. Jolley, T. Juliusdottir, L. Kinnunen, W. Koenig, M. Koskenvuo, W. Kratzer, J. Laitinen, C. Lamina, K. Leander, N. R. Lee, P. Lichtner, L. Lind, J. Lindström, K. S. Lo, S. Lobbens, R. Lorbeer, Y. Lu, F. Mach, P. K. E. Magnusson, A. Mahajan, W. L. McArdle, S. McLachlan, C. Menni, S. Merger, E. Mihailov, L. Milani, A. Moayyeri, K. L. Monda, M. A. Morken, A. Mulas, G. Müller, M. Müller-Nurasyid, A. W. Musk, R. Nagaraja, M. M. Nöthen, I. M. Nolte, S. Pilz, N. W. Rayner, F. Renstrom, R. Rettig, J. S. Ried, S. Ripke, N. R. Robertson, L. M. Rose, S. Sanna, H. Scharnagl, S. Scholtens, F. R. Schumacher, W. R. Scott, T. Seufferlein, J. Shi, A. V. Smith, J. Smolonska, A. V. Stanton, V. Steinthorsdottir, K. Stirrups, H. M. Stringham, J. Sundström, M. A. Swertz, A. J. Swift, A.-C. Syvänen, S.-T. Tan, B. O. Tayo, B. Thorand, G. Thorleifsson, J. P. Tyrer, H.-W. Uh, L. Vandenput, F. C. Verhulst, S. H. Vermeulen, N. Verweij, J. M. Vonk, L. L. Waite, H. R. Warren, D. Waterworth, M. N. Weedon, L. R. Wilkens, C. Willenborg, T. Wilsgaard, M. K. Wojczynski, A. Wong, A. F. Wright, Q. Zhang, LifeLines Cohort Study, E. P. Brennan, M. Choi, Z. Dastani, A. W. Drong, P. Eriksson, A. Franco-Cereceda, J. R.

Gådin, A. G. Gharavi, M. E. Goddard, R. E. Handsaker, J. Huang, F. Karpe, S. Kathiresan, S. Keildson, K. Kiryluk, M. Kubo, J.-Y. Lee, L. Liang, R. P. Lifton, B. Ma, S. A. McCarroll, A. J. McKnight, J. L. Min, M. F. Moffatt, G. W. Montgomery, J. M. Murabito, G. Nicholson, D. R. Nyholt, Y. Okada, J. R. B. Perry, R. Dorajoo, E. Reinmaa, R. M. Salem, N. Sandholm, R. A. Scott, L. Stolk, A. Takahashi, T. Tanaka, F. M. van 't Hooft, A. A. E. Vinkhuyzen, H.-J. Westra, W. Zheng, K. T. Zondervan, ADIPOGen Consortium, AGEN-BMI Working Group, CARDIOGRAMplusC4D Consortium, CKDGen Consortium, GLGC, ICBP, MAGIC Investigators, MuTHER Consortium, MIGen Consortium, PAGE Consortium, ReproGen Consortium, GENIE Consortium, International Endogene Consortium, A. C. Heath, D. Arveiler, S. J. L. Bakker, J. Beilby, R. N. Bergman, J. Blangero, P. Bovet, H. Campbell, M. J. Caulfield, G. Cesana, A. Chakravarti, D. I. Chasman, P. S. Chines, F. S. Collins, D. C. Crawford, L. A. Cupples, D. Cusi, J. Danesh, U. de Faire, H. M. den Ruijter, A. F. Dominiczak, R. Erbel, J. Erdmann, J. G. Eriksson, M. Farrall, S. B. Felix, E. Ferrannini, J. Ferrières, I. Ford, N. G. Forouhi, T. Forrester, O. H. Franco, R. T. Gansevoort, P. V. Gejman, C. Gieger, O. Gottesman, V. Gudnason, U. Gyllenstein, A. S. Hall, T. B. Harris, A. T. Hattersley, A. A. Hicks, L. A. Hindorf, A. D. Hingorani, A. Hofman, G. Homuth, G. K. Hovingh, S. E. Humphries, S. C. Hunt, E. Hyppönen, T. Illig, K. B. Jacobs, M.-R. Jarvelin, K.-H. Jöckel, B. Johansen, P. Jousilahti, J. W. Jukema, A. M. Jula, J. Kaprio, J. J. P. Kastelein, S. M. Keinanen-Kiukkaanniemi, L. A. Kiemeny, P. Knekt, J. S. Kooner, C. Kooperberg, P. Kovacs, A. T. Kraja, M. Kumari, J. Kuusisto, T. A. Lakka, C. Langenberg, L. L. Marchand, T. Lehtimäki, V. Lyssenko, S. Männistö, A. Marette, T. C. Matise, C. A. McKenzie, B. McKnight, F. L. Moll, A. D. Morris, A. P. Morris, J. C. Murray, M. Nelis, C. Ohlsson, A. J. Oldehinkel, K. K. Ong, P. A. F. Madden, G. Pasterkamp, J. F. Peden, A. Peters, D. S. Postma, P. P. Pramstaller, J. F. Price, L. Qi, O. T. Raitakari, T. Rankinen, D. C. Rao, T. K. Rice, P. M. Ridker, J. D. Rioux, M. D. Ritchie, I. Rudan, V. Salomaa, N. J. Samani, J. Saramies, M. A. Sarzynski, H. Schunkert, P. E. H. Schwarz, P. Sever, A. R. Shuldiner, J. Sinisalo, R. P. Stolk, K. Strauch, A. Tönjes, D.-A. Trégouët, A. Tremblay, E. Tremoli, J. Virtamo, M.-C. Vohl, U. Völker, G. Waeber, G. Willemssen, J. C. Witteman, M. C. Zillikens, L. S. Adair, P. Amouyel, F. W. Asselbergs, T. L. Assimes, M. Bochud, B. O. Boehm, E. Boerwinkle, S. R. Bornstein, E. P. Bottinger, C. Bouchard, S. Cauchi, J. C. Chambers, S. J. Chanock, R. S. Cooper, P. I. W. de Bakker, G. Dedoussis, L. Ferrucci, P. W. Franks, P. Froguel, L. C. Groop, C. A. Haiman, A. Hamsten, J. Hui, D. J. Hunter, K. Hveem, R. C. Kaplan, M. Kivimäki, D. Kuh, M. Laakso, Y. Liu, N. G. Martin, W. März, M. Melbye, A. Metspalu, S. Moebus, P. B. Munroe, I. Njølstad, B. A. Oostra, C. N. A. Palmer, N. L. Pedersen, M. Perola, L. Pérusse, U. Peters, C. Power, T. Quertermous, R. Rauramaa, F. Rivadeneira, T. E. Saaristo, D. Saleheen, N. Sattar, E. E. Schadt, D. Schlessinger, P. E. Slagboom, H. Snieder, T. D. Spector, U. Thorsteinsdottir, M. Stumvoll, J. Tuomilehto, A. G. Uitterlinden, M. Uusitupa, P. van der Harst, M. Walker, H. Wallaschofski, N. J. Wareham, H. Watkins, D. R. Weir, H.-E. Wichmann, J. F. Wilson, P. Zanen, I. B. Borecki, P. Deloukas, C. S. Fox, I. M. Heid, J. R. O'Connell, D. P. Strachan, K. Stefansson, C. M. van Duijn, G. R. Abecasis, L. Franke, T. M. Frayling, M. I. McCarthy, P. M. Visscher, A. Scherag, C. J. Willer, M. Boehnke, K. L. Mohlke, C. M. Lindgren, J. S. Beckmann, I. Barroso, K. E. North, E. Ingelsson, J. N. Hirschhorn, R. J. F. Loos, E. K. Speliotes, Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

52. A. Okbay, J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers, C. A. Rietveld, P. Turley, G.-B. Chen, V. Emilsson, S. F. W. Meddens, S. Oskarsson, J. K. Pickrell, K. Thom, P. Timshel, R. de Vlaming, A. Abdellaoui, T. S. Ahluwalia, J. Bacelis, C. Baumbach, G. Bjornsdottir, J. H. Brandsma, M. Pina Concas, J. Derringer, N. A. Furlotte, T. E. Galesloot, G. Girotto, R. Gupta, L. M. Hall, S. E. Harris, E. Hofer, M. Horikoshi, J. E. Huffman, K. Kaasik, I. P. Kalafati, R. Karlsson, A. Kong, J. Lahti, S. J. van der Lee, C. deLeeuw, P. A. Lind, K.-O. Lindgren, T. Liu, M. Mangino, J. Marten, E. Mihailov, M. B. Miller, P. J. van der Most, C. Oldmeadow, A. Payton, N. Pervjakova, W. J. Peyrot, Y. Qian, O. Raitakari, R. Rueedi, E. Salvi, B. Schmidt, K. E. Schraut, J. Shi, A. V. Smith, R. A. Poot, B. St Pourcain, A. Teumer, G. Thorleifsson, N. Verweij, D. Vuckovic, J. Wellmann, H.-J. Westra, J. Yang, W. Zhao, Z. Zhu, B. Z. Alizadeh, N. Amin, A. Bakshi, S. E. Baumeister, G. Biino, K. Bønnelykke, P. A. Boyle, H. Campbell, F. P. Cappuccio, G. Davies, J.-E. De Neve, P. Deloukas, I. Demuth, J. Ding, P. Eibich, L. Eisele, N. Eklund, D. M. Evans, J. D. Faul, M. F. Feitosa, A. J. Forstner, I. Gandin, B. Gunnarsson, B. V. Halldórsson, T. B. Harris, A. C. Heath, L. J. Hocking, E. G. Holliday, G. Homuth, M. A. Horan, J.-J. Hottenga, P. L. de Jager, P. K. Joshi, A. Jugessur, M. A. Kaakinen, M. Kähönen, S. Kanoni, L. Keltigangas-Järvinen, L. A. L. M. Kiemeny, I. Kolcic, S. Koskinen, A. T. Kraja, M. Kroh, Z. Kutalik, A. Latvala, L. J. Launer, M. P. Lebreton, D. F. Levinson, P. Lichtenstein, P. Lichtner, D. C. M. Liewald, LifeLines Cohort Study, A. Loukola, P. A. Madden, R. Mägi, T. Mäki-Opas, R. E. Marioni, P. Marques-Vidal, G. A. Meddens, G. McMahon, C. Meisinger, T. Meitinger, Y. Milaneschi, L. Milani, G. W. Montgomery, R. Myhre, C. P. Nelson, D. R. Nyholt, W. E. R. Ollier, A. Palotie, L. Paternoster, N. L. Pedersen, K. E. Petrovic, D. J. Porteous, K. Räikkönen, S. M. Ring, A. Robino, O. Rostapshova, I. Rudan, A. Rustichini, V. Salomaa, A. R. Sanders, A.-P. Sarin, H. Schmidt, R. J. Scott, B. H. Smith, J. A. Smith, J. A. Staessen, E. Steinhausen-Thiessen, K. Strauch, A. Terracciano, M. D. Tobin, S. Ulivi, S. Vaccargiu, L. Quaye, F. J. A. van Rooij, C. Venturini, A. A. E. Vinkhuyzen, U. Völker, H. Völzke, J. M. Vonk, D. Vozzi, J. Waage, E. B. Ware, G. Willemsen, J. R. Attia, D. A. Bennett, K. Berger, L. Bertram, H. Bisgaard, D. I. Boomsma, I. B. Borecki, U. Bültmann, C. F. Chabris, F. Cucca, D. Cusi, I. J. Deary, G. V. Dedoussis, C. M. van Duijn, J. G. Eriksson, B. Franke, L. Franke, P. Gasparini, P. V. Gejman, C. Gieger, H.-J. Grabe, J. Gratten, P. J. F. Groenen, V. Gudnason, P. van der Harst, C. Hayward, D. A. Hinds, W. Hoffmann, E. Hyppönen, W. G. Iacono, B. Jacobsson, M.-R. Jarvelin, K.-H. Jöckel, J. Kaprio, S. L. R. Kardina, T. Lehtimäki, S. F. Lehrer, P. K. E. Magnusson, N. G. Martin, M. McGue, A. Metspalu, N. Pendleton, B. W. J. H. Penninx, M. Perola, N. Pirastu, M. Pirastu, O. Polasek, D. Posthuma, C. Power, M. A. Province, N. J. Samani, D. Schlessinger, R. Schmidt, T. I. A. Sørensen, T. D. Spector, K. Stefansson, U. Thorsteinsdottir, A. R. Thurik, N. J. Timpson, H. Tiemeier, J. Y. Tung, A. G. Uitterlinden, V. Vitart, P. Vollenweider, D. R. Weir, J. F. Wilson, A. F. Wright, D. C. Conley, R. F. Krueger, G. Davey Smith, A. Hofman, D. I. Laibson, S. E. Medland, M. N. Meyer, J. Yang, M. Johannesson, P. M. Visscher, T. Esko, P. D. Koellinger, D. Cesarini, D. J. Benjamin, Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
53. A. Mahajan, J. Wessel, S. M. Willems, W. Zhao, N. R. Robertson, A. Y. Chu, W. Gan, H. Kitajima, D. Taliun, N. W. Rayner, X. Guo, Y. Lu, M. Li, R. A. Jensen, Y. Hu, S. Huo, K. K. Lohman, W. Zhang, J. P. Cook, B. P. Prins, J. Flannick, N. Grarup, V. V. Trubetskoy, J. Kravic, Y. J. Kim, D. V. Rybin, H. Yaghootkar, M. Müller-Nurasyid, K. Meidtner, R. Li-Gao, T. V. Varga, J. Marten, J. Li, A. V. Smith, P. An, S. Ligthart, S. Gustafsson, G. Malerba,

- A. Demirkan, J. F. Tajes, V. Steinthorsdottir, M. Wuttke, C. Lecoeur, M. Preuss, L. F. Bielak, M. Graff, H. M. Highland, A. E. Justice, D. J. Liu, E. Marouli, G. M. Peloso, H. R. Warren, ExomeBP Consortium, MAGIC Consortium, GIANT Consortium, S. Afaq, S. Afzal, E. Ahlqvist, P. Almgren, N. Amin, L. B. Bang, A. G. Bertoni, C. Bombieri, J. Bork-Jensen, I. Brandslund, J. A. Brody, N. P. Burt, M. Canouil, Y.-D. I. Chen, Y. S. Cho, C. Christensen, S. V. Eastwood, K.-U. Eckardt, K. Fischer, G. Gambaro, V. Giedraitis, M. L. Grove, H. G. de Haan, S. Hackinger, Y. Hai, S. Han, A. Tybjærg-Hansen, M.-F. Hivert, B. Isomaa, S. Jäger, M. E. Jørgensen, T. Jørgensen, A. Käräjämäki, B.-J. Kim, S. S. Kim, H. A. Koistinen, P. Kovacs, J. Kriebel, F. Kronenberg, K. Läll, L. A. Lange, J.-J. Lee, B. Lehne, H. Li, K.-H. Lin, A. Linneberg, C.-T. Liu, J. Liu, M. Loh, R. Mägi, V. Mamakou, R. McKean-Cowdin, G. Nadkarni, M. Neville, S. F. Nielsen, I. Ntalla, P. A. Peyser, W. Rathmann, K. Rice, S. S. Rich, L. Rode, O. Rolandsson, S. Schönherr, E. Selvin, K. S. Small, A. Stančáková, P. Surendran, K. D. Taylor, T. M. Teslovich, B. Thorand, G. Thorleifsson, A. Tin, A. Tönjes, A. Varbo, D. R. Witte, A. R. Wood, P. Yajnik, J. Yao, L. Yengo, R. Young, P. Amouyel, H. Boeing, E. Boerwinkle, E. P. Bottinger, R. Chowdhury, F. S. Collins, G. Dedoussis, A. Dehghan, P. Deloukas, M. M. Ferrario, J. Ferrières, J. C. Florez, P. Frossard, V. Gudnason, T. B. Harris, S. R. Heckbert, J. M. M. Howson, M. Ingelsson, S. Kathiresan, F. Kee, J. Kuusisto, C. Langenberg, L. J. Launer, C. M. Lindgren, S. Männistö, T. Meitinger, O. Melander, K. L. Mohlke, M. Moitry, A. D. Morris, A. D. Murray, R. de Mutsert, M. Orho-Melander, K. R. Owen, M. Perola, A. Peters, M. A. Province, A. Rasheed, P. M. Ridker, F. Rivadineira, F. R. Rosendaal, A. H. Rosengren, V. Salomaa, W. H.-H. Sheu, R. Sladek, B. H. Smith, K. Strauch, A. G. Uitterlinden, R. Varma, C. J. Willer, M. Blüher, A. S. Butterworth, J. C. Chambers, D. I. Chasman, J. Danesh, C. van Duijn, J. Dupuis, O. H. Franco, P. W. Franks, P. Froguel, H. Grallert, L. Groop, B.-G. Han, T. Hansen, A. T. Hattersley, C. Hayward, E. Ingelsson, S. L. R. Kardia, F. Karpe, J. S. Kooner, A. Köttgen, K. Kuulasmaa, M. Laakso, X. Lin, L. Lind, Y. Liu, R. J. F. Loos, J. Marchini, A. Metspalu, D. Mook-Kanamori, B. G. Nordestgaard, C. N. A. Palmer, J. S. Pankow, O. Pedersen, B. M. Psaty, R. Rauramaa, N. Sattar, M. B. Schulze, N. Soranzo, T. D. Spector, K. Stefansson, M. Stumvoll, U. Thorsteinsdottir, T. Tuomi, J. Tuomilehto, N. J. Wareham, J. G. Wilson, E. Zeggini, R. A. Scott, I. Barroso, T. M. Frayling, M. O. Goodarzi, J. B. Meigs, M. Boehnke, D. Saleheen, A. P. Morris, J. I. Rotter, M. I. McCarthy, Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571 (2018).
54. M. Nikpay, A. Goel, H.-H. Won, L. M. Hall, C. Willenborg, S. Kanoni, D. Saleheen, T. Kyriakou, C. P. Nelson, J. C. Hopewell, T. R. Webb, L. Zeng, A. Dehghan, M. Alver, S. M. Armasu, K. Auro, A. Bjornd, D. I. Chasman, S. Chen, I. Ford, N. Franceschini, C. Gieger, C. Grace, S. Gustafsson, J. Huang, S.-J. Hwang, Y. K. Kim, M. E. Kleber, K. W. Lau, X. Lu, Y. Lu, L.-P. Lyytikäinen, E. Mihailov, A. C. Morrison, N. Pervjakova, L. Qu, L. M. Rose, E. Salfati, R. Saxena, M. Scholz, A. V. Smith, E. Tikkanen, A. Uitterlinden, X. Yang, W. Zhang, W. Zhao, M. de Andrade, P. S. de Vries, N. R. van Zuydam, S. S. Anand, L. Bertram, F. Beutner, G. Dedoussis, P. Frossard, D. Gauguier, A. H. Goodall, O. Gottesman, M. Haber, B.-G. Han, J. Huang, S. Jalilzadeh, T. Kessler, I. R. König, L. Lannfelt, W. Lieb, L. Lind, C. M. Lindgren, M.-L. Lokki, P. K. Magnusson, N. H. Mallick, N. Mehra, T. Meitinger, F.-U.-R. Memon, A. P. Morris, M. S. Nieminen, N. L. Pedersen, A. Peters, L. S. Rallidis, A. Rasheed, M. Samuel, S. H. Shah, J. Sinisalo, K. E. Stirrups, S. Trompet, L. Wang, K. S. Zaman, D. Ardisino, E. Boerwinkle, I. B. Borecki, E. P. Bottinger, J. E. Buring, J. C.

- Chambers, R. Collins, L. A. Cupples, J. Danesh, I. Demuth, R. Elosua, S. E. Epstein, T. Esko, M. F. Feitosa, O. H. Franco, M. G. Franzosi, C. B. Granger, D. Gu, V. Gudnason, A. S. Hall, A. Hamsten, T. B. Harris, S. L. Hazen, C. Hengstenberg, A. Hofman, E. Ingelsson, C. Iribarren, J. W. Jukema, P. J. Karhunen, B.-J. Kim, J. S. Kooner, I. J. Kullo, T. Lehtimäki, R. J. F. Loos, O. Melander, A. Metspalu, W. März, C. N. Palmer, M. Perola, T. Quertermous, D. J. Rader, P. M. Ridker, S. Ripatti, R. Roberts, V. Salomaa, D. K. Sanghera, S. M. Schwartz, U. Seedorf, A. F. Stewart, D. J. Stott, J. Thiery, P. A. Zalloua, C. J. O'Donnell, M. P. Reilly, T. L. Assimes, J. R. Thompson, J. Erdmann, R. Clarke, H. Watkins, S. Kathiresan, R. McPherson, P. Deloukas, H. Schunkert, N. J. Samani, M. Farrall, A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
55. Schizophrenia Working Group of the Psychiatric Genomics Consortium, Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
 56. International Schizophrenia Consortium, S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, P. Sklar, Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
 57. T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, P. C. Sham, Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480 (2017).
 58. T. Ge, C.-Y. Chen, Y. Ni, Y.-C. A. Feng, J. W. Smoller, Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
 59. C. J. Hoggart, S. W. Choi, J. García-González, T. Souaiaia, M. Preuss, P. F. O'Reilly, BridgePRS leverages shared genetic effects across ancestries to increase polygenic risk score portability. *Nat. Genet.* **56**, 180–186 (2024).
 60. Y. Ruan, Y.-F. Lin, Y.-C. A. Feng, C.-Y. Chen, M. Lam, Z. Guo, Stanley Global Asia Initiatives, L. He, A. Sawa, A. R. Martin, S. Qin, H. Huang, T. Ge, Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022).
 61. L. Xu, G. Zhou, W. Jiang, H. Zhang, Y. Dong, L. Guan, H. Zhao, JointPRS: A data-adaptive framework for multi-population genetic risk prediction incorporating genetic correlation. *Nat. Commun.* **16**, 3841 (2025).
 62. G. Zhou, T. Chen, H. Zhao, SDPRX: A statistical method for cross-population prediction of complex traits. *Am. J. Hum. Genet.* **110**, 13–22 (2023).
 63. G. Wang, A. Sarkar, P. Carbonetto, M. Stephens, A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).
 64. Z. Gerber, M. Fisun, H. Aschard, S. Djebali, PainterPipe: a pipeline for genetic variant fine-mapping using functional annotations. *Bioinform. Adv.* **4**, vbad188 (2024).

65. C. M. Lewis, E. Vassos, Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* **12**, 44 (2020).

1 Chapter 1: A deep population reference panel of tandem repeat variation

Abstract: Tandem repeats (TRs) represent one of the largest sources of genetic variation in humans and are implicated in a range of phenotypes. Here we present a deep characterization of TR variation based on high coverage whole genome sequencing from 3550 diverse individuals from the 1000 Genomes Project and H3Africa cohorts. We develop a method, EnsembleTR, to integrate genotypes from four separate methods resulting in high-quality genotypes at more than 1.7 million TR loci. Our catalog reveals novel sequence features influencing TR heterozygosity, identifies population-specific trinucleotide expansions, and finds hundreds of novel eQTL signals. Finally, we generate a phased haplotype panel which can be used to impute most TRs from nearby single nucleotide polymorphisms (SNPs) with high accuracy. Overall, the TR genotypes and reference haplotype panel generated here will serve as valuable resources for future genome-wide and population-wide studies of TRs and their role in human phenotypes.

1.1 Introduction

The availability of whole genome sequencing (WGS) datasets from thousands of individuals has enabled characterization of human genetic variation at unprecedented scale. Initial variant discovery efforts using low-coverage WGS were focused on single nucleotide polymorphisms (SNPs) and short insertions or deletions (indels)¹. More recently, high-coverage WGS has enabled more accurate catalogs of short indels and structural variation². Although multiple large WGS datasets now exist²⁻⁴, variants in tandem repeat (TR) regions are largely underrepresented, in part because they require more specialized bioinformatics approaches.

Here we consider two types of TRs: short tandem repeats (STRs) consist of repeat units of 1-6bp in tandem, whereas variable number tandem repeats (VNTRs) have longer repeat units. TRs experience rapid mutation rates that result in frequent changes in copy number⁵. Collectively, they comprise around 3% of the human genome⁶ and occur at more than 2 million distinct loci^{7,8}. TRs have been implicated in a variety of Mendelian disorders⁹ and complex traits¹⁰. Although TRs represent one of the largest sources of human genetic variation, they are technically challenging to genotype, and are only partially captured by general SNP and indel genotyping tools used in standard variant calling pipelines.

Over the last decade, TR genotyping has rapidly matured. Variants, including large expansions, at STRs and VNTRs can now be reliably detected by multiple methods from high-coverage short read WGS^{7,8,11-14}. These methods have been applied to catalog genome-wide TR variation across thousands of individuals from diverse populations. One of the earliest catalogs profiled nearly 700,000 STRs using low-coverage WGS from phase 1 of the 1000 Genomes Project (1000GP) cohort¹⁵. Subsequent studies have analyzed TR variation in deep WGS from other cohorts^{4,10,16-20}. However, these have faced important limitations. Most available large WGS datasets have been biased toward European individuals. Those including more diverse populations were either low-coverage, resulting in low accuracy and high rates of missing genotypes, or had relatively small sample size.

Another important limitation of existing TR catalogs is that none provides a comprehensive view of TR variation. Most tools begin with a reference set of TRs based on a reference genome. However, reference sets vary dramatically across tools due to differences in parameters used to define repeats, which are often based on limitations of individual genotyping approaches. For example, GangSTR¹³ can identify large expansions but only genotypes TRs with no sequence

imperfections, whereas imperfect repeats are considered by HipSTR⁸. ExpansionHunter¹² models imperfect repeats, but the reference set must be semi-manually defined by the user and may differ from that used by other tools. Further, the set of repeat unit lengths considered differs by tool (HipSTR considers 1-6bp units, GangSTR 1-20bp, adVNTR 6+bp). Thus, no single tool captures the full spectrum of TR variation.

Here, we develop a new method, EnsembleTR, which takes TR genotypes output by existing tools (currently ExpansionHunter, adVNTR, HipSTR, and GangSTR) as input, and outputs a consensus TR callset by converting TR genotypes to a consistent internal representation and using a voting-based scheme. We apply EnsembleTR to genotype 1.7 million TRs based on the hg38 reference genome across deep PCR-free WGS for 3,202 individuals from the 1000GP² and PCR+ WGS data for 348 individuals from H3Africa Project²¹. We apply this resource to characterize population-specific TR variants, identify novel sequence-context features contributing to TR variability, identify TRs associated with gene expression, and generate an improved phased SNP-TR reference haplotype panel. The full set of phased genotypes as well as population-specific summary statistics are made publicly available to facilitate use by the genomics community. Overall, we envision this will be a powerful resource enabling study of TR variation across a wide range of future applications.

1.2 Results

1.2.1 A genome-wide catalog of TR variation

We performed genome-wide genotyping of TRs using high-coverage PCR-free WGS data available for 3,202 samples from the 1000 Genomes Project (1000GP) and PCR+ WGS data for 348 samples from the H3Africa Project (Methods). Both datasets were sequenced to an average of

approximately 30x coverage. We applied four separate TR genotyping methods which consider a variety of TR classes, including short STRs (HipSTR, ExpansionHunter, GangSTR), STR expansions (GangSTR, ExpansionHunter), and VNTRs (GangSTR, adVNTR). All four methods take as input a reference set of TRs and output inferred diploid repeat lengths in each sample. HipSTR additionally identifies sequence differences between repeat alleles. Genotypes from each method were filtered to remove poor quality calls (Methods).

We next developed a novel ensemble calling method (EnsembleTR) which takes as input VCF files from different TR genotypers and outputs a consensus set of genotypes (Fig. 1.1a). TRs genotyped by a single tool do not require merging and are simply added to the output. EnsembleTR then identifies overlapping TR regions genotyped by two or more tools, infers a mapping between alternate allele sets reported by each method, and outputs a consensus genotype and quality score for each call (Methods). We applied EnsembleTR to jointly genotype all samples, which resulted in consensus calls at 1,785,572 unique TRs (42% homopolymers) on autosomal chromosomes. After removing TRs called in fewer than 75% of samples, 1,714,353 TRs (41% homopolymers) remained. Of those, 55% were only genotyped by a single method (Fig. 1.1b), largely reflecting differences in TR reference sets published for each tool.

We evaluated whether the resulting consensus genotypes capture the expected patterns of genetic variation in our cohort. We first examined patterns of Mendelian inheritance (MI) in the 602 available trios (Methods). Overall, 94% of calls follow MI, and this rate increases with increasingly stringent score thresholds (Fig. 1.1c). TRs called by multiple methods typically show higher MI rates (Fig. 1.1b). Further, TRs called by HipSTR and ExpansionHunter have higher overall MI rates than TRs called by GangSTR and adVNTR (Fig. 1.1b).

Some improvement in MI at TRs genotyped by multiple methods could result from differences in TR composition across reference sets for each method, with TRs considered by multiple methods being potentially easier to genotype compared to those genotyped by a single method. To test this, we computed MI separately for each genotyper at calls on Chromosome 1 that are shared vs. unique across methods. adVNTR was excluded from this analysis since it genotypes a largely distinct set of TRs (Fig. 1.1b). For all methods, MI was highest at calls genotyped by all 3 methods compared to those called by a single method (Supplementary Data 1). This difference was most pronounced for GangSTR, which showed MI of 98.2% at shared calls vs. 88% for calls made by GangSTR alone.

Next, to determine whether ensemble-based calling improves genotype quality beyond what can be explained by differences in TR composition alone, we compared the MI of EnsembleTR calls vs. that obtained by individual methods at TRs on Chromosome 1 genotyped by multiple tools. For most combinations of methods, although in most cases the different genotypers were highly concordant (range 80-95%), we still observed that EnsembleTR calls either exhibited greater Mendelian consistency or achieved comparable MI to the genotyper with the highest MI level (Supplementary Data 2). For example, at 15,250,413 calls genotyped by both HipSTR and GangSTR, the EnsembleTR MI rate is 99.3%, compared to 99.1% for HipSTR and 98.1% for GangSTR. The only exception was for TRs called only by GangSTR and ExpansionHunter but not HipSTR, for which EnsembleTR calls showed lower MI than ExpansionHunter alone. However, these TRs make up a small percentage (approximately 0.1%) of the total callset, and typically are longer (mean 30bp in hg38 vs. 19bp for other TRs) making them more difficult to genotype precisely. We also observed that in 72% of calls that were genotyped by multiple methods but where at least one method did not follow MI, EnsembleTR

calling resulted in Mendelian consistency, compared to 65% obtained by a naïve approach of always choosing the HipSTR genotype. Overall, our results suggest that while TRs called by multiple methods tend to be easier to genotype in general, EnsembleTR still improves call quality over any single method at these loci. For downstream analyses, we filtered calls from TRs with Mendelian error rates above 5% which resulted in 1,443,686 unique TRs.

To further evaluate our callset we performed fragment analysis via capillary electrophoresis to genotype 48 TRs on a subset of samples. Our validation panel includes 11 TRs implicated in repeat expansion disorders, plus an additional 37 TRs spanning a range of repeat classes. Each TR in our panel was tested on 48 samples, including 25 samples chosen to represent diverse population groups from the 1000GP, 17 additional samples from the “Platinum Genomes” multi-generational pedigree (6 of which are included as 1000GP trios), and 6 samples from the Genome In a Bottle project (Supplementary Data 3-7; Figs. 1.2-1.3; Methods). Out of 1,395 mutual calls between EnsembleTR and fragment analysis, 1,362 (98%) were concordant (96% concordance for GangSTR and 98% for HipSTR calls). Of the 33 discordant calls, 9 were from a single TR (C9orf72), a GC-pure repeat (which comprise 0.67% of our catalog). Nearly all errors including those at C9orf72 resulted either from discrepancies of a single unit or by dropout of one of the alleles at a heterozygous locus in either technology. Notably, our validation focused on TRs that could be readily genotyped by PCR, and thus excluded more complex repeats as well as homopolymer loci, for which error rates are likely higher. Still, our results suggest that most non-homopolymer TR genotypes in our catalog based on WGS are of comparable accuracy to those obtained by the experimental gold standard of fragment analysis.

Next, we examined population-specific allele frequencies at well-characterized TRs, including known pathogenic loci and those used for forensics analysis, and found that EnsembleTR

results recapitulated published results for these loci (Methods; Figs. 1.4-1.5), with mean Jensen-Shannon divergence ranging between 0.03-0.21 for forensic loci and 0.1-0.31 for pathogenic loci across all superpopulations (compared to mean 0.43 and 0.55 when permuting loci; Methods). We then examined genome-wide patterns of TR variation across populations. Initial inspection of the number of variant TR alleles per sample showed that H3Africa samples had far higher rates of polymorphism even compared to African samples in the 1000GP (Fig. 1.6). However, we hypothesized this could be driven by the PCR+ nature of the H3Africa samples, which induces high error rates, in particular at homopolymer TRs²². Repeating this analysis, but excluding homopolymer loci, revealed similar patterns as were observed for other classes of variants² (Fig. 1.1d). Individuals from African populations had the highest number of variant TR alleles compared to the reference, whereas Europeans had the fewest. Further, admixed African individuals showed the highest variability in the number of variants per sample. As expected, the rate of discovery of new TR alleles slows with each new individual, but this rate increases after the addition of African samples and continues to increase when H3Africa samples are added (Fig. 1.7). Performing principal component analysis (PCA) on a matrix of the sum of repeat lengths at each TR for each sample captured expected patterns of population structure (Fig. 1.8).

1.2.2 Characterizing population-specific TR variation

We next characterized patterns of TR variation and how they vary across populations. After filtering, we identified an average of 184,056 and 184,759 TRs in each sample for which one or both alleles did not match the reference genome, respectively (Supplementary Data 8). Our callset contains 6,758 TRs entirely inside coding exons, corresponding to 0.5% of TRs genotyped genome-wide (Supplementary Data 9). On average, each sample contained at least one non-reference allele at 295 coding TRs. As expected, TRs with repeat unit lengths that are multiples of

3 are over-represented in coding exons whereas mononucleotide, dinucleotide, and tetranucleotide TRs are far more prevalent in non-coding regions of the genome (Fig. 1.9a). Additionally, a far lower percentage of TRs in coding regions are polymorphic (51% for coding exons compared to 78% genome-wide; Fig. 1.9b).

We then summarized the variability in the length of each TR by computing the heterozygosity (H) (and counting the number of alleles with frequency $\geq 1\%$; Fig. 1.10; Methods). TRs in our catalog show a wide range of polymorphism rates, with 56% of non-homopolymer TRs (6% of homopolymers) fixed or nearly fixed at a single allele length ($H < 0.001$), 11% (29% of homopolymers) with two common alleles, and 17% (50% of homopolymers) with three or more common alleles. TR heterozygosity and the number of common alleles are highly correlated across populations (Figs. 1.11-1.12), with few TRs being polymorphic in one population but not others.

The majority of alleles identified in each sample differ in length from the reference genome by only a small number of repeat units, and this trend is consistent across populations (Fig. 1.13a-b). All populations show a slight bias toward alleles that are shorter than the reference allele, and exhibit the highest rates of variation at homopolymer TRs. Overall, alleles closest in length to the reference allele tend to be common, whereas alleles tend to decrease in frequency as a function of their length difference from the reference (Fig. 1.13c). However, we observed 196 TRs at which more than 95% of observed alleles differed from the reference by more than 2 repeats (Supplementary Data 10). Many of these consisted of highly imperfect repeats or TRs with multiple distinct repeat units. We manually inspected available PacBio “HiFi” reads from a single sample (Methods) overlapping each of these TRs, and found that the EnsembleTR allele was supported at 194/196 loci (examples shown in Fig. 1.14). We further investigated these in the new T2T reference genome²³ and found that for 194/195 TRs that could be successfully lifted over, the

T2T reference matched the most common allele called by EnsembleTR. Overall, this suggests that a subset of complex TRs may not be correctly represented in the hg38 reference but are resolved in T2T.

We also observed a subset of common alleles with large expansions compared to the reference (Fig. 1.13c). To identify population-specific polymorphic repeat expansions, we searched for TRs with common expansions in either Africans or non-Africans but not both. We filtered homopolymer TRs and only considered expansions as outlier alleles with copy number >10 (Methods). This method identified 264 candidate TRs (Supplementary Data 11). Of these, 198 were specifically expanded in Africans and 66 in non-Africans. We additionally applied two methods specifically designed to detect pathogenic repeat expansions (STRetch²⁴ and ExpansionHunter Denovo¹¹) to identify candidate expansions in the H3Africa cohort. Of the 264 candidate TRs, 11 were supported by at least one and 5 were supported by both methods. Two TRs had particularly dramatic Africa-specific expansion alleles (Fig. 1.13d-e), both of which were supported by STRetch and ExpansionHunter Denovo. These include an intronic CAG repeat in *CA10* (1.6% of African alleles have >+65 copies relative to hg38 compared to 0.13% in non-Africans, originally genotyped by ExpansionHunter, HipSTR, and GangSTR) and a CTT repeat upstream of *NEXN* (14% of African alleles have >+39.7 copies relative to hg38 compared to 1.2% in non-Africans, originally genotyped by ExpansionHunter and HipSTR). Notably, previous studies of the *CA10* locus reported expansions at up to 30% of alleles²⁵, far more than reported for any population here. While we still likely underestimate expansion rates (see below), we note these studies used a threshold of 40 total copies to define expansions, compared to 86 total copies here (+65 relative to the 21 copies in hg38) based on our outlier detection approach.

To further validate the *CA10* and *NEXN* expansions, we compared EnsembleTR calls to genotypes obtained by manual inspection of PacBio HiFi reads available for 27 1000GP samples (Supplementary Data 12). Notably, long alleles at both repeats are much longer than the Illumina read length, and so repeat length estimates are inexact. Still, allele lengths estimated by EnsembleTR are strongly correlated with length estimates based on HiFi reads (Pearson $r=0.78/0.89$, two-sided $p=1.5e-6/3.6e-10$ for *CA10* and *NEXN* respectively). Further, all large expansion alleles identified by EnsembleTR were supported by PacBio, although some large expansions were missed as a result of a bias in EnsembleTR's voting scheme which down-weights lower confidence alleles (see Discussion). We additionally performed PCR amplification of the *CA10* repeat in four 1000GP samples with a range of allele lengths, which confirmed EnsembleTR genotypes including a large expansion at this locus (Fig. 1.15). Interestingly, manual inspection of both TRs in HiFi reads revealed common variation not only in TR length, but also in TR sequence. At the *CA10* TR, which is annotated as a CAG repeat in hg38, most alleles are perfect CAG repeats but expansions of CCG or CGG were also observed (Fig. 1.15c). Similarly, at the *NEXN* TR, which is annotated as a CTT repeat, many observed alleles instead consist of repeats of the hexamer sequence CTTCTC. This alternate repeat unit was observed on both expanded and normal range allele lengths.

Finally, we assessed whether these sequence imperfections are captured by HipSTR, the only method used here that attempts to infer the sequence as well as the length of TR alleles. We used TRViz²⁶ to visualize sequence structure for alleles at the *CA10* and *NEXN* TRs (Fig. 1.16). HipSTR successfully identified the CTTCTC motif in shorter alleles at *NEXN* as well as CAC and CAT motifs at the *CA10* locus. Longer alleles could only be genotyped by ExpansionHunter or GangSTR, which report perfect repeats even for genotypes where manual inspection of PacBio

reads confirmed the presence of alternate motifs. Notably, EnsembleTR output annotates the method of origin of each call, which could be used to determine in which cases repeat allele sequence information is reliable.

1.2.3 Sequence determinants of TR heterozygosity

We next used our catalog to examine determinants of polymorphism patterns across different TRs by correlating sequence features with TR heterozygosity. As widely observed previously^{15,27,28}, we found that TR heterozygosity is most strongly correlated with total repeat length and the length of the repeat unit (Fig. 1.17a), with TRs with longer total stretches of uninterrupted repeat sequence and shorter repeat units being typically the most polymorphic. This trend is consistently observed across populations (Fig. 1.18). Among TRs in our catalog with the same repeat unit length, heterozygosity also varies to a lesser extent across different repeat unit sequences (Fig. 1.17b-e). For example, CG and AT dinucleotide repeats have higher average heterozygosities at a given length compared to AG or AC repeats. For tetranucleotides and pentanucleotides, AGAT and AAAAG repeats tend to have the highest heterozygosities across a range of repeat lengths. When visualizing reference TR length in bp vs. abundance in the genome, we additionally observed an unexpected periodic pattern for multiple repeat classes. Trinucleotides with length $0 \bmod 3$ are less abundant than those consisting of a non-integer number of total repeat copies, consistent with a previous observation²⁹. Similarly, dinucleotides with an even total length (e.g. ACACAC) tend to be slightly less abundant than those with an odd total length (e.g. ACACACA). Similar periodic trends were observed for other repeat unit lengths (Fig. 1.17a).

Beyond the sequence of the TR, we reasoned that features of the genomic sequence flanking a TR may also impact its variability. To investigate this further, we focused on non-'GC' dinucleotide STRs in our catalog with repeat units AC/GT, AT, or AG/CT with reference length

between 12-17bp. We first classified TRs as either “stable” (major allele frequency = 1) or “polymorphic” (major allele frequency <0.99), resulting in a set of 3,435 polymorphic STRs and 6,922 stable STRs (Fig. 1.17f). Of these 4,829 had an AC/GT repeat unit, 3,051 had AG/CT, and 2,477 had AT. We then applied two methods to identify sequence features characteristic of stable TRs. First, we applied HOMER³⁰, a motif discovery tool, to sequences extracted from a 64bp window on each side of the TRs of each repeat unit separately. For AC repeats, HOMER identified five motifs enriched in the context of polymorphic vs. stable TRs. Of these, four contain a repetitive motif with a dinucleotide repeat unit (Fig. 1.17g). Similar top motifs were identified for AT, but not AG/CT repeats (Fig. 1.19).

Second, we trained a convolutional neural network (CNN) using 58bp flanking each TR plus 6 directly adjacent bases that make up the TR from both sides, and used gradient-based attribution scores to quantify the importance of each input base (Methods). Our model achieved an overall accuracy of 73% on a held out test set with area under the precision recall curve 0.82/0.62 when considering stable/polymorphic STRs as the target class (precision=0.76/0.64 for stable/polymorphic and recall=0.87/0.46 for stable/polymorphic). Visualization of attribution scores for the TRs most confidently and correctly predicted to be polymorphic identified that nearby dinucleotide repeat-like sequences have the strongest influence on whether the model predicted an STR to be polymorphic (Fig. 1.17h). This pattern was not visible in TRs confidently correctly predicted to be stable (Fig. 1.20). This result is consistent with our HOMER findings that dinucleotide repeat-like sequences in the flanking regions of dinucleotide TRs results in increased heterozygosity.

To validate these results and quantify the strength of the relationship, we found the count of all 4-mers composed of adjacent dinucleotide or mononucleotide motifs (e.g. ATAT, ACAC,

AAAA, etc.) in a 64 bp window around all dinucleotide STRs. When dividing the counts of these repetitive 4-mers in context regions into quintiles, polymorphic STRs are highly prevalent in the upper quintile (27.6% of polymorphic STRs fall into this bin, compared to 15.5% of stable STRs) and depleted in the lowest quintile (which contains 16.1% of polymorphic STRs and 23.1% of stable STRs). Overall, the count of these context 4-mers is significantly correlated with STR heterozygosity both overall and individually for STRs with AT, AC/GT, and AG/CT repeat units (Bonferroni corrected P-values 1.5×10^{-124} , 9.7×10^{-10} , 5.3×10^{-30} , and 1.0×10^{-5}). Although these context dinucleotides are predictive, in all cases the strength of the correlation with these sequence context features is less than the correlation with copy number (Fig. 1.17i; Supplementary Data 13).

1.2.4 Detecting TRs associated with gene expression

To assess the utility of our catalog in identifying trait-associated TRs, we performed expression quantitative trait loci (eQTL) discovery in 452 unrelated samples (363 EUR and 89 AFR) with available RNA-sequencing derived from lymphoblastoid cell lines (LCLs) from the Geuvadis project³¹. We tested for association between the sum of the repeat length of the alleles of each individual and gene expression for each TR within 100kb of each gene (Methods; Supplementary Data 14-15; Data Availability; Fig. 1.21a). Tests were performed separately in the EUR and AFR cohorts. In total, we identified 74,340 (EUR) and 400 (AFR) individual significant TR-gene pairs (FDR<0.05) and 3,664 (EUR) and 81 (AFR) total eGenes (gene-level FDR<0.05). Effect sizes of eTRs significant in at least one cohort (EUR or AFR) were strongly correlated across these two cohorts (Pearson $r=0.40$; $p<10^{-200}$; $n=64,760$; Fig. 1.21b). For comparison, a previous eTR analysis we performed in this cohort³² based on low-coverage WGS from the 1000GP identified only 2,060 eGenes at the same FDR threshold.

We compared eTR effects measured here to eSTRs we identified previously across 17 tissues in the Genotype-Tissue Expression (GTEx) cohort¹⁶. Effect sizes computed for overlapping sets of TR-gene pairs across studies were significantly correlated in all tissues ($p < 10^{-200}$ and $p = 5.7 \times 10^{-12}$ in all tissues considering eTRs significant in European and African Geuvadis cohorts, respectively) but were most strongly correlated with Cultured Fibroblasts (Supplementary Data 16, Fig. 1.21c-d). The previous GTEx analysis excluded LCLs due to low sample size, and so we could not directly compare to data from the same cell type. eQTLs discovered here recapitulate known signals, and also identify novel trait-associated TRs. For example, one of our top eTR signals is a dodecamer repeat in the promoter of *CSTB*, which has been reported by multiple previous studies^{16,33} and is associated with myoclonus epilepsy³⁴ (Fig. 1.21e). We identified a total of 2,996 eTRs significant at the gene-level that were either not previously tested (2,917) or did not reach at least nominal significance (79) in any tissue tested in GTEx. An example novel association of a dinucleotide AT repeat with *TIMM10* expression is shown in Fig. 1.21f.

1.2.5 Phased reference panel allows accurate imputation of TR variants

Finally, we generated a phased reference haplotype panel of SNPs/short indels and TRs from the 1000GP samples. We used our previously published pipeline³⁵ to phase each TR separately onto a backbone of phased SNPs in a +/-50kb window, resulting in a single panel containing both phased SNPs and TRs (Methods). The resulting panel contains a total of 1,091,550 TRs, compared to 453,671 TRs in the previously published panel. We assessed the utility of this panel for imputing TRs by performing a leave-one-out analysis at TRs on chromosome 21 and observed an average concordance of 99% between imputed genotypes and observed genotypes in all 5 superpopulations of 1000GP. For comparison, we performed a naive imputation method in which each genotype is imputed as the most common diploid genotype, which resulted in an

average concordance of 87%. As expected, imputation performance is strongest at the least polymorphic TRs, and weakest at those that are highly multi-allelic and/or have the highest heterozygosity (Fig. 1.22; Fig. 1.23a).

We then evaluated the effectiveness of our panel in imputing medically relevant TRs. Utilizing our panel, we imputed genotypes for 92 repeats, comprising 16 widely recognized pathogenic loci involved in expansion disorders as well as 76 potential causal variants associated with blood traits that were included in our panel (Supplementary Data 17). Notably, our panel is comprised of healthy controls, and thus imputation is restricted to common normal range alleles for the known pathogenic loci. Our analysis revealed an average concordance range 0.95-0.97 across populations for all examined loci. Performance was lower at known pathogenic TRs, with highly polymorphic TRs implicated in Huntington's Disease and spinocerebellar ataxia 3 being most challenging (concordance range 0.68-0.90). Overall the average concordance across all pathogenic loci remained high (range 0.89-0.92 across populations) indicating that common, non-pathogenic alleles at a range of medically relevant repeats can be accurately imputed.

Multiple TRs have recently been implicated as causal drivers of genotype-phenotype associations discovered using genome-wide association studies (GWAS)^{10,36}. In these cases, although the TRs are likely causal, the signals were originally identified using nearby tagging SNPs in at least moderate linkage disequilibrium (LD) with the TR. We used our haplotype panel to explore the ability of nearby SNPs or small indels to tag TRs. We determined the best tag SNP for each TR as the SNP/indel within a $\pm 50\text{kb}$ window with the strongest LD (Supplementary Data 18). As expected, TRs that are largely bi-allelic are often well-tagged by nearby SNPs (mean best tag SNP $r^2=0.90$), whereas the LD of the best tag SNP decreases for TRs with an increasing number of common alleles (Fig. 1.23b). For example, for TRs with five common alleles the mean r^2 of the

best tag SNP in Europeans is 0.68, with generally weaker tagging in the African superpopulation. The majority of tag SNPs are located within a small window around the TR, and in some cases the top tag SNP is within or directly adjacent to the TR (Fig. 1.23c). Overall, these results indicate that while bi-allelic TRs are likely well captured by existing variant panels used by GWAS and other studies, more polymorphic TRs are often not well tagged by a single nearby SNP or indel.

1.3 Discussion

TRs represent some of the most polymorphic regions of the genome, but have so far not been systematically included in large genetic variation databases, in large part due to technical challenges in genotyping as well as discrepancies in how TRs are defined by different tools. Here, we developed a novel framework, EnsembleTR, which uses an ensemble approach to integrate the output of multiple TR genotypers and generate a deep catalog of TR variation in the 1000GP and H3Africa cohorts. After filtering low-quality calls, our ensemble approach genotyped more than 1.7 million TR loci. We applied EnsembleTR to identify population-specific repeat expansions, characterize sequence determinants of TR stability, perform eQTL analysis, and generate a phased TR-SNP reference haplotype panel.

Our TR catalog based on EnsembleTR contains substantially more TRs than can be genotyped with any single method. The largest gain comes from the fact that recommended reference TR sets have low overlap across tools (Fig. 1.1b), with many TRs included only by a single method. This lack of overlap stems from the algorithm design of each tool. For example, whereas HipSTR can genotype TRs with sequence imperfections, GangSTR is restricted to perfect repeats. On the other hand, adVNTR is designed specifically for VNTRs and not STRs. Beyond improving the quantity of TRs called, ensemble-based genotyping also increases genotype quality at TRs called by more than one method. For example, at TRs called by multiple methods in almost

all cases EnsembleTR genotypes showed Mendelian consistency levels that were either better than any individual method or comparable to the best-performing method (Supplementary Data 2).

The TR dataset presented here provides important improvements over previous population-wide TR panels and their applications. Previous TR panels were primarily based on hg19 and on a single TR genotyper^{4,15,37}, or are under controlled access restrictions and have limited representation of diverse ancestries^{16,18,20,36}. In contrast, this dataset is made freely available, is based on the hg38 reference genome, and integrates TR calls from four different tools. We also improve upon published efforts to identify TRs acting as eQTLs. We previously identified eTRs using low-coverage Phase 3 1000GP data³², which was underpowered due to low TR genotyping quality. A separate study performed eTR analysis based on targeted sequencing of promoter TRs but focused on TRs within 1kb of transcription start sites³⁸, therefore missing the majority of TRs. More recently, we analyzed eSTRs¹⁶ and eVNTRs¹⁷ in the GTEx dataset, but excluded LCLs in the STR analysis due to low sample number. The current study shows high concordance with these previous eTR efforts, but identifies 2,996 novel TR-gene associations in LCLs. Finally, we present a new TR-SNP reference haplotype panel, with 1,091,550 loci compared to our previous panel of 453,671 TRs³⁵. This panel can be used for imputing TRs into external GWAS datasets, an approach which has already proven successful by us and others to identify novel trait-associated STRs³⁶ and VNTRs¹⁰.

While overall patterns of TR variation are highly similar across populations, detailed analysis of individual TRs revealed individual loci with population-specific patterns. For example, we identified multiple Africa-specific repeat expansions, including common trinucleotide expansions in an intron of *CA10* and in the promoter of *NEXN* which often involve expansions containing multiple distinct repeat units. These expansions are supported by both the African

cohorts within 1000GP as well as within H3Africa. Common expansions of the repeat in *CA10*, a brain-expressed gene, have been previously reported (previously referred to as the ERDA1 locus³⁹), and have been speculated to be associated with psychiatric disorders⁴⁰. *NEXN* mutations have previously been shown to result in dilated cardiomyopathy⁴¹, which is particularly prevalent in Africa⁴². Because 1000GP and H3Africa do not have phenotype information available, future efforts are needed to determine the potential phenotypic impacts of these population-specific expansions.

This study faced several limitations, many of which will be overcome as sequencing technology and genotyping algorithms continue to improve. First, our TR catalog is based on genotypes obtained from short reads. While this enables reliable genotyping of nearly 2 million TRs, including TRs longer than Illumina read lengths, it is far from comprehensive and many long and complex TRs are still missing. In particular, long alleles with complex structures consisting of multiple repeat units, such as pathogenic alleles in *RFC1* linked to CANVAS⁴³, or the imperfect repeats at long alleles at the *NEXN* and *CA10* repeats described above, are still challenging to resolve with short reads. Long reads such as PacBio HiFi show great promise to genotype the majority of these loci. Intriguingly, the new TRGT method⁴⁴ can reliably infer both repeat length and sequence structure even at complex repeats such as *RFC1*. However, PacBio data is currently only available for several dozen 1000GP samples. Second, while the 1000GP data is PCR-free, WGS from H3Africa is PCR+, likely resulting in high error rates in particular at homopolymer TRs and preventing reliable assessment of repeat expansions specific to that cohort. Third, several technical improvements to the EnsembleTR pipeline may improve future genotyping efforts. For example, it currently only merges TR records from two or more methods if the repeat unit is determined to be identical for that locus across all methods. This occasionally fails, for example at

the *CSTB* promoter TR which is called as a 5-mer by HipSTR but a 12-mer by adVNTR. Further, EnsembleTR currently prioritizes allele lengths that can be most precisely estimated, which in some cases such as the *CA10* TR results in incorrectly choosing high-confidence HipSTR calls over large but inexact expansions identified by ExpansionHunter. Another future improvement is to incorporate ensemble-calling of TRs into pangenome-based methods, which have resulted in important improvements to variant calling at other variant types but are still not optimized for TRs⁴⁵.

Overall, this study presents a dataset of 1.7 million TRs across 3,550 diverse individuals, as well as a phased TR-SNP reference haplotype panel. These calls are made publicly available (see Data Availability) and will serve as an important resource for future efforts to identify population-wide patterns of TR variation and study the effect of genetic variation at TRs on human phenotypes.

1.4 Methods

1.4.1 Dataset description

Whole genome sequencing CRAM files for 3,202 1000GP samples, including 1,598 males and 1,604 females, aligned to GRCh38 were obtained from European Nucleotide Archive accessions PRJEB31736 [<https://www.ebi.ac.uk/ena/browser/view/PRJEB31736>] (unrelated samples) and PRJEB36890 [<https://www.ebi.ac.uk/ena/browser/view/PRJEB36890>] (related samples). Population and superpopulation labels for each sample were obtained from the 1000GP data portal (<https://www.internationalgenome.org/data-portal/sample>). As described on the 1000GP data portal (<https://www.internationalgenome.org/1000-genomes-summary/>), all collections included in the 1000GP Project followed their ethical guidelines and model informed consent language.

CRAM files for 348 H3Africa samples are available from the European Genome-Phenome Archive with accession code EGAS00001002976 [<https://ega-archive.org/studies/EGAS00001002976/>] and were generated as part of the H3AChip Design project. The samples in the H3Africa dataset represent individuals from West, Central, and South African countries. Gender information for samples in the H3Africa cohort was not available to the authors. CRAM files were accessed through the H3Africa Genome Analysis Working Group. As previously described in the original paper²¹, all samples were collected after appropriate approvals had been obtained from local Ethics Boards and Committees in each of the represented countries, and participants gave informed consent.

Information on age of samples was not available for any of the cohorts. Gender of samples was not used in any of the downstream analyses.

1.4.2 TR genotyping with published tools

We first used each tool (HipSTR, GangSTR, adVNTR, ExpansionHunter) to genotype TRs and generate raw calls in VCF format, with a single VCF file per population.

GangSTR: GangSTR¹³ v2.4.5 was run on each sample separately with non-default parameters `--str-info str_info_file` (see below), `--bam-samps sample_id`, `--samp-sex sample_sex`, and `--grid-threshold 250`. We generated an initial set of reference TRs for the hg38 assembly using Tandem Repeats Finder⁴⁶ with the following parameters: `match=2`, `mismatch=5`, `indel=17`, `maxperiod=20`, `pm=80`, `pi=10` and `minscore=24`. We then refined the reference set by applying a series of filtering steps. First, we removed repeats longer than 1kb. Then, we kept a single repeat with the shortest repeat unit length among those with identical start or stop coordinates. Compound and imperfect repeats were removed and any extra bases not matching the repeat motif were trimmed from both sides. Any duplicated repeats were discarded post-trimming. We then removed

any repeats from the reference that did not have a minimum number of 10, 5, 4, and 3 copies for homopolymers, di-, tri-, and tetra/penta/hexa-nucleotide repeats respectively. Finally, we filtered out any overlapping repeats if their motifs consisted of identical nucleotide types. The reference set is available at https://s3.amazonaws.com/gangstr/hg38/genomewide/hg38_ver17.bed.gz.

The file `str_info_file` contains the per-locus stutter parameters obtained by training the stutter model on 19 samples using a modified version of HipSTR v0.6.2 (<https://github.com/mikmakxi/HipSTR>) with non-default parameters `--stutter-model-only` (to skip genotyping), `--chrom` (to run separately for each chromosome), `--min-reads 20`, and `--output-filters.mergeSTR`⁴⁷ v3.0.3 was used to merge the VCF files of each sample into a unified VCF file for each population.

HipSTR: We used HipSTR⁸ v0.6.2 with non-default parameter `--max-reads 2000000` to perform joint autosomal STR genotyping separately for each population using the hg38 STR reference available at https://github.com/HipSTR-Tool/HipSTR-references/blob/master/human/hg38.hipstr_reference.bed.gz.

adVNTR: adVNTR⁷ v1.4.0 was run on each sample separately with a custom reference TR set (https://cseweb.ucsd.edu/~mbakhtia/adVNTR/vntr_data_recommended_loci_hg38.zip). For adVNTR's reference set, a total of 10,264 loci were selected. We started with TRs detected by Tandem Repeats Finder⁴⁶ to identify an initial set of VNTRs located in coding, untranslated, or promoter regions. To identify VNTRs in coding exons and UTRs, we used RefSeq gene coordinates downloaded from UCSC Table Browser⁴⁸. For VNTRs within promoter regions, we considered 500 bp upstream of the transcriptional start sites of genes as the promoter regions. A total of 13,081 VNTRs were identified, of which 10,262 VNTRs were within the size range for short-read genotyping. We subsequently added two VNTRs known to be linked to human disease⁴⁹.

mergeSTR v4.0.1 was used to merge the VCF files of each sample into a single VCF file for each population.

ExpansionHunter: ExpansionHunter¹¹ v5.0.0 was run separately on each sample using a variant catalog of polymorphic STRs (https://github.com/Illumina/RepeatCatalogs/blob/release-v0.1.x/polymorphic_STR/hg38/polymorphic_STR.json). mergeSTR v4.0.1 was used to merge the VCF files of each sample into a unified VCF file for each population.

1.4.3 Filtering initial TR genotypes

Prior to EnsembleTR calling, all population-level VCF files from each tool were leniently filtered with dumpSTR⁴⁷ v3.0.3 with the following options: `--min-locus-callrate 0.75` (to remove TRs with low call rate), `--min-locus-hwep 0.000001` (to remove TRs whose genotypes do not follow Hardy-Weinberg Equilibrium), and `--filter-regions hg38_segdup.sorted.bed.gz --filter-regions-names SEG Dup` (to remove TRs overlapping segmental duplications obtained from the UCSC Genome Browser⁵⁰). For GangSTR, we additionally used options `--gangstr-filter-spanbound-only` and `--gangstr-filter-badCI` to remove low quality calls. For AdVNTR calls, we additionally devised a motif complexity filter, which discards TRs where the consensus motif has imperfect internal repeat structures. For example, a VNTR with consensus motif TTTTCTT may actually be capturing a homopolymer repeat resulting in incorrect VNTR calls.

For this filter, we iteratively mask out characters in the motif and compute the Hamming distance between the masked motif against a sliding window of itself, while allowing the masked character to match any other character in the motif. Let M be a motif of length n , we first compute M' by masking k characters. For a pre-selected k and $i \in [0, \binom{n}{k} - 1]$ (using 0-based string indexing), $M'_{i,k}$ is defined a string matching the original motif except characters that are in the i 'th subset of length k of the motif are masked.

For each $M'_{i,k}$, we concatenate it to itself to form $M''_{i,k}$ to identify any possible internal repetitions. The motif score is then computed as follows:

$$\text{Score}(M) = \max_{i \in (0, \binom{n}{k}-1), j \in (1, n-1)} \text{dist}(M'_{i,k}, M''_{i,k}[j:j+n-1]) \quad (1)$$

where $\text{dist}(x, y)$ gives the Hamming distance between two strings x and y with equal length. The value k is determined based on the motif length n : $k = \left\lceil \frac{n}{10} \right\rceil$ if $n \leq 40$, else 1.

Finally, $\text{Score}(M)$ is normalized by motif length n , to ensure Values are between $[0,1]$, with values closer to 1 indicating the presence of an internal repeat structure. We filtered all VNTRs with motif score > 0.8 .

1.4.4 Merging all populations

Finally, filtered population-level VCF files from each tool were merged using mergeSTR v4.0.1 to generate a single VCF file containing all samples. HipSTR in some cases adjusts the coordinates of an STR region to encompass polymorphic flanking regions around the repeat. In some cases, this can lead to the same STR having slightly different genomic coordinates in the VCF output for different populations. Thus, merging HipSTR VCF files across populations required specific modifications in mergeSTR code. When mergeSTR tries to merge records from different populations, three scenarios can happen: 1) A TR has the same starting position and reference allele sequence in populations A and B, in which case mergeSTR correctly merges them into one record. 2) A TR has different starting positions in populations A and B, in which case mergeSTR writes two distinct records for repeat X in the output. 2) A TR has the same starting position in populations A and B, but the end coordinate and reference allele sequence is different

in these populations, in which case mergeSTR will skip the locus due to the inconsistency in the reference allele. To address this issue, we first modified mergeSTR to write the loci with different reference alleles and identical starting position as multiple records. The modified version of mergeSTR is available at (https://github.com/gymreklab/TRTools/tree/conf_ref). Then, a python script (https://github.com/gymreklab/1000Genomes-STRs/blob/main/Hipstr_correction.py) was used to correct the output VCF file of mergeSTR. First, all the records with the same repeat ID are collected, then the largest overlapping region among all reference alleles is identified and all alleles are trimmed accordingly. If an allele sequence is empty after trimming, all genotypes with that allele were considered as no call. Genotypes are updated based on the new list of alleles and a corrected merged record is written in the output VCF file.

Finally, for GangSTR calls, after merging samples from all populations, we identified repeats with overlapping coordinates and among them, we only kept the first one.

1.4.5 Ensemble genotyping

EnsembleTR takes VCF files from multiple TR genotypes as input and outputs a merged consensus callset. The specific steps of EnsembleTR are described below.

Identifying overlapping TRs between callsets: EnsembleTR starts by finding the mutual samples across all callers. Then EnsembleTR walks through the list of TR loci (records) called by each method in sorted order to identify sets of mergeable calls. Records are deemed mergeable if they have overlapping coordinates and identical repeat unit sequences. EnsembleTR allows at most one record from each caller in each mergeable set.

Matching alleles: Mergeable sets may contain records from multiple callers, each of which might have genotyped a locus using slightly different representations of the possible alleles (see

Fig. 1.1a for an example). To overcome this issue, EnsembleTR forms an internal representation of the consensus set of alleles such that alleles will be directly comparable across methods. It first extends all alleles to the maximum region spanned by all records. In this way, all alleles from different callers will start and end at the same position on the genome. It then extends the original alleles to span this entire region by prepending or appending flanking sequences extracted from the reference genome.

After identifying mergeable alleles, a representative sequence is determined for each allele set. If the mergeable set contains a HipSTR record, EnsembleTR uses the HipSTR allele sequences and discards alleles from other methods with the same length as the HipSTR alleles. This is done because HipSTR is the only method of the four used which reports the actual allele nucleotide sequence rather than only copy numbers. If there are two HipSTR alleles with the same length but different sequences in the allele set, both allele sequences are stored, and the original allele called by HipSTR for each sample is retrieved. In the case that an allele set contains two different HipSTR alleles, but a sample does not have a HipSTR call at that locus, we choose the most common allele of that length output by HipSTR. If a HipSTR record is not in the mergeable set, allele sequences are retrieved from one of the available callers randomly.

Ensemble calling: For each sample at each locus in a mergeable set, EnsembleTR matches calls from each method to the consensus alleles determined in the previous step. It then determines a consensus genotype by choosing the diploid genotype with the highest score as defined below. In case of ties, it gives priority to callers with the order of HipSTR, GangSTR, ExpansionHunter, and adVNTR. Let S_g be the score for a diploid genotype g . S_g is computed as:

$$S_g = \frac{\sum_{m \in \mathbf{M}} Q_{g,m}}{\sum_{g' \in \mathbf{G}} \sum_{m \in \mathbf{M}} Q_{g',m}} * \max_{m \in \mathbf{M}} Q_{g,m} \quad (2)$$

Where $\mathbf{M}$ is the set of methods considered, $\mathbf{G}$ is the set of possible diploid genotypes (pairs of consensus alleles), and $Q_{g,m}$ is the quality score for method m for genotype g . If the genotype returned by method m is not equal to g , then $Q_{g,m}$ is set to 0. Otherwise, $Q_{g,m}$ is set to a quality score specific to each method. For HipSTR, AdVNTR, and GangSTR the quality score is obtained from the Q score of the original VCF file, which ranges from 0 to 1. For ExpansionHunter, we defined a quality score based on the copy numbers and confidence intervals of each allele in the called genotype. Each allele's score is calculated by the formula:

$$\frac{1}{\exp(4 * \frac{CI}{CN})} \quad (3)$$

where CN is the copy number and CI is the length of the confidence interval. Then a final score for the ExpansionHunter genotype is calculated as a weighted average between two alleles' scores. A coefficient of 0.8 is used for the lower score and 0.2 for the higher one to give prominence to the low-quality genotype. We tried coefficients other than (0.8, 0.2) in score definition and compared their performance in terms of alignment with the Mendelian Inheritance rates in ExpansionHunter calls. While all settings of score coefficients are effective in capturing the true quality of calls according to MI error rates, differences in their performance are negligible. EnsembleTR outputs a new VCF file with the final genotypes with the highest score, along with the score S_g for each call.

1.4.6 Experimental validation of TR genotypes

For each candidate TR, we obtained primers to amplify the TR and surrounding region (Supplementary Data 3). A universal M13(-21) sequence (5'-TGTAACGACGGCCAGT-3') was appended to each forward primer. We then amplified each TR using a three-primer reaction previously described⁵¹ consisting of the forward primer with the M13(-21) sequence, the reverse primer, and a third primer consisting of the M13(-21) sequence labeled with a fluorophore.

The forward (with M13(-21) sequence) and reverse primers for each TR were purchased through IDT. The labeled M13 primers were obtained through ThermoFisher (#450007) with fluorescent labels added to the 5' ends (either FAM, VIC, NED, or PET). TRs were amplified using the forward and reverse primers plus an M13 primer with one of the four fluorophores with GoTaq polymerase (Promega #PRM7123) using PCR program: 94°C for 5 minutes, followed by 30 cycles of 94°C for 30 seconds, 58°C for 45 seconds, 72°C for 45 seconds, followed by 8 cycles of 94°C for 30 seconds, 53°C for 45 seconds, 72°C for 45 seconds, followed by 72°C for 30 minutes.

For several loci which were difficult to amplify using the above conditions, we used separate PCR conditions. Full experimental details for these loci are provided in Supplementary Methods. For HTT, C9orf72, and FMR1 we used available kits from Asuragen for genotyping (HTT: AmpliX® PCR/CE HTT Kit⁵², C9orf72: AmpliX® PCR/CE C9orf72 Kit⁵³, FMR1: AmpliX® PCR/CE FMR1 Kit⁵⁴).

Fragment analysis of PCR products was performed on a ThermoFisher SeqStudio instrument using the GSLIZ1200 ladder, G5 (DS-33) dye set, and long fragment analysis options. Raw PCR product sizes are given in Supplementary Data 3. Product sizes were converted to allele lengths using a binning process described in the Supplementary Methods.

Asuragen results are reported in Supplementary Data 5. To make Asuragen results (reported in total repeat copy number) comparable to WGS calls (reported as the number of repeat units relative to hg38), we applied an offset of -3 and -19 for C9orf72 and HTT genotypes, respectively. While AmplideX® PCR/CE FMR1 Kit results are also reported for FMR1, we did not include that locus in our validation analysis since our WGS calls include only autosomal loci.

For HTT only, we performed a sequence-specific analysis to compare WGS and experimentally validated genotypes. The repeat region in HTT consists of the sequence (CAG)_nCAACAGCCGCCA(CCG)_n. While EnsembleTR identifies changes in either the CAG or CCG repeat, the AmplideX® PCR/CE HTT Kit calls specifically analyzes only the CAG repeat. Therefore, we extracted the total number of CAG repeats, rather than the entire repeat length, from EnsembleTR before performing comparisons. Notably, the SCA1 locus consists of an imperfect repeat. While EnsembleTR, HipSTR, and the capillary electrophoresis calls measure the total change in repeat length, GangSTR considers only the longest perfect repeat stretch, which likely accounts for the discrepancy with GangSTR calls at this locus.

1.4.7 Published allele frequencies for forensics and disease-associated TRs

Population-specific allele frequencies for the CODIS forensics TRs in EUR, AMR, AFR, and EAS populations were obtained from NIST STRBase (<https://strbase.nist.gov/>, <https://strbase.nist.gov/1036-Revised-Allele-Freqs-PopStats-July-19-2017.xlsx>). SAS allele frequencies were obtained from literature sources^{55,56}. Control allele frequencies for disease-associated TRs were obtained from various sources: (1) HTT: Validated repeat lengths were previously obtained¹³ from Huntington’s Disease patients (dbGaP accession “phs000371.v2.p1 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000371.v2.p1]”). We used non-expanded alleles from table phs002988.v1.p1.c1 to estimate control allele frequencies in

European samples. For other populations, allele frequencies were extracted from Masuda, et al.⁵⁷ (EAS), Baine, et al. (AFR and H3Africa)⁵⁸, Saleem, et al.⁵⁹ (SAS), and Paradisi, et al.⁶⁰ (AMR); (2) DMPK: Allele frequencies were obtained from Ambrose, et al.⁶¹ (EAS), Acton, et al.⁶² (AFR), and Magana, et al.⁶³ (AMR); (3) PPP2R2B: Allele frequencies for Europeans were obtained from Majounie, et al.⁶⁴. Similarity between published allele frequencies and those obtained from EnsembleTR was measured using Jensen-Shannon divergence, implemented in the python scipy library⁶⁵ v1.5.2 to obtain a null value for the divergence. For each locus/population pair, we first calculated the average Jensen-Shannon divergence between its EnsembleTR frequency and the published frequency for all other locus/population pairs. We then calculated the mean over all pairs.

1.4.8 Mendelian inheritance analysis

We analyzed 602 trios available in 1000GP. For each trio, we only assess the Mendelian Inheritance if 1) calls were available for all three samples and 2) at least one of the samples is not homozygous for the reference allele. The score assigned to each trio is the minimum score reported by EnsembleTR among all samples in the trio. In all analyses except the TR expansion analysis, TRs with Mendelian error rates >5% were filtered, leaving 1,443,686 total TR loci.

1.4.9 Characterization of population-specific TR variation

We performed principal component analysis on a matrix $\mathbf{X}_{n,m}$ where n is the number of TR loci and m is the number of samples. Each cell $c_{i,j}$ of $\mathbf{X}$ denotes the sum of allele lengths for a diploid call of j th sample at i th locus. In the case of a no call, $c_{i,j}$ is set to NaN. Due to the large size of $\mathbf{X}$, we used a memory-efficient method, Incremental Principal Component Analysis (IPCA), implemented in the python scikit-learn library⁶⁶ v1.0.2.

1.4.10 Inspecting TRs not matching hg38

To validate the 196 TRs for which the majority of alleles (>95%) differ by more than 2 repeat units from the hg38 reference (Supplementary Data 10), we examined the length of those TRs separately in both the T2T²³ reference and PacBio HiFi reads for sample HG00438 obtained from the Human Pangenome Reference Consortium⁶⁷ (HPRC) (<https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC/>). For the PacBio HiFi dataset, we used the Integrative Genomics Viewer⁶⁸ to manually inspect reads aligning to each TR. To compare to the T2T reference v1.1 (https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/chm13.draft_v1.1.fasta.gz), hg38 coordinates of the 196 TRs were converted to T2T v1.0 first and then converted to v1.1 using the UCSC liftOver⁵⁰ utility with the corresponding chain files at <http://t2t.gi.ucsc.edu/chm13/hub/t2t-chm13-v1.0/hg38Lastz/hg38.t2t-chm13-v1.0.over.chain.gz> and

https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/changes/v1.0_to_v1.1/v1.0_to_v1.1_rdna_merged.chain.

For TRs that failed to convert due to partial deletion, we added additional flanking sequences (up to 1,000bp) to the start and end coordinates and reattempted liftOver, which resulted in successful conversion of 195/196 TRs to T2T v1.1. One TR failed due to deletion in the T2T v1.1 reference. For each TR, We used samtools⁶⁹ v1.5 to extract its sequence with flanking regions from both the T2T v1.1 and hg38 references and compared the two sequences using BLASTN⁷⁰ v2.13.0+. We used a similar method to compare the TR lengths between the T2T v1.1 reference and the major alleles called in EnsembleTR by replacing the hg38 reference with the EnsembleTR major allele at each TR.

1.4.11 Detecting population-specific expansions

For each TR with repeat unit length >1bp, we defined an expanded allele to be any allele with copy number greater than $Q3 + 3 \times IQR$, where $Q3$ is the third quartile, and IQR is the difference between the third and first quartile. We calculated the frequency of expansions in African (1000GP AFR and H3Africa) and Non-African populations (all other 1000GP super-populations). We defined TRs with population-specific expansions as those for which 1) the expansion threshold copy number ($Q3 + 3 \times IQR$) is greater than 10, 2) the frequency of expansions is greater than 0.01 in at least one population and, 3) the expansion frequency in one population is at least 10 times larger than the other population. Gene annotations for repeat expansions were based on Ensembl version 108 available at https://ftp.ensembl.org/pub/release-108/gtf/homo_sapiens/Homo_sapiens.GRCh38.108.gtf.gz.

To support these results, we applied two additional TR genotypers (STRetch²⁴ and ExpansionHunter Denovo¹¹) to identify expansions in the H3Africa cohort. STRetch v.0.4.0 takes as input a reference genome with decoy STR contigs of length 2000bp, a CRAM or BAM file and a bed file with genome locations of TRs. A STRetch STR catalog was generated for GRCh38 and the recommended pipeline for WGS was run for each sample. We ran STRetch twice, once using 143 control samples provided and once without any controls and using all samples in one batch. However, due to the distinct nature of the sequencing data for control samples in STRetch, which were PCR-free unlike the H3Africa samples, we consider the results from the second analysis to be more reliable and therefore use them for downstream analysis. STRetch results for each sample were then merged into a single file and filtered using the criteria $p_{adj} < 0.05$, locuscoverage ≥ 3 and outlier Z-score ≥ 8 . The filtered loci were annotated using the OrganismDbi⁷¹ R package v1.40.0.

ExpansionHunter Denovo v.0.9.0 was used to generate genome-wide STR profiles for each sample with default parameters of `--min-unit-len 2`, `--max-unit-len 20`, `--min-anchor-mapq 50`, and `--max-irr-mapq 40` in order to restrict the search of motif lengths of up to 20 bp. Unlike STRetch, ExpansionHunter Denovo does not require prior knowledge of the location of repeats in the genome. A manifest file was synthesized where each sample was labeled as a case and STR profiles were merged to allow comparisons among samples after read depth normalization. Outlier locus and motif analyses were performed using scripts available in the ExpansionHunter Denovo package and the output was ranked using Z scores. Annotation of locus-based analysis results was done using ANNOVAR⁷². To find the overlap between our set of expansions and expansions found by ExpansionHunter Denovo and STRetch, for each repeat expansion r in our list, we checked if there is any expanded repeat reported by either ExpansionHunter Denovo or STRetch that meets two conditions: 1) it is located in the surrounding ± 1000 bp window of the r , 2) the repeat unit sequence of both expansions are the same.

To validate the candidate repeat expansion in an intron of *CA10*, we designed primers to amplify the TR and surrounding region using the three-primer reaction described above. consisting of the forward primer with the M13(-21) sequence (5'-TGTAACGACGGCCAGTTGGCTCCAAGTAGCACATCTT-3'), the reverse primer (5'-TGCAACTAGCGGTGACCTTA-3'), and a third primer consisting of the M13(-21) sequence labeled with a fluorophore. Primers were purchased through IDT. The labeled M13 primers were obtained through ThermoFisher (#450007) with FAM fluorescent labels added to the 5' ends. The locus was amplified with GoTaq polymerase (Promega #PRM7123) using the PCR program: 94°C for 5 minutes, followed by 30 cycles of 94°C for 30 seconds, 59°C for 45 seconds, 72°C for 45 seconds, followed by 8 cycles of 94°C for 30 seconds, 53°C for 45 seconds, 72°C for 45 seconds,

either side of the TRs. The sequences before and after each STR were separate examples but part of the same variable or stable set. HOMER was run to find motifs 4-12 bases long without GC correction with the command `findMotifs.pl <targetSequences.fa> fasta <output directory> -fasta <background.fa> -len 4,5,6,7,8,9,10,11,12 -noweight`.

Our neural network model consists of a 1-D CNN with inception blocks⁷³ implemented with Pytorch. Instead of applying convolutional kernels of a single width, inception blocks in parallel apply filters of multiple widths in addition to a pooling and single-width convolution. Our best performing model used six inception blocks with kernel sizes 5, 9, and 15 followed by global average pooling and a single linear layer. The data was split 70:15:15 into train, validation, and test splits and reverse complements were added to the same split as the forward strand sample.

To blind the model from STR length and focus on nearby sequence information, the model input is the 58 bp flanking the STR and the 6 directly adjacent bases that make up the STR from both flanks. The model input is then these two flank sequences concatenated together into a single sequence. The input was represented using one-hot encoding, so a zero matrix was used as a baseline for generating attribution scores with Integrated Gradients⁷⁴. We found that using a global average pooling layer instead of a global max pooling layer led to more informative attribution scores.

1.4.14 eQTL analysis in the GEUVADIS cohort

We obtained gene-level reads per kilobase of transcript per million mapped reads (RPKM) values for 452 unrelated individuals generated from lymphoblastoid cell lines by the GEUVADIS project³¹ (<https://www.internationalgenome.org/data-portal/data-collection/geuvadis>). Duplicated samples were removed by arbitrarily keeping the first dataset for each. Genes with RPKM above 0.1 in more than 10 samples were kept for downstream analysis. Expression values for remaining

genes were quantile-normalized on sample level followed by quantile-normalization to a standard normal distribution separately for each gene. Genes overlapping segmental duplications were removed, and analysis was restricted to protein-coding genes based on GENCODE v12 annotation. Gene coordinates were adjusted from hg19 to hg38 using the liftOver available from the UCSC Genome Browser⁵⁰. After filtering, 12,607 genes remained for analysis.

To control for population structure, we obtained publicly available genotype data on 2,318 unrelated individuals from the 1000GP genotyped with Omni 2.5 SNP genotyping arrays (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/). We removed all indels, multiallelic SNPs, and SNPs with a minor allele frequency of less than 5%. We then used plink⁷⁵ v.1.90b3.44 to subset these remaining SNPs to a set of SNPs in approximate linkage equilibrium with the command `--indep 50 5 2`. We excluded any remaining SNPs with a missingness rate of 5% or greater. We lastly ran principal component analysis using smartpca^{76,77} v.1.3050 with default parameters.

Association tests were performed separately on the African and European populations. For each TR, we tested for association with each gene within 100kb. We performed a linear regression for each test between the TR dosage (the sum of allele lengths relative to the hg38 reference genome) and normalized gene expression. We included the top 10 genotype principal components as computed above, 44 PEER factors⁷⁸, and sex as covariates. The number of PEER factors was chosen based on the recommended 1/10 of the sample size. PEER factors were calculated using PEER v1.0 based on the normalized gene expression data of all 452 samples. In cases where fewer than 50 samples had non-missing TR genotypes, the TR was removed from analysis in that population.

To identify individual significant eTRs, we obtained adjusted p-values using the Benjamini-Hochberg approach for controlling the false discovery rate⁷⁹ applied to p-values for all TR-gene pairs separately in Europeans and Africans. To identify gene-level significant eTRs, we followed the steps in our previous eTR analyses^{16,32}. For each gene, we determined the TR association with the strongest P-value. This P-value was adjusted using a Bonferroni correction for the number of TRs tested per gene to give a P-value for observing a single eTR association for each gene. We then used the list of adjusted P-values (one per gene) as input to the Benjamini-Hochberg method to obtain a q-value for the best eTR for each gene.

eTR summary statistics based on the GTEx dataset used for effect size comparisons were obtained from Supplementary Dataset 2 of Fotsing et al¹⁶. TR coordinates were lifted over to hg38 for comparison. Because coordinates can vary slightly between callsets, we identified TRs as overlapping if their coordinates were within 20bp.

1.4.15 Phasing and imputation

Phased SNPs for 1000GP samples were downloaded from the 1000GP FTP server (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/workin_g/20201028_3202_phased). We used Beagle v4.0⁸⁰ to phase each TR separately. To produce a high-quality TR callset for phasing, we performed additional filtering to remove all calls with quality score below 0.9, TRs with call rate below 0.8, and TRs with average Mendelian Error rate > 5%. We manually included three pathogenic loci (HTT, SCA1, and SCA3) that did not meet the filtering criteria.

Our pipeline is based on our previously published framework³⁵ and takes the unphased TR and surrounding phased SNPs from a 50kb window centered at the TR as input (--gt). We set the -usephase parameter to True to allow Beagle to use the phase information of provided phased SNPs

to phase the target TR. This step outputs a phased VCF file containing both SNPs and the target TR. We apply a custom script to ensure the phase order matches the original SNP input. We then extract and concatenate phased TR genotypes from each locus and combine them with the original phased SNPs into a single phased VCF file. In order to comply with Beagle's requirement of a maximum of 126 alleles per record, manual modifications were made to the VCF entries for HTT and SCA1. The alteration involved unifying alleles with the same length but different sequences, and subsequently updating the corresponding genotypes to meet the specified requirement.

We then used Beagle v5.4⁸¹ to perform a Leave-One-Out analysis to assess concordance. This analysis was restricted to chromosome 21 due to the high computational burden. For each sample S , phased SNPs+TRs for all samples except S are given to Beagle as --ref and phased SNPs for sample S are given to Beagle as reference panel (--gt). Beagle will use these inputs to impute the missing TRs for S . After performing imputation for $n = 100$ randomly chosen samples from each population (excluding trio samples), concordance for each locus is computed as follows: For each sample $S_i, i \in \{1..n\}$, let $\mathbf{x}_{ij}$ be the EnsembleTR genotype and $\mathbf{y}_{ij}$ be the imputed TR genotype for sample S_i at the j th locus. Each genotype for a diploid sample contains two alleles, therefore we will define $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2})$ and $\mathbf{y}_{ij} = (y_{ij1}, y_{ij2})$. Then concordance c_{ij} for S_i at the j th locus is computed as: 1 if both genotypes match: $\text{sorted}(x_{ij1}, x_{ij2}) == \text{sorted}(y_{ij1}, y_{ij2})$; 0 if neither imputed allele matched an EnsembleTR allele; else 0.5 if one but not both imputed alleles matched the EnsembleTR alleles. Total concordance for the j th locus c_j is computed by averaging over concordance values for each sample $c_j = \frac{1}{n} \sum_i c_{ij}$.

To compute LD between each TR and a nearby SNP, we calculated the squared Pearson correlation coefficient between the SNP and TR genotype vectors, where each vector has $2n$

elements, n is the number of samples and each sample has two alleles. We used the phased and imputed TR genotypes for this analysis and we selected the SNP with the highest r^2 as the candidate tag SNP for each target TR.

To evaluate imputation of trait-associated TRs (Supplementary Data 17), we obtained coordinates of 95 blood-trait associated loci from Supplementary Table 4 of Margoliash, et al³⁶. TR coordinates were converted from hg19 to hg38 using the UCSC liftOver tool⁵⁰.

1.5 Data Availability

TR genotypes, the phased TR-SNP reference panel, and population-specific summary statistics generated in this study are available from the EnsembleTR Github webpage [<https://github.com/gymrek-lab/EnsembleTR>]. Summary statistics are also made available in browsable format at WebSTR [webstr.ucsd.edu]. Summary statistics of all tested TR-gene pairs in African and European samples are deposited in the Figshare database [https://figshare.com/articles/dataset/1000GenomesH3Africa_SuppData16_zip/24164367]. The WGS datasets for the 1000GP samples used in this study are available from the European Nucleotide Archive under accessions PRJEB31736 [<https://www.ebi.ac.uk/ena/browser/view/PRJEB31736>] (unrelated samples) and PRJEB36890 [<https://www.ebi.ac.uk/ena/browser/view/PRJEB36890>] (related samples). The H3Africa WGS datasets used in this study are available in the European Genome-Phenome Archive under accession EGAS00001002976 [<https://ega-archive.org/studies/EGAS00001002976/>]. Geuvadis datasets used in this study are available from the 1000GP website [<https://www.internationalgenome.org/data-portal/data-collection/geuvadis>]. Analyses are based on the GRCh38 reference genome [https://storage.googleapis.com/genomics-public-data/resources/broad/hg38/v0/Homo_sapiens_assembly38.fasta].

1.6 Code Availability

EnsembleTR is available on GitHub: <https://github.com/gymrek-lab/EnsembleTR>⁸².

1.7 Acknowledgments

Chapter 1, in full, is a reprint of the material “A deep population reference panel of tandem repeat variation” by Helyaneh Ziaei Jam, Yang Li, Ross DeVito, Nima Mousavi, Nichole Ma, Ibra Lujumba, Yagoub Adam, Mikhail Maksimov, Bonnie Huang, Egor Dolzhenko, Yunjiang Qiu, Fredrick Elishama Kakembo, Habi Joseph, Blessing Onyido, Jumoke Adeyemi, Mehrdad Bakhtiari, Jonghun Park, Sara Javadzadeh, Daudi Jjingo, Ezekiel Adebisi, Vineet Bafna & Melissa Gymrek as published in Nature Communications, 2023. This work was partially funded by NIH/NHGRI grants 1R01HG010149 (M.G. and V.B.), R01HG010885 (M.G.), 1RM1HG011558 (M.G.), 5U24HG006941 (Y.A. and E.A.), 5U2RTW010679 (E.A.), 1U2RTW010672-01 for (D.J, I.L, F.K and H.J) and 1U2CEB032224-01 (D.J). B.O. and J.A. were partially funded by the World Bank ACE Impact funding for CApiC-ACE. V.B., J.P., and S.J. were supported in part by grants GM114362, HG010149, and RM1HG011558. The authors acknowledge Brian Haynes and Sarah Statt for providing Asuragen kits and results and Rahel Wachs for help with illustrations. The authors also acknowledge support from the H3Africa genome analysis working group, as well as TrypanoGEN, CAfGEN and Baylor for data from the H3Africa consortium.

1.8 Appendix

Experimental validation of TR genotypes

The following loci were amplified using separate PCR conditions than those described in **Methods**:

- SCA1(ATXN1): 95°C for 5 mins, 28 cycles of 95°C for 1 min, 61°C for 1 min, 72°C for 1.5 min; 8 cycles of 95°C for 1 min, 53°C for 1 min, 72° C for 1.5 mins, and a final extension at 72°C for 10 min.

- ATN1: 95°C for 15 min, 27 cycles at 94°C for 20 s, 65° C for 30 s, 72°C for 1.5 min; 8 cycles of 94°C for 20 s, 53°C for 30 s, 72°C for 1.5 mins and a final extension at 72°C for 10 min.

Several additional loci were initially analyzed by capillary electrophoresis but were filtered from downstream analysis since we determined they produced unreliable genotypes:

- Known pathogenic repeats at *HOXD13*
(F:tgtaaaacgacggccagtAAGGAGAGGAGGGAGGAGG,
R:GACATACGGCAGCTGTAGTAGC and *PAPBN1*
(F:tgtaaaacgacggccagtCCAGTGCCCCGCCTTAGA, R:
ACAAGATGGCGCCGCGCCCGCCCGGC) were excluded since all samples were called as homozygous reference by both EnsembleTR and capillary electrophoresis.
- A known pathogenic repeat in FXN was genotyped but excluded from comparisons since it was not genotyped by EnsembleTR.
- A known pathogenic repeat for SCA2 (F:6FAM-GGGCCCCCTCACCATGTCG,
R:CGGGCTTGCGGACATTGG) was excluded since many capillary calls had evidence of three alleles in a single sample, and we could not determine the correct diploid call.
- A known pathogenic repeat in *PHOX2B*
(F:tgtaaaacgacggccagtCCAGGTCCCAATCCCAAC, R:GAGCCCAGCCTTGTCCAG)
was excluded since all capillary calls were heterozygous for the same alleles, which is unlikely to be correct.
- We additionally excluded known pathogenic loci on chrX, since our calls focus on autosomal TRs. Results for the FMR1 repeat, which is on chrX, are reported from Asuragen but are not included in comparison.

Fragment analysis was performed in separate batches for Platinum Genomes (17 samples; “PG”) and the remaining 31 samples (“Coriell”). Product sizes were converted to repeat differences from reference to be comparable to our WGS genotypes. In some cases, product sizes were offset by a constant value from the expected values. We binned product sizes to allele lengths, with binsets manually set for each locus in each batch. For each locus, we set the minimum (minval) and maximum (maxval) product size, the repeat unit length (period), and allele assignment of the minimum product size (start). We then determined the allele for a given product size (psize) using the following rule:

1. Set allele=start and i=minval
2. While i<maxval: if psize>=i and psize<i+period return allele. Else allele+=1, i+=period
3. If no matching allele was found, return NA

The bin definitions for each locus/batch are provided in **Supplementary Table 7**.

Supplementary Data 1-17 can be found here:

https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-023-42278-3/MediaObjects/41467_2023_42278_MOESM4_ESM.xlsx

Supplementary Data 18 can be found here:

https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-023-42278-3/MediaObjects/41467_2023_42278_MOESM5_ESM.zip

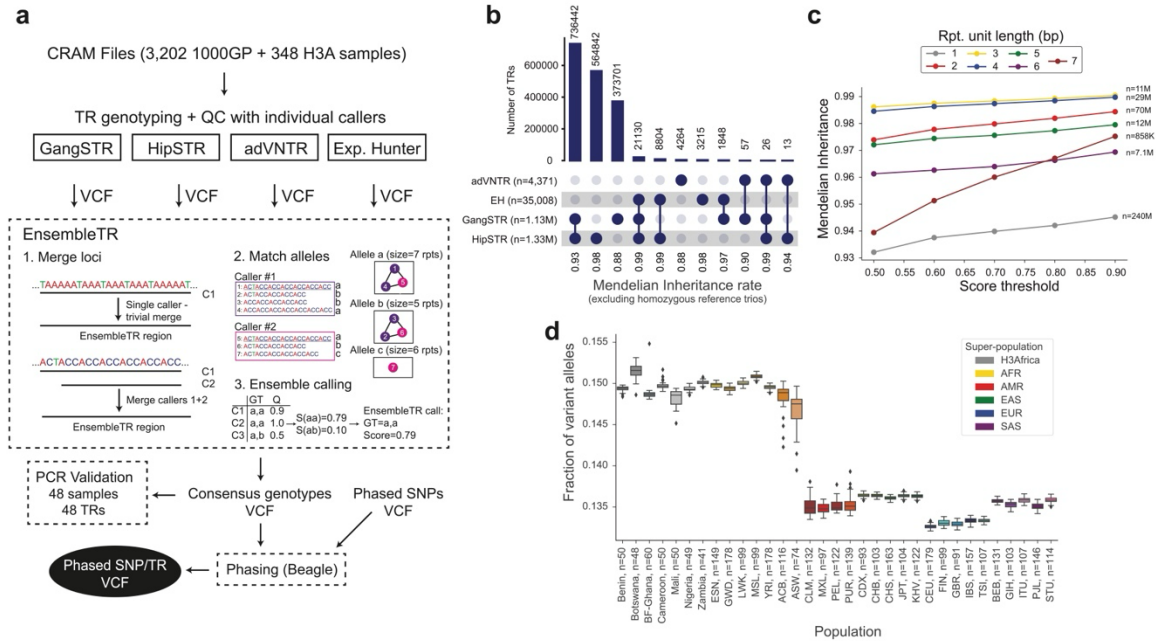


Figure 1.1: A deep catalog of TR variation across human populations

a. Overview of EnsembleTR workflow. Aligned reads are input to four TR genotyping tools (GangSTR, HipSTR, adVNTR, and ExpansionHunter). Filtered VCFs are input to EnsembleTR. EnsembleTR first identifies sets of mergeable loci (step 1) and identifies sets of compatible alleles between callers (step 2). Finally, it scores each possible diploid genotype (step 3) and outputs the best genotype and its score. The resulting VCF file is used to generate a phased SNP+TR reference haplotype panel. **b. Overlap of TRs called by each method.** Annotations below the bars indicate the combination of methods a TR was called in. Numbers next to each method indicate the number of unique TRs in each category. Numbers below the plot indicate the Mendelian Inheritance (MI) rate across all calls in each category. Categories with fewer than 10 total TRs were excluded. **c. Mendelian Inheritance as a function of EnsembleTR quality score.** The x-axis gives the EnsembleTR quality score threshold used, and the y-axis gives the percent of genotyped trios which follow MI. Line colors denote repeat unit lengths. Each trio/TR pair was only included in each category if all calls in the trio passed the score threshold. Trio/TR pairs for which all samples were homozygous for the reference allele were excluded from analysis. **d. Distribution of the fraction of non-reference alleles in individuals by population.** Boxplots summarize the distribution of the fraction of variant alleles in each sample. Horizontal lines show median values, boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to $Q1 - 1.5 \times IQR$ (bottom) and $Q3 + 1.5 \times IQR$ (top), where IQR gives the interquartile range ($Q3 - Q1$). Homopolymer TRs are excluded. Box colors denote superpopulations. Gray denotes H3Africa. Other colors denote 1000 Genomes superpopulations.

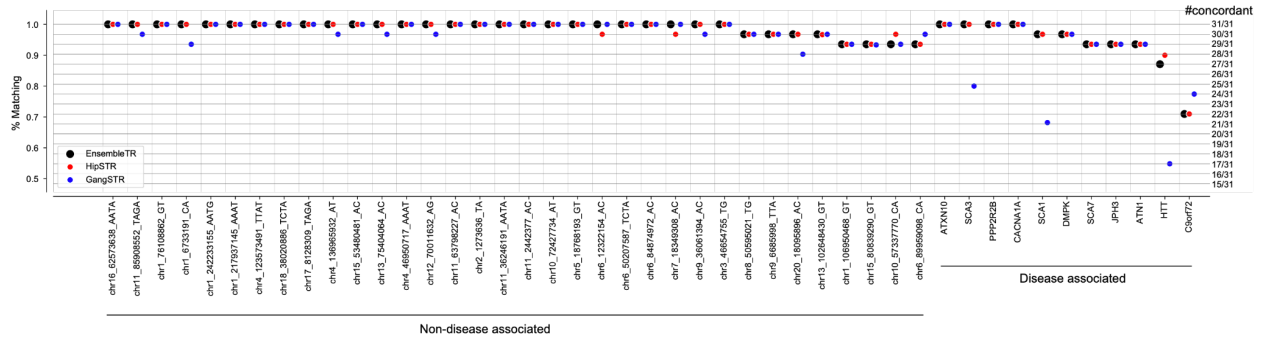


Figure 1.2: Per-locus comparison between WGS-based TR genotypes and capillary electrophoresis results

Each point gives the percent of genotypes that were concordant (y-axis) between capillary electrophoresis and WGS (black=EnsembleTR; red=HipSTR; blue=GangSTR) for each TR (x-axis). TRs are arranged by category (left=non-disease associated TRs; right=disease-associated TRs). Within each category, TRs are sorted by EnsembleTR concordance. Horizontal lines denote the % matching corresponding to the different possible numbers of concordant genotypes (out of a maximum of 31 samples). Full genotype comparison results are provided in **Supplementary Table 6**. Note, for HTT, GangSTR does not distinguish between the CAG_n and CCG_n repeats, whereas the Asuragen test used for PCR counts only CAG copies. Similarly, SCA1 consists of an imperfect repeat, and GangSTR only considers the length of perfect repeat stretches. This likely accounts for lower concordance of GangSTR compared to EnsembleTR and HipSTR at those loci.

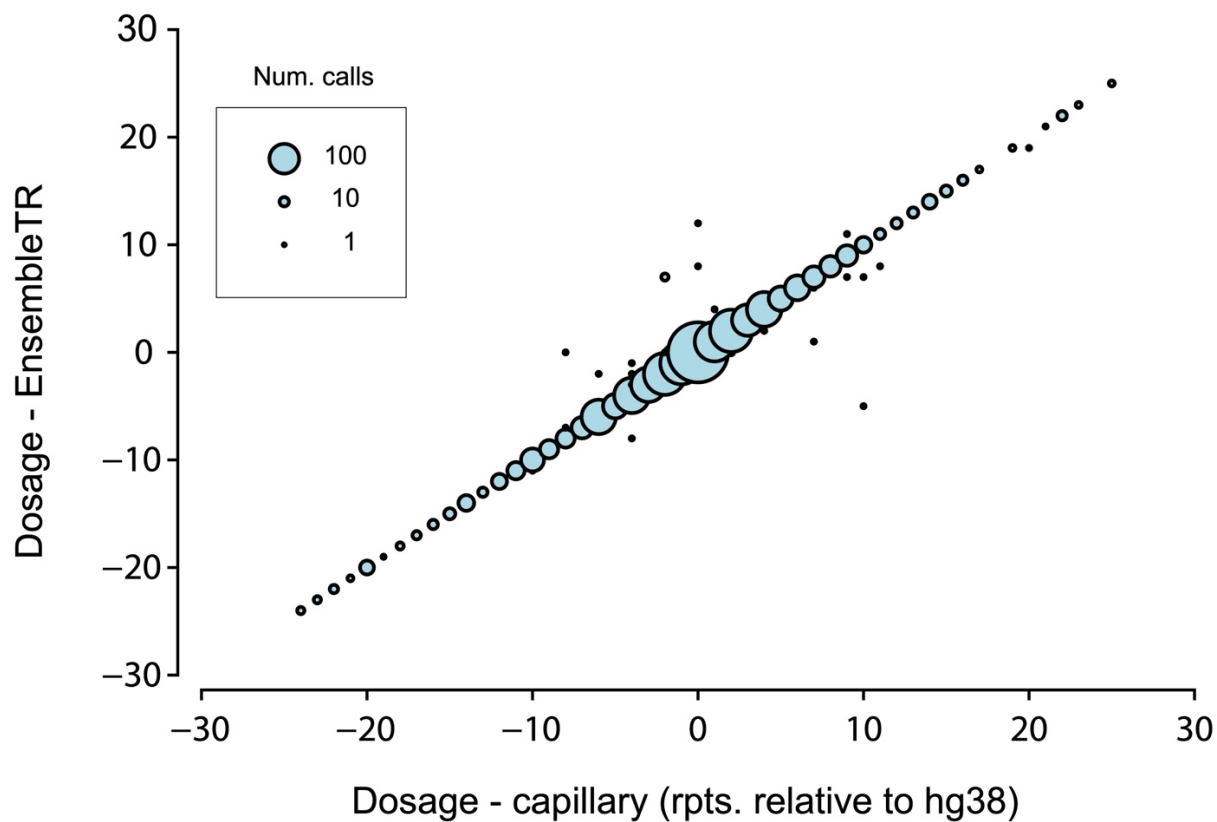


Figure 1.3: Comparison of genotypes called by capillary electrophoresis vs. EnsembleTR

The length of each allele relative to the hg38 reference was computed and lengths of the two alleles were summed for each call. The x-axis gives the capillary result, and the y-axis gives the result called by EnsembleTR. Bubble size scales with the number of genotype calls at each coordinate.

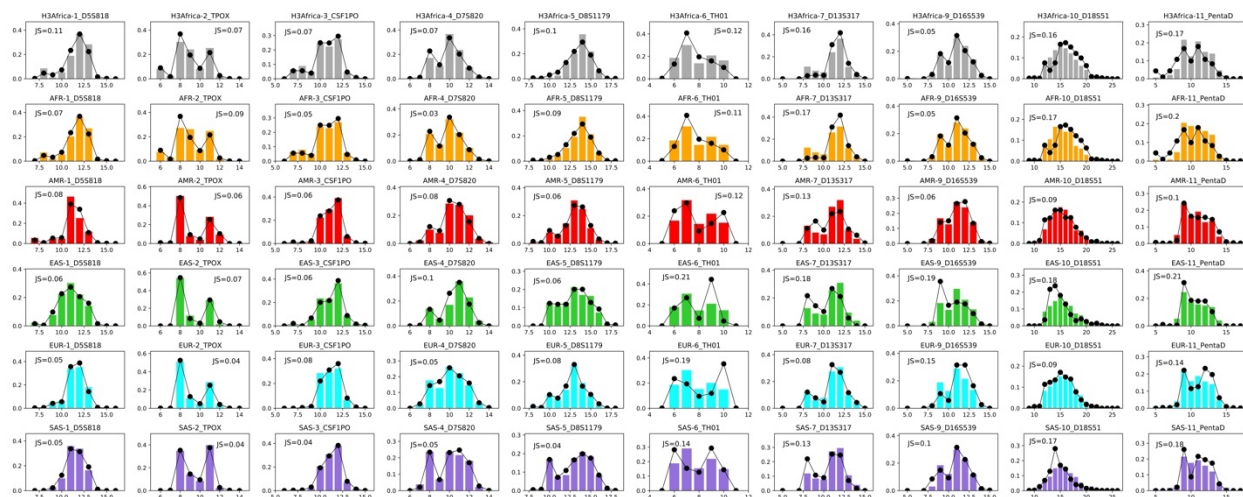


Figure 1.4: Comparison of population-specific allele length distributions called by EnsembleTR vs. published frequencies at the CODIS forensic loci

Rows represent 1000GP superpopulations and H3Africa (gray=H3Africa, blue=EUR, green=EAS, orange=AFR, red=AMR, purple=SAS). Each column represents a different CODIS locus. Colored bars give the frequency (y-axis) of each allele length (x-axis) observed in EnsembleTR results in terms of the number of repeat units. EnsembleTR alleles at the TH01 locus, a tetranucleotide repeat with a common 3bp interruption, were rounded down to the nearest repeat unit. All EnsembleTR alleles at D5S818 were adjusted down by one repeat unit since the EnsembleTR call contains an extra imperfect repeat unit not counted in published frequency data. Black dots give the expected frequency of each length based on previous literature (see **Methods**). For each locus/population pair, the Jensen-Shannon divergence between expected frequencies and EnsembleTR frequencies is annotated at the top of the plot.

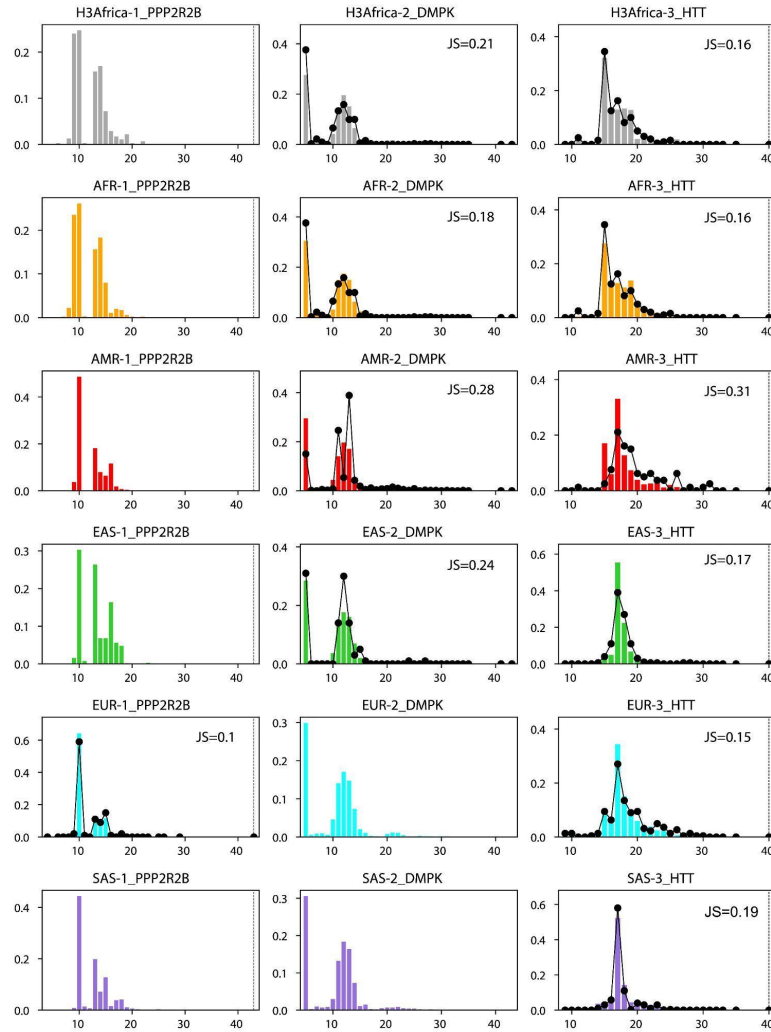


Figure 1.5: Comparison of population-specific allele length distributions called by EnsembleTR vs. published frequencies at the disease-associated loci

Rows represent 1000GP superpopulations and H3Africa (gray=H3Africa, blue=EUR, green=EAS, orange=AFR, red=AMR, purple=SAS). Each column represents a different locus implicated in a TR expansion disorder. Colored bars give the frequency (y-axis) of each allele length (x-axis) observed in EnsembleTR results in terms of the number of repeat units. Black dots give the expected frequency of each length based on previous literature (see **Methods**). For each locus/population pair, the Jensen-Shannon divergence between expected frequencies and EnsembleTR frequencies is annotated at the top of the plot. For HTT, we counted only the number of CAG repeats in alleles returned by EnsembleTR to be consistent with counts obtained in the literature, which do not count the adjacent CCG repeat.

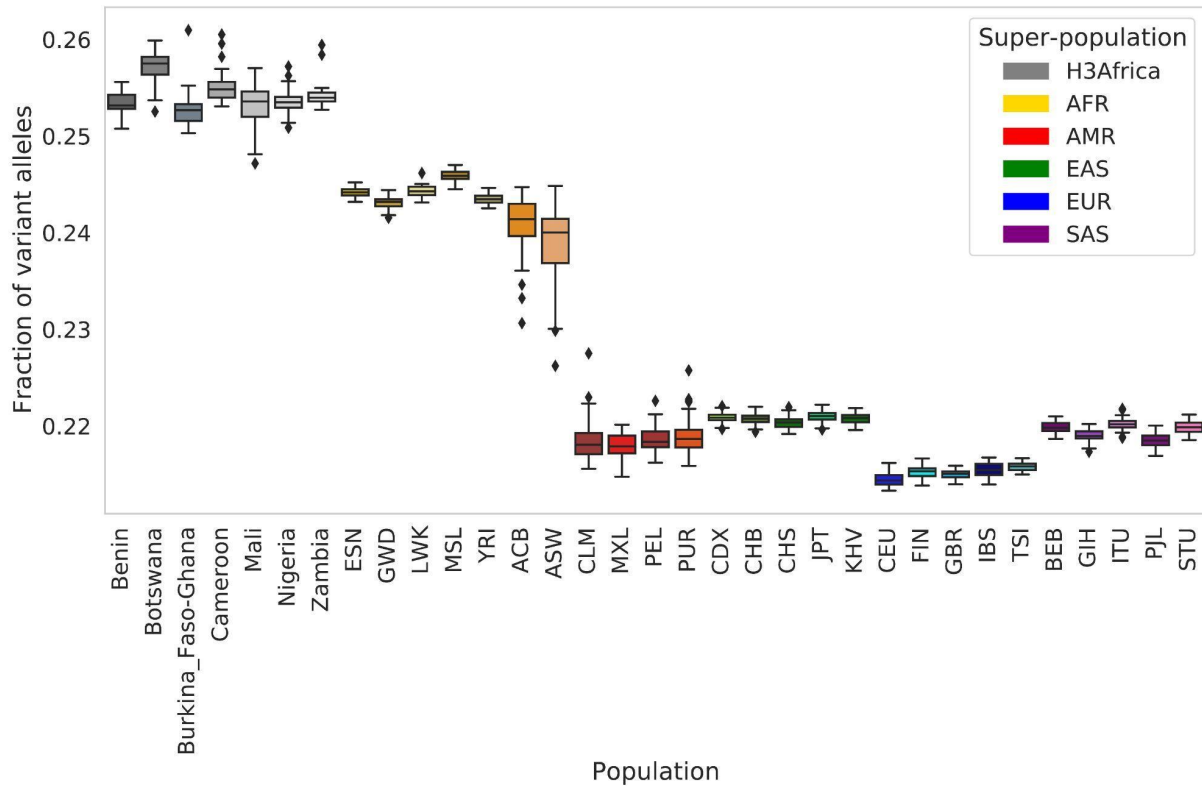


Figure 1.6: Distribution of the fraction of non-reference alleles in individuals by population

Boxplots summarize the distribution of the fraction of variant alleles in each sample. Horizontal lines show median values, boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to $Q1 - 1.5 \times IQR$ (bottom) and $Q3 + 1.5 \times IQR$ (top), where IQR gives the interquartile range ($Q3 - Q1$). Data is the same as in **Fig. 1.1d** except homopolymer TRs are included.

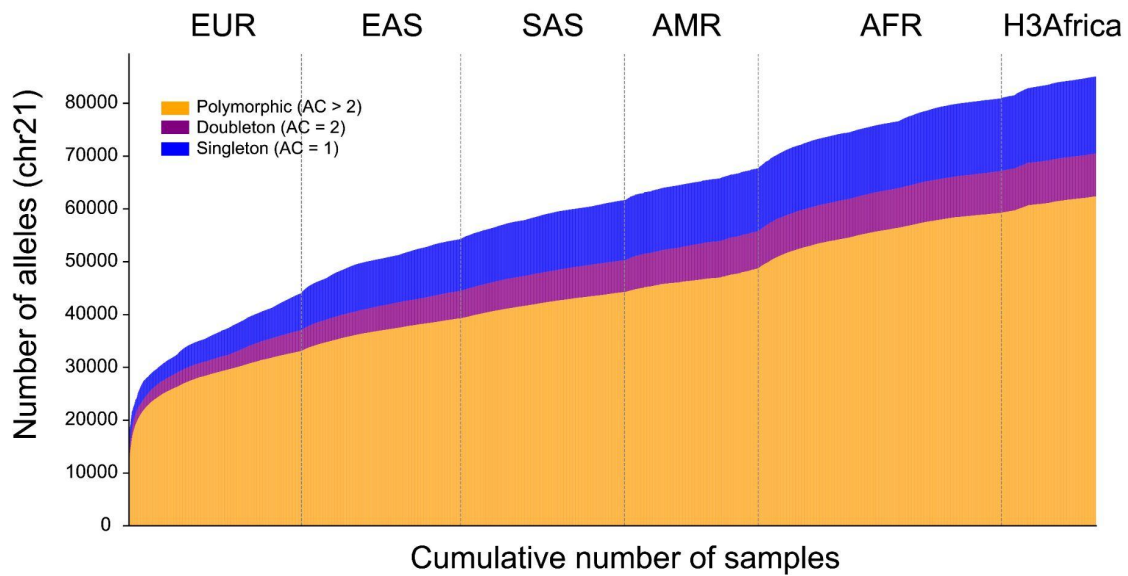


Figure 1.7: Cumulative number of TR alleles with each additional sample

The y-axis gives the number of unique TR alleles discovered with each additional sample. Samples on the x-axis are sorted by superpopulation. Orange=polymorphic alleles (allele count > 2), purple=doubleton (observed twice), blue=singleton (observed once). Data is based on chromosome 21.

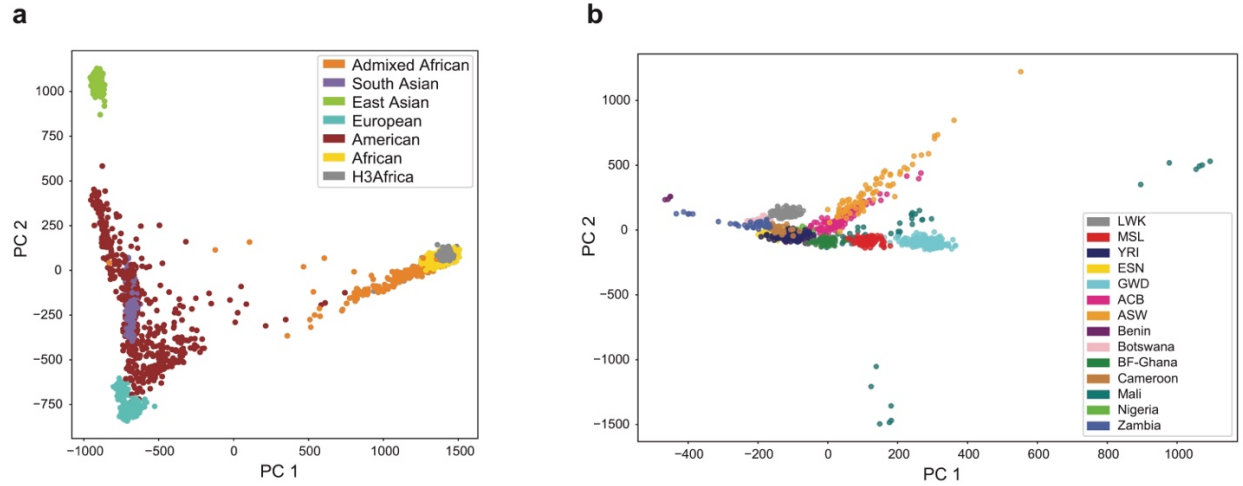


Figure 1.8: TR variants capture known population structure. a. Principal components analysis of all samples

PCA was performed using TR genotypes on all samples (**Methods**). The projection of each sample along PC1 and PC2 are shown. **b. Principal components analysis of African populations.** A separate PCA was performed on only the H3Africa and 1000GP AFR samples. Samples of similar origin cluster despite being sequenced and genotyped separately (e.g. YRI [Yorubans from Nigeria] from 1000G and Nigeria from H3Africa are clustered together). For **a-b**, each dot represents a single individual. Dots are colored based on the population or superpopulation of origin.

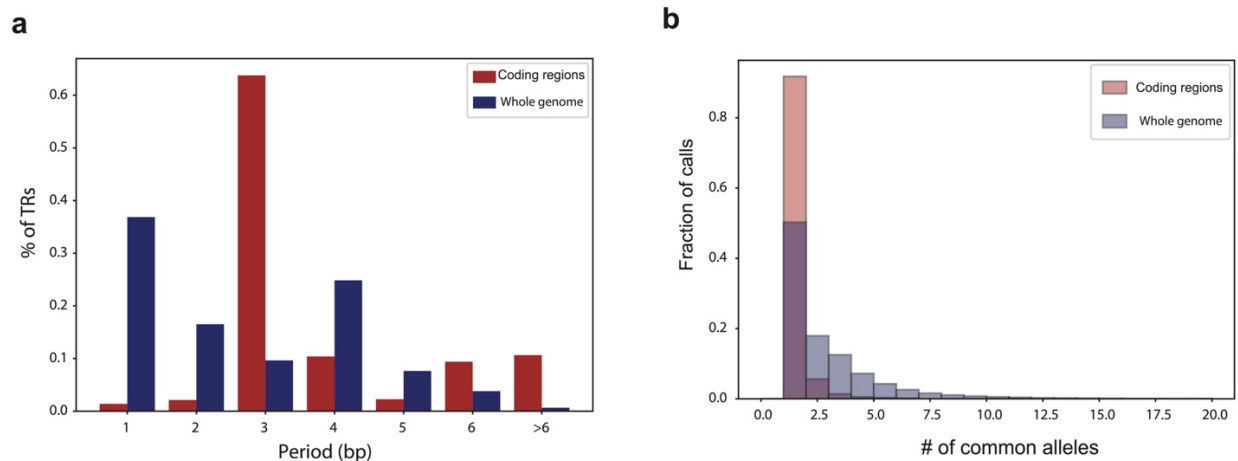


Figure 1.9: Comparison of TR patterns in coding regions vs. genome-wide. a. Comparison of TR repeat unit length distributions

While homopolymer and dinucleotide repeats are most prevalent genome-wide, trinucleotide TRs are over-represented in coding regions. **b. Comparison of the distribution of the number of common alleles at each locus.** Common alleles are defined as having a frequency > 1%. As expected, coding TRs are less polymorphic and have fewer common alleles per TR compared to genome-wide TRs. In both panels, red bars denote TRs in coding regions and blue bars denote genome-wide TRs.

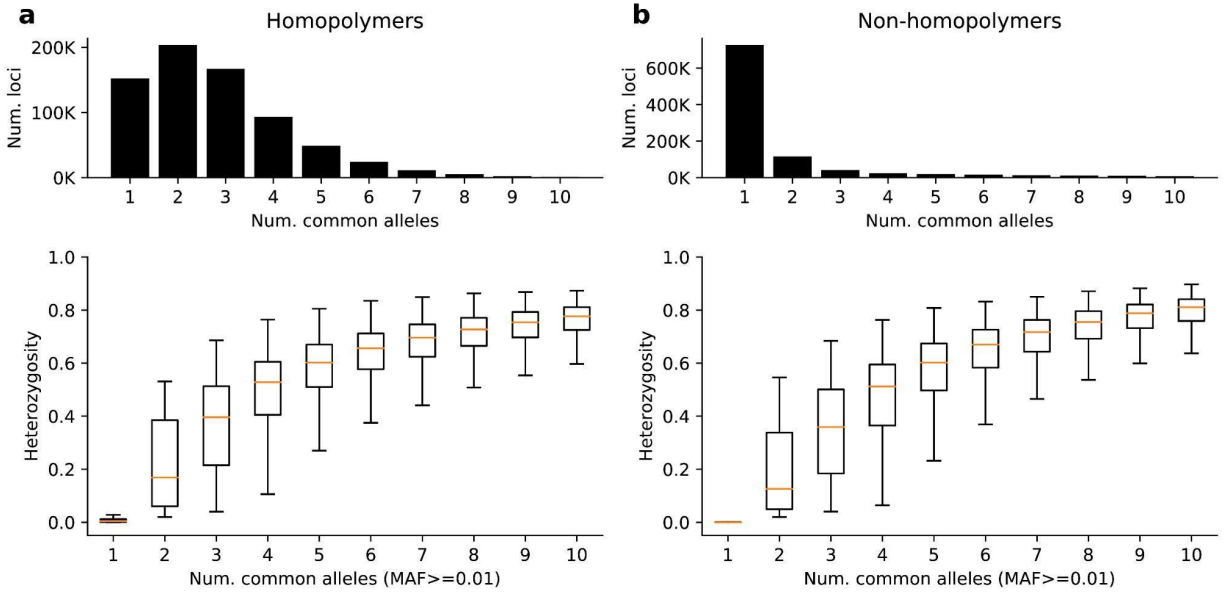


Figure 1.10: TRs exhibit a range of polymorphism levels across loci

a. shows data for only homopolymer TRs, and **b.** shows data for non-homopolymer TRs. Both panels show heterozygosity as a function of the number of common alleles. The y-axis gives the distribution of heterozygosity values for TRs with the number of common alleles specified in the x-axis. Boxplots summarize the distribution of heterozygosity values. Horizontal lines show median values, boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to $Q1 - 1.5 \cdot IQR$ (bottom) and $Q3 + 1.5 \cdot IQR$ (top), where IQR gives the interquartile range ($Q3 - Q1$). In **a-b**, top panels show the number of TRs with each number of common alleles.

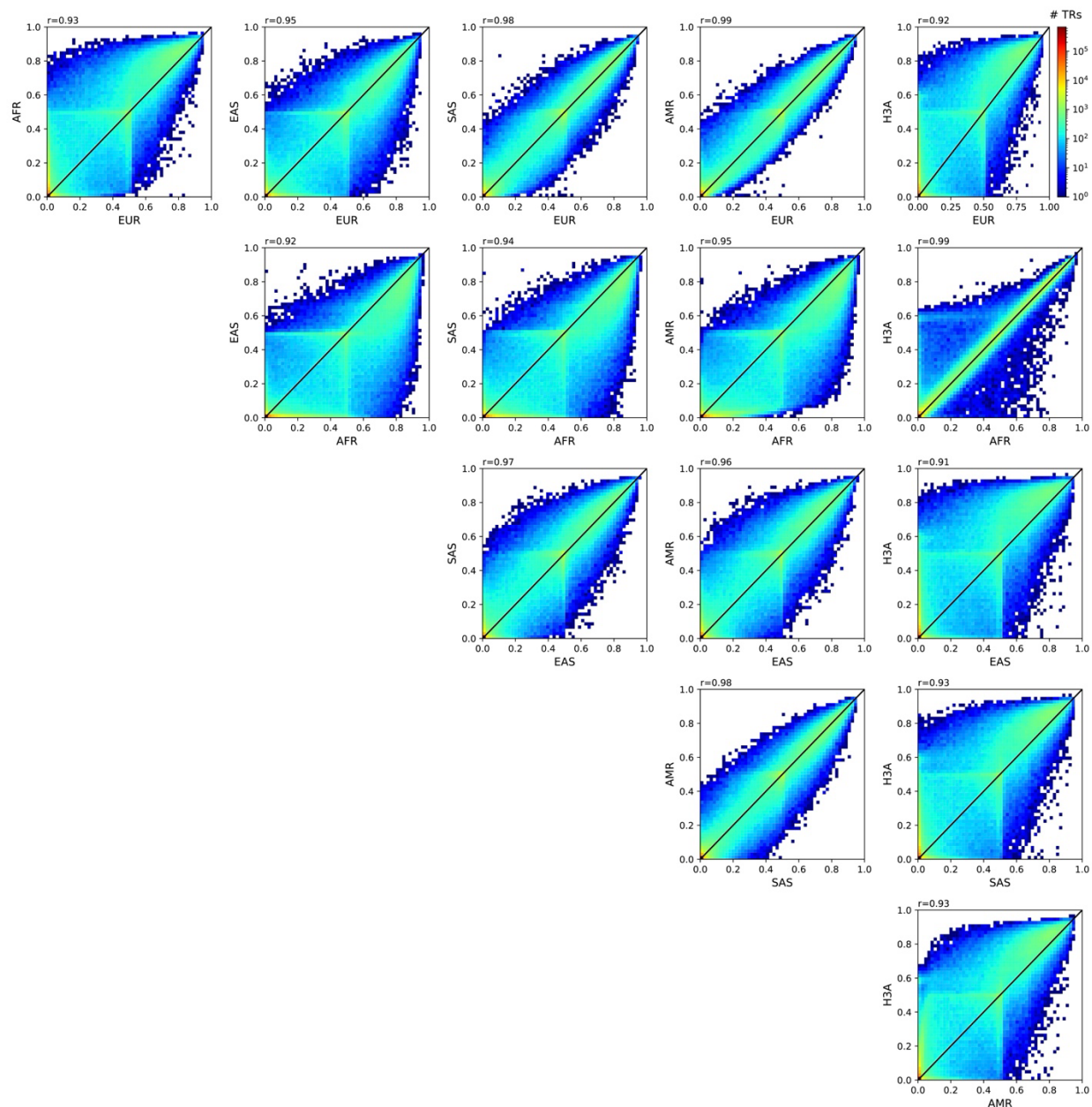


Figure 1.11: Comparison of TR heterozygosity values across populations

Each scatterplot compares the heterozygosity of each TR across two superpopulations. The black line denotes the diagonal. Heatmap shading indicates the number of TRs falling into each bin. The Pearson correlation comparing heterozygosity values in each pair of superpopulations is annotated above each panel. Data is shown for non-homopolymer TRs.

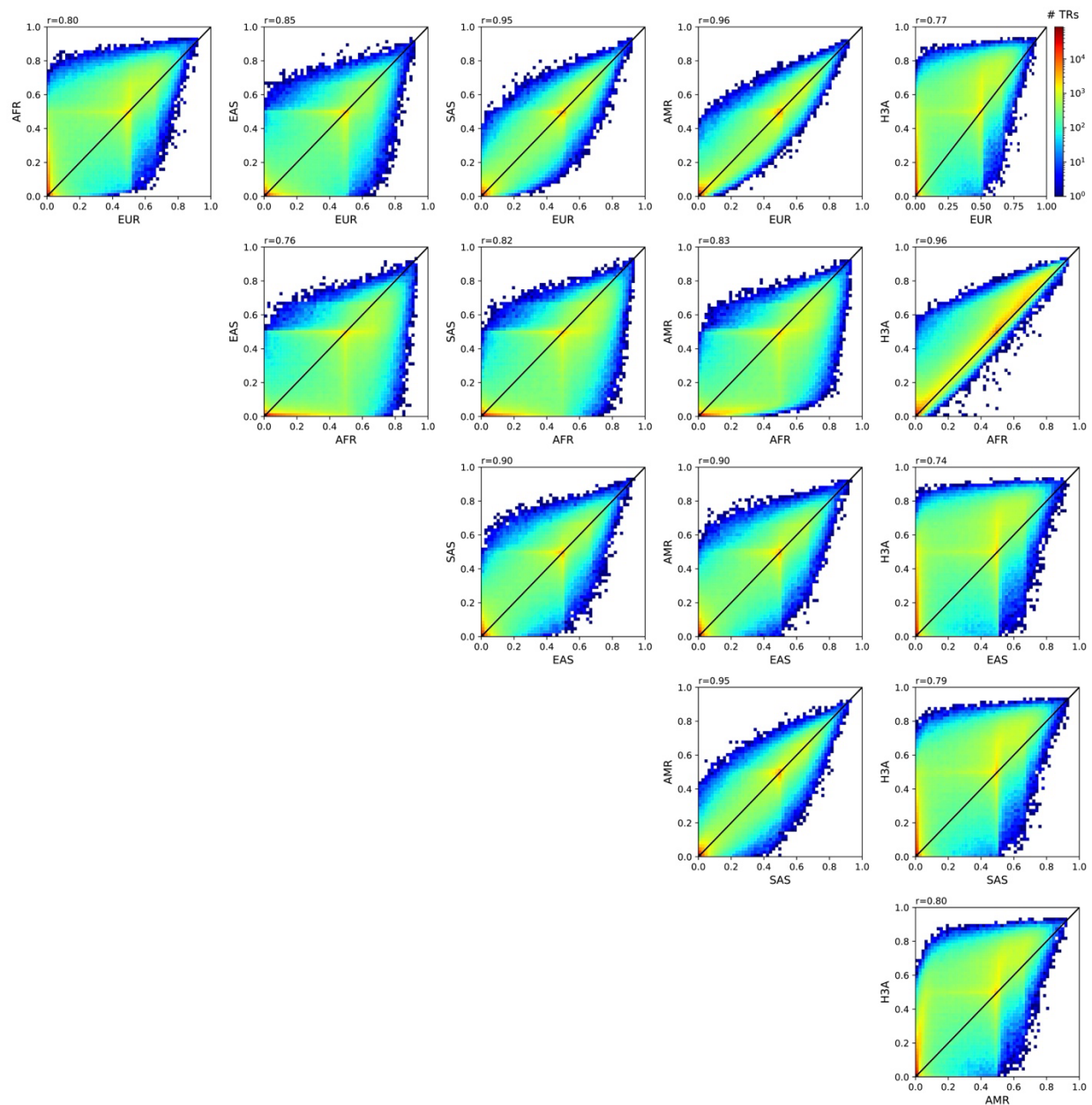


Figure 1.12: Comparison of TR heterozygosity values across populations

Each scatterplot compares the heterozygosity of each TR across two superpopulations. The black line denotes the diagonal. Heatmap shading indicates the number of TRs falling into each bin. The Pearson correlation comparing heterozygosity values in each pair of superpopulations is annotated above each panel. Data is shown for homopolymer TRs.

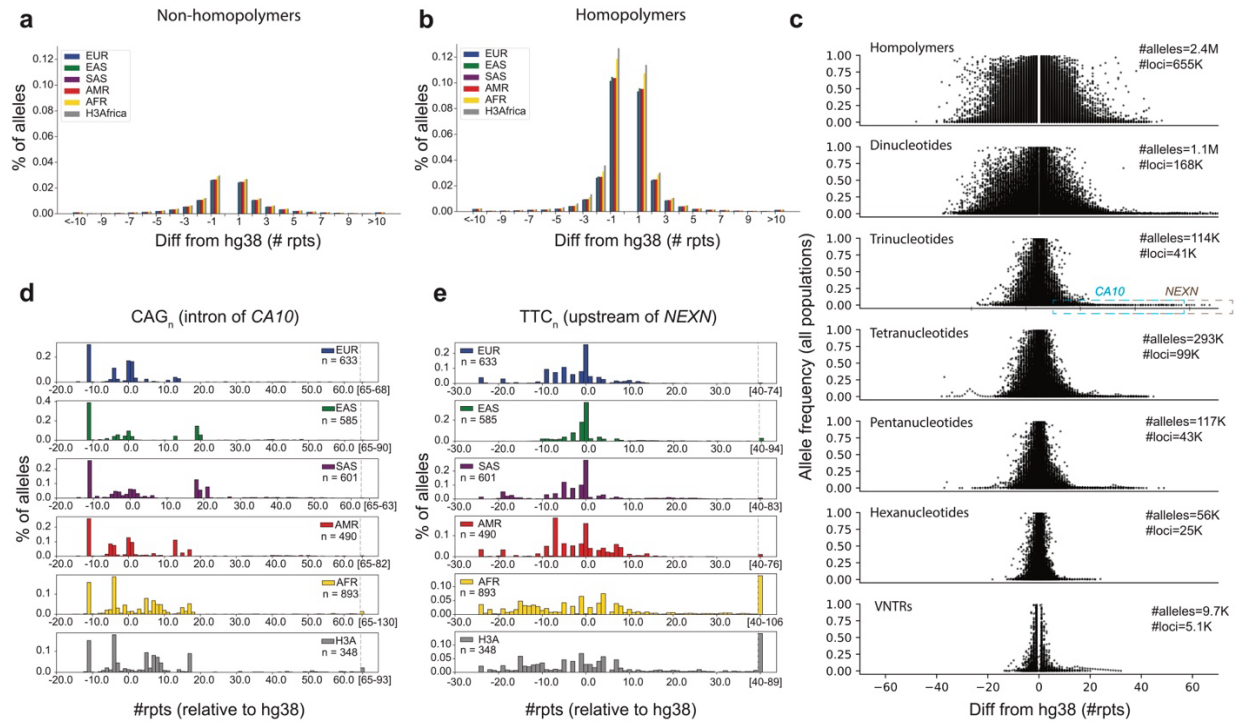


Figure 1.13: Characterizing population-specific TR variation. a-b. Distribution of variant allele sizes

Bars show the percent of variant alleles that have a specified difference in length compared to the hg38 reference. Positive numbers indicate expansions and negative numbers indicate contractions relative to the reference. Panel **a** shows data for all non-homopolymer TRs and **b** shows data for homopolymer TRs. Bar colors denote superpopulations. Gray denotes H3Africa. Other colors denote 1000 Genomes superpopulations. **c. Allele frequency vs. allele length.** The x-axis shows allele lengths relative to the reference genome and the y-axis shows the frequency of each allele across all populations. Different panels denote different repeat unit lengths. Dots corresponding to expansion alleles highlighted in the text are annotated with dashed boxes. Only alleles with frequency at least 0.1% are shown. Alleles with the same length as the reference allele are excluded. **d-e. Population-specific allele distributions at example loci.** In each panel, the x-axis denotes allele length (number of repeats relative to hg38) and the y-axis denotes the frequency of each allele. Each panel shows a different superpopulation. Panel **d** shows a trinucleotide repeat in intron 4 of *CA10*. Panel **e** shows a trinucleotide repeat upstream of *NEXN*. Both repeats have expansion alleles common in African populations compared to non-Africans.

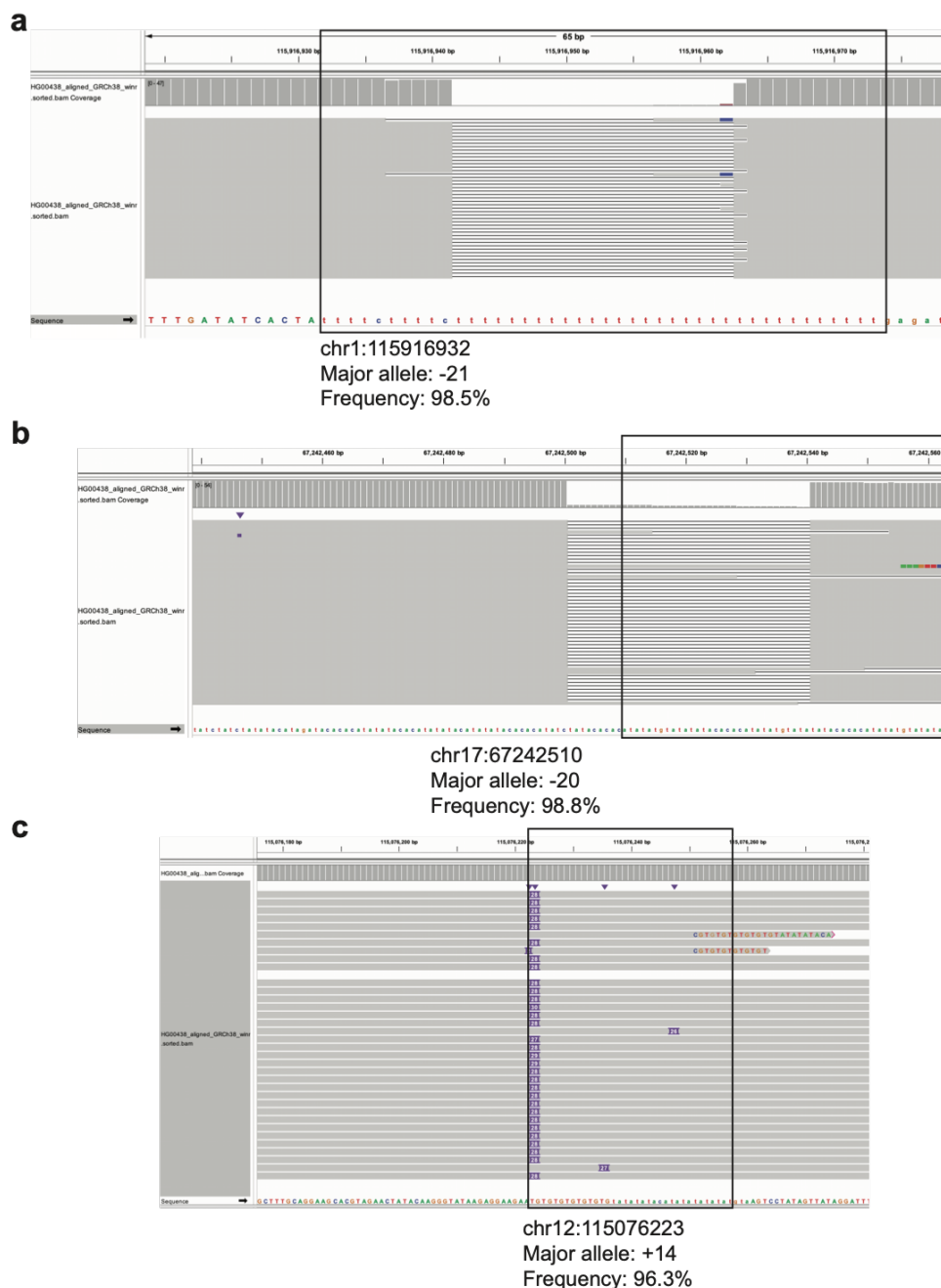


Figure 1.14: Example TRs for which the majority of observed alleles show dramatic deviations from the hg38 reference

In each panel, the TR region is denoted by a black box. Aligned PacBio HiFi reads from HPRC for sample HG00438 are visualized using the Integrative Genomics Viewer (IGV). The major allele length (repeat units relative to hg38) and frequency are annotated in each panel.

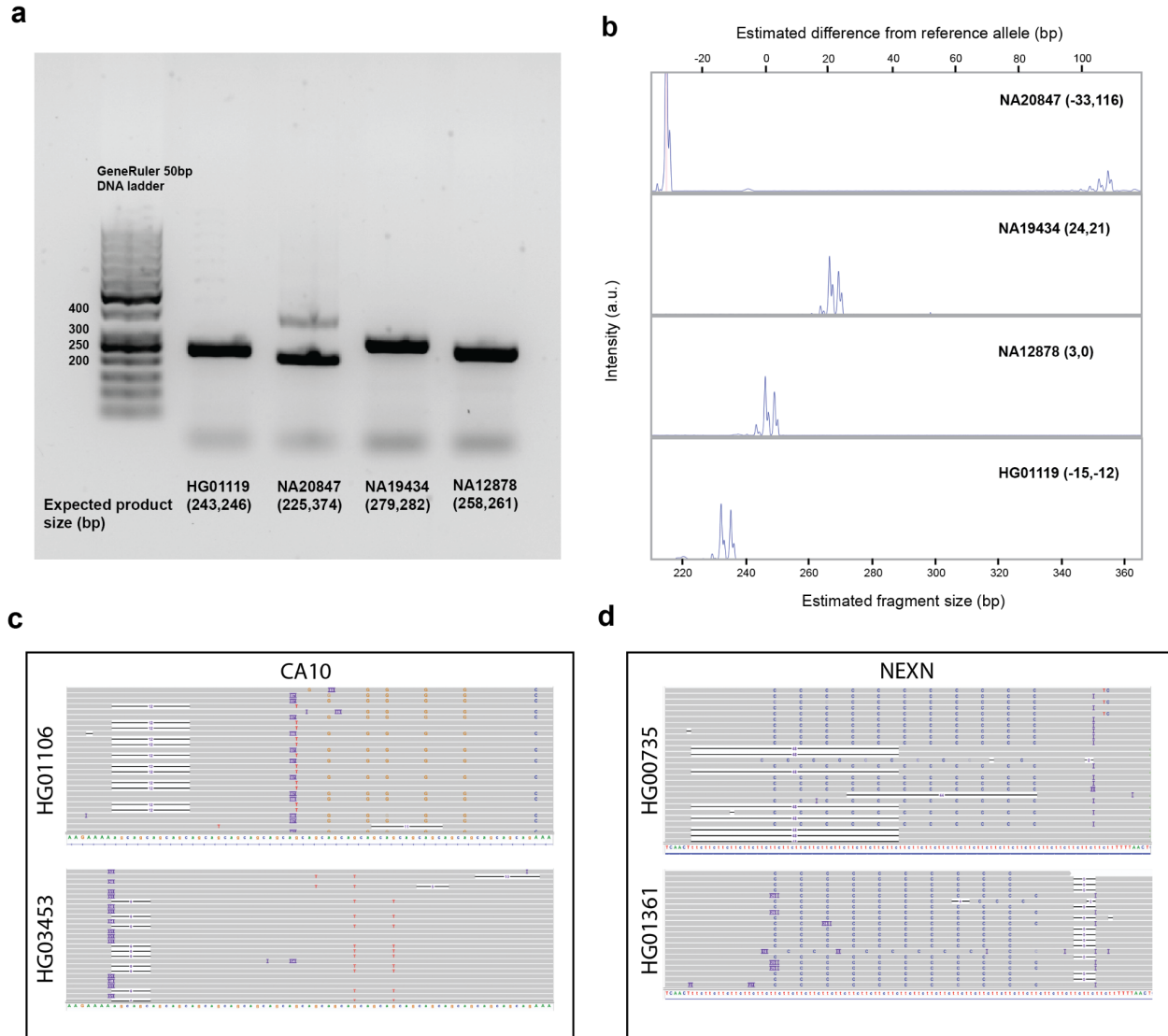


Figure 1.15: Experimental validation of genotypes at the CAG repeat in an intron of CA10

In **a.**, PCR was used to amplify the TR region in four 1000GP samples. Sample names and EnsembleTR calls (expected product size in bp) are annotated below each lane. We identified an expansion in NA20847, which is validated by PCR. Panel **b.** Shows results of fragment analysis on the same PCR products using the GeneMapper software. Samples are annotated with EnsembleTR calls (bp difference from hg38). Panels **c-d** show IGV screenshots of Pacbio HiFi reads aligned to the TRs in *CA10* and *NEXN* for which Africa-specific common repeat expansions were identified.

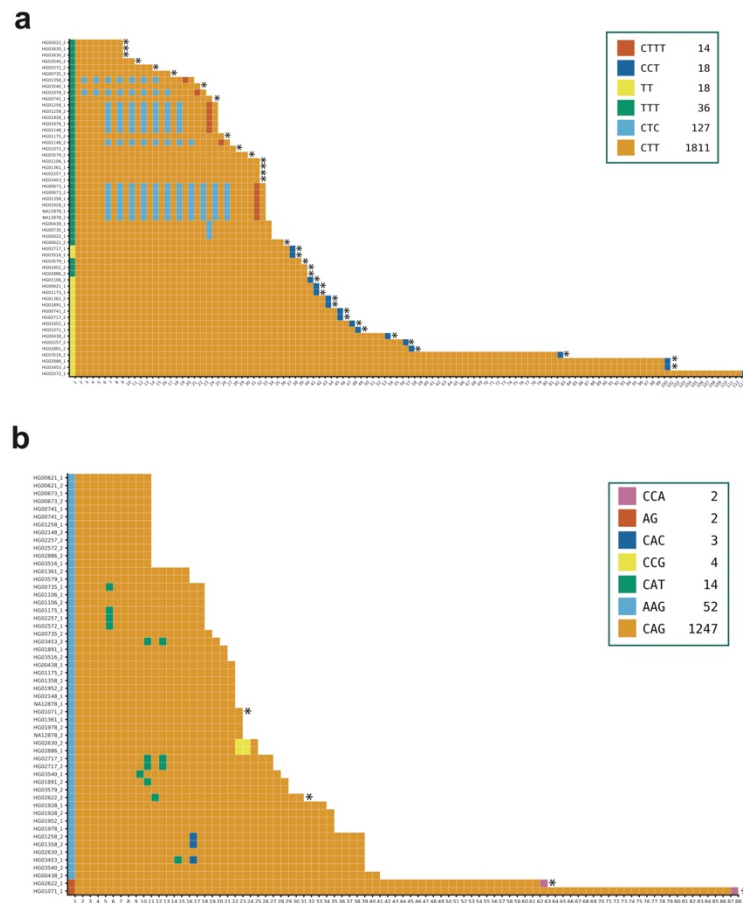
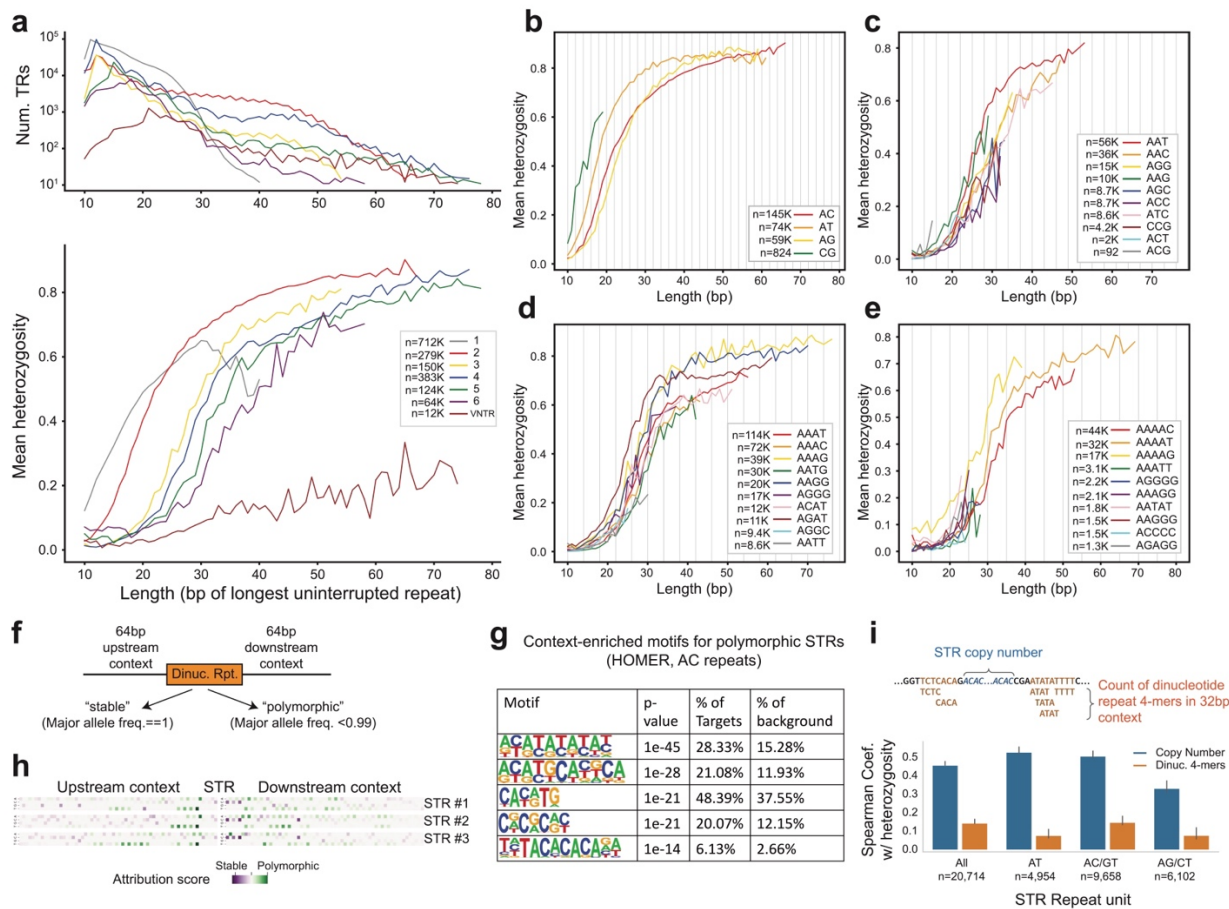


Figure 1.16: Allele sequence visualization with TRviz

We used TRviz to visualize allele sequences for 27 samples in 1000Genomes samples. **a.** Allele sequences for the repeat upstream of *NEXN* described in the text are shown. Although it is annotated as a CTT repeat, many observed alleles consist of repeats of the hexamer sequence CTTCTC. For longer alleles, ExpansionHunter calls were chosen, therefore sequence imperfections are not reported. **b.** Allele sequences for the intronic repeat in *CA10* described in the text are shown. While the repeat is annotated as CAG, other motif sequences such as CAT and CAC are frequently seen. Similar to the *NEXN* repeat, long alleles are genotyped by ExpansionHunter where the sequence imperfections are not reported. A “*” next to a haplotype indicates it was obtained from ExpansionHunter, and therefore no imperfections could be detected.

Figure 1.17: Sequence determinants of TR polymorphism. a. Heterozygosity is correlated with total TR length

The x-axis denotes the length of each TR in hg38 (bp of the longest perfect repeat). The top panel gives the number of repeats in each category. The bottom panel shows the mean heterozygosity for TRs with each length. **b-e** are the same as the bottom panel of **a**, except for different repeat unit sequences. Homopolymers are not shown separately as most have the same repeat unit (A_n). Gray lines are shown every other bp in **b**, every third bp in **c**, every fourth bp in **d**, and every fifth bp in **e**. **f. Overview of approach to classify TRs as stable vs. polymorphic based on sequence context.** We used two approaches (HOMER and convolutional neural networks) to classify dinucleotide TRs based on 64bp of sequence context upstream and downstream of the TR. **g. Top motifs enriched in the context of AC dinucleotide TRs.** All other discovered motifs were flagged as likely false positives by HOMER. One-sided unadjusted binomial P-values are shown. **h. Attribution scores of three example AC TRs predicted to be polymorphic.** Each row denotes a different TR. Within each row, the matrix has a row for each nucleotide and a column for each position (centered on the TR). Color denotes the attribution score of each base in each position (green and purple denote positive and negative contributions toward polymorphism prediction, respectively). **i. Correlation of TR and context features with heterozygosity.** Blue bars denote the Spearman correlation of TR length (reference copy number) with heterozygosity. Orange denotes correlation of the counts of dinucleotide-like or homopolymer-like 4-mers in the context region (± 64 bp) with heterozygosity. Error bars give 99% confidence intervals around the estimated Spearman correlation found by bootstrapping with 1,000 70% subsets.



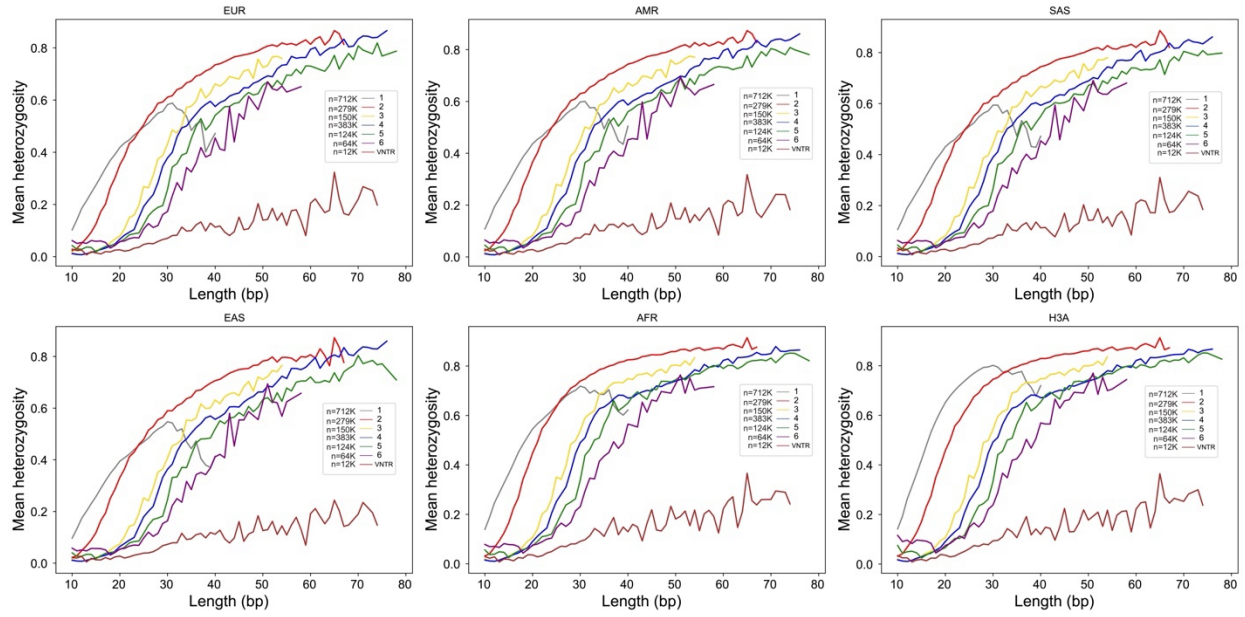


Figure 1.18: Heterozygosity is correlated with total TR length

The x-axis denotes the length of each TR in the reference genome (in total bp of the longest uninterrupted repeat) and the y-axis shows the mean heterozygosity for TRs with each length. Figures are the same as **Fig. 1.17a** but plotted separately based on heterozygosity values in each superpopulation. Colors denote different repeat unit lengths. Superpopulation labels are annotated on the top of each panel.

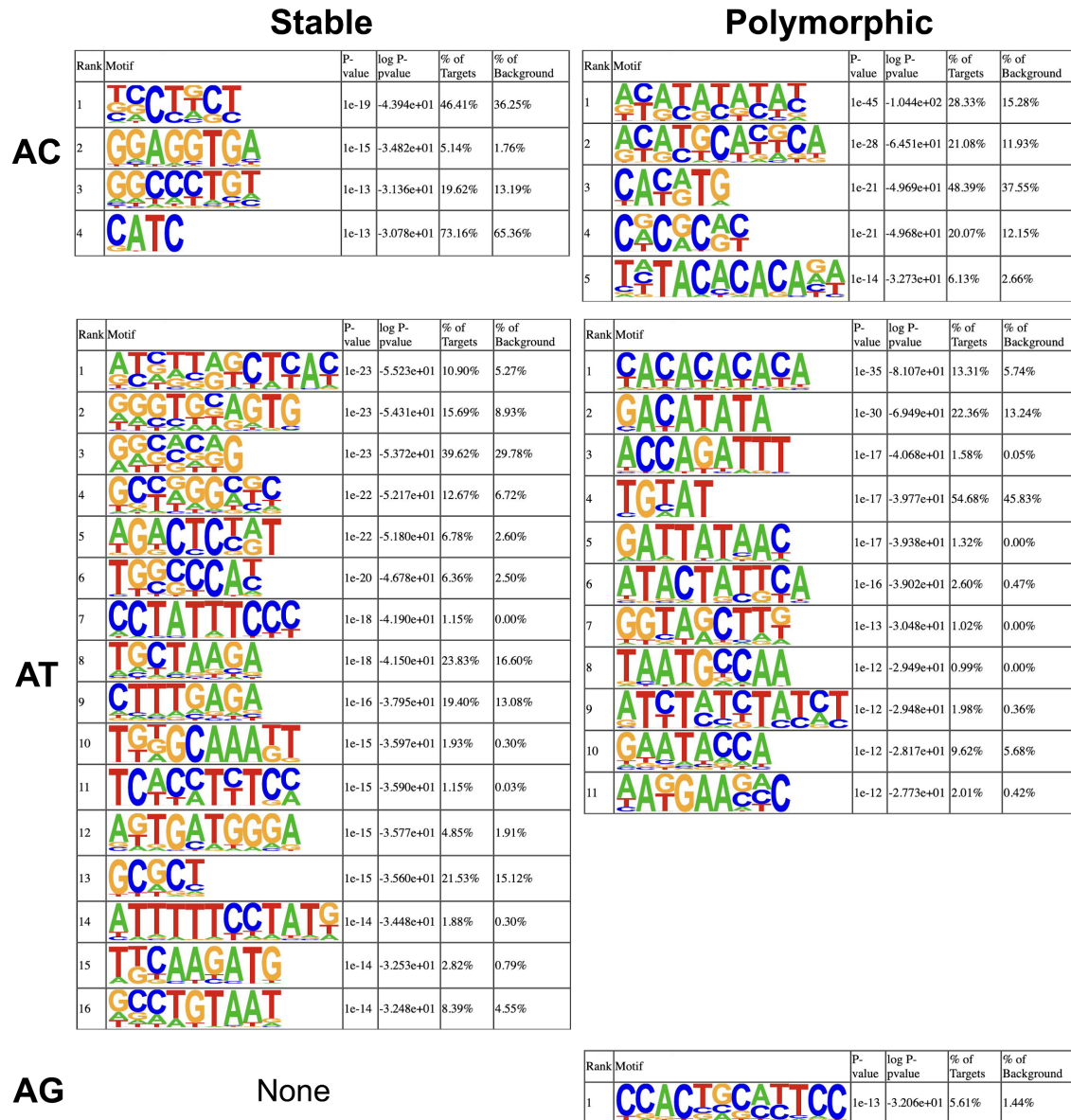


Figure 1.19: Significant HOMER results not marked as likely false positives

Motifs enriched in the context regions for polymorphic vs. stable were identified using HOMER. Results marked as significant by HOMER are shown. Reverse complements were used such that the AC motif category includes information from all AC and GT motif STRs on the forward strand.

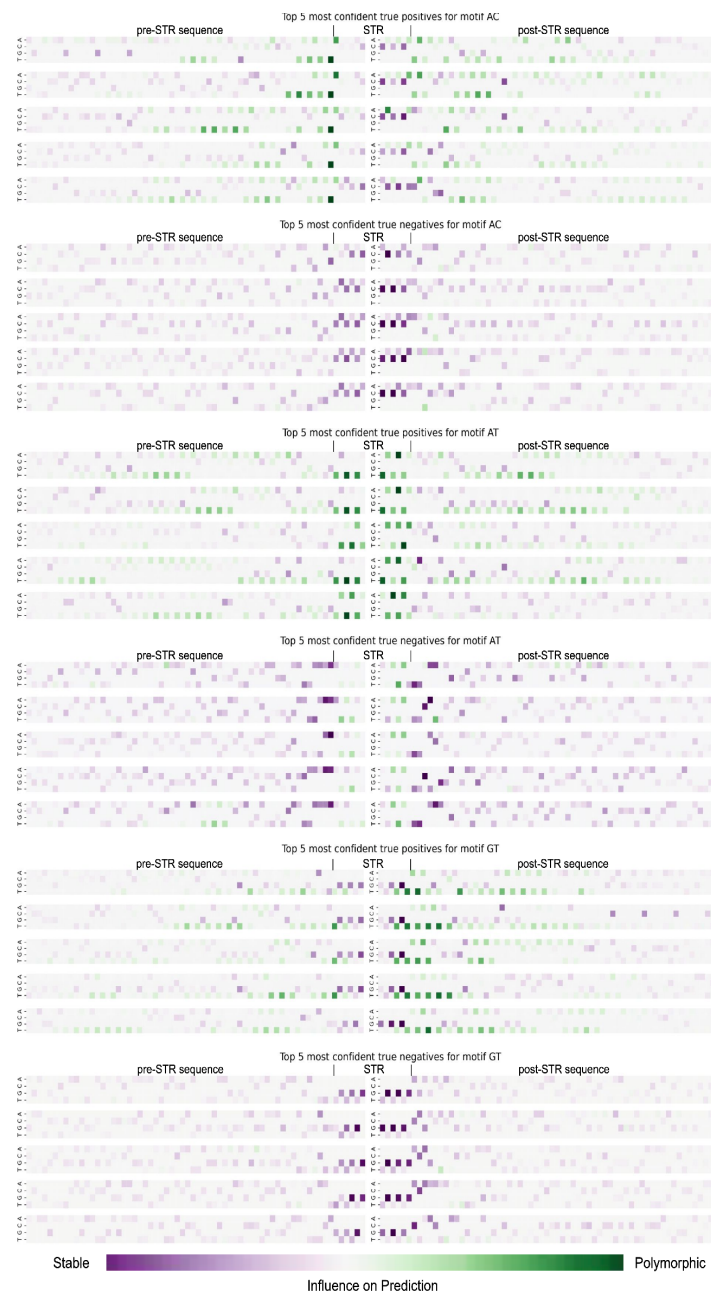


Figure 1.20: Integrated Gradients attribution scores for most confidently predicted true positive (correctly predicted polymorphic) and true negative (correctly predicted stable) STRs by the CNN model

Each row denotes a different TR. Within each row, the matrix has a row for each nucleotide (A, C, G, T) and a column for each position (centered on the TR). Color denotes the attribution score of each base in each position, where green indicates a base positively contributed towards the model predicting polymorphic and purple indicates contributing towards the model predicting stable.

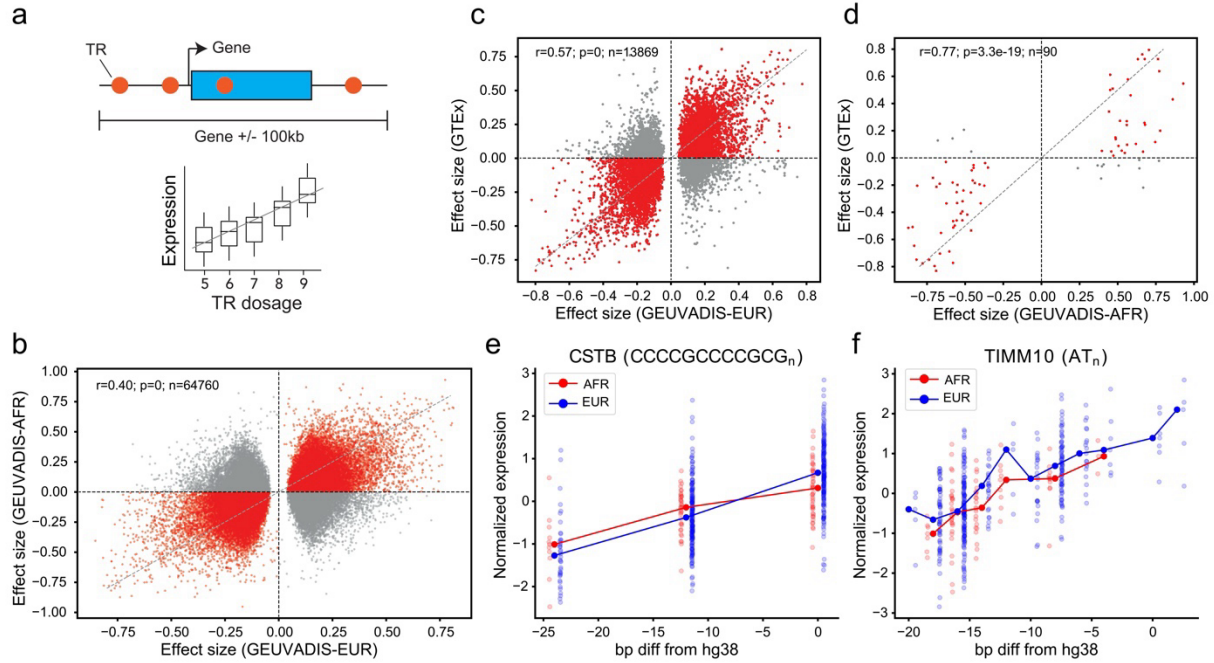


Figure 1.21: TRs associated with gene expression in LCLs. a. Schematic overview of eTR detection

A separate association test between TR dosage (sum of repeat lengths) and expression is performed for each TR within 100kb of a gene. **b. Comparison of effect sizes across populations.** The x-axis gives effect sizes based on European samples and the y-axis gives effect sizes based on African samples from GEUVADIS. Each dot represents a TR-gene pair (eTR). eTRs with consistent effect directions are colored in red. Only eTRs reaching $FDR < 0.05$ in at least one population are included. r gives the Pearson correlation between effect sizes. The P-value is two-sided and not adjusted for multiple comparisons. **c-d. Comparison of effect sizes in GEUVADIS vs. GTEx.** The x-axis gives effect sizes measured in GEUVADIS in Europeans (c) or Africans (d). The y-axis of each plot gives the effect sizes measured in Fotsing et al.¹⁶ in cultured fibroblasts. Each dot represents a TR-gene pair (eTR). Only eTRs with adjusted P-values < 0.05 in the GEUVADIS analysis are shown. r gives Pearson correlation between effect sizes. The P-value is two-sided and not adjusted for multiple comparisons. **e. Example replication of a previously identified eTR.** The x-axis gives the number of repeats of a TR upstream of the gene *CSTB*. The y-axis gives normalized *CSTB* expression. **f. Example novel eTR.** The x-axis gives the number of repeats of a TR near *TIMM10*. The y-axis gives normalized *TIMM10* expression.

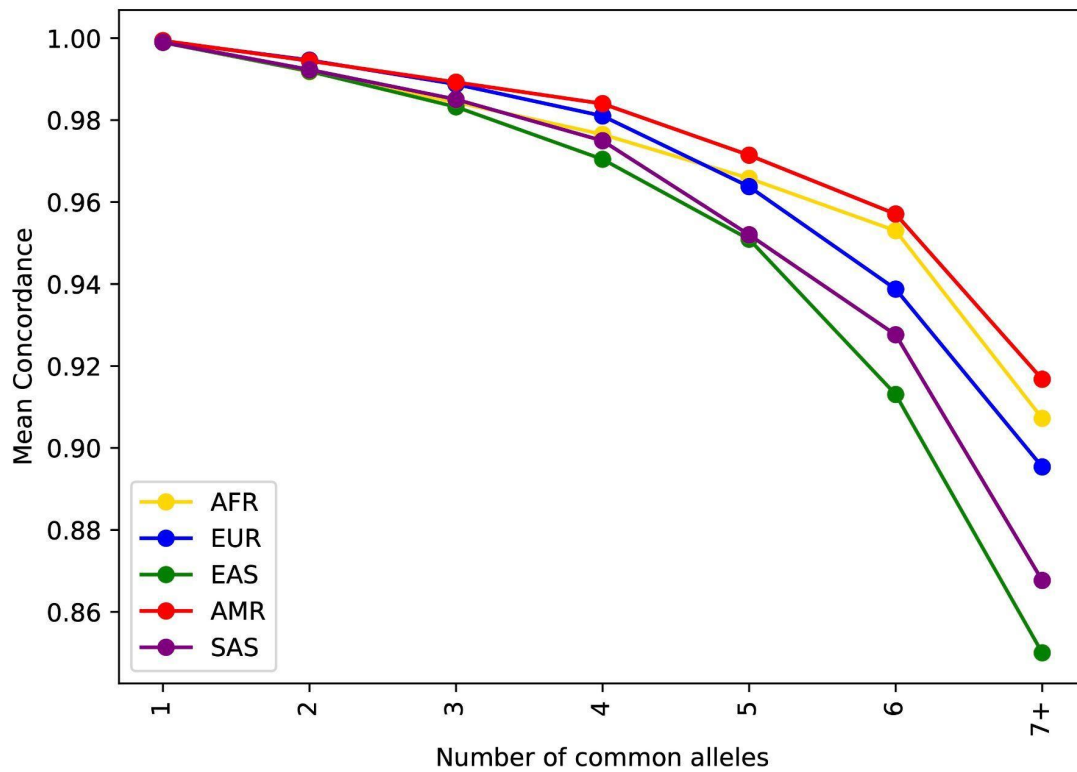


Figure 1.22: Imputation accuracy decreases with the number of common alleles

The x-axis denotes the number of common alleles (frequency > 1%) for each TR. The y-axis denotes the mean imputation concordance for TRs in each bin based on a Leave-One-Out analysis on chromosome 21.

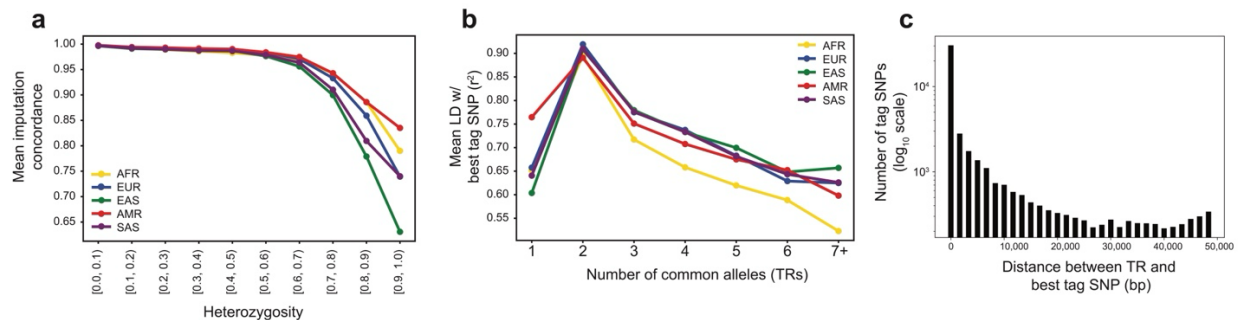


Figure 1.23: Phasing and imputation at TRs. a. Imputation accuracy decreases with heterozygosity

The x-axis denotes TR heterozygosity. The y-axis denotes the mean concordance for TRs in each heterozygosity bin based on a Leave-One-Out analysis on chromosome 21. **b. TRs are often tagged by common SNPs.** The x-axis denotes the number of common alleles (frequency >0.01) for each TR. The y-axis denotes the mean LD (r^2) of the best tag SNP for TRs in each bin. For **a-b**, colors denote 1000G superpopulation. **c. Distribution of the distance between each TR and its best tag SNP.** The y-axis is given on a log₁₀ scale.

1.9 References

1. 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korb, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. M. Byrska-Bishop, U. S. Evani, X. Zhao, A. O. Basile, H. J. Abel, A. A. Regier, A. Corvelo, W. E. Clarke, R. Musunuri, K. Nagulapalli, S. Fairley, A. Runnels, L. Winterkorn, E. Lowy, Human Genome Structural Variation Consortium, Paul Flicek, S. Germer, H. Brand, I. M. Hall, M. E. Talkowski, G. Narzisi, M. C. Zody, High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
3. Whole-genome sequencing of the UK Biobank. [Preprint] (2022). <https://doi.org/10.1038/d41586-022-01984-6>.
4. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Vilems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, doi: 10.1038/nature18964 (2016).
5. J. L. Weber, C. Wong, Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2**, 1123–1128 (1993).
6. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczy, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E.

- Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kasprzyk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
7. M. Bakhtiari, S. Shleizer-Burko, M. Gymrek, V. Bansal, V. Bafna, Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res.* **28**, 1709–1719 (2018).
 8. T. Willems, D. Zielinski, J. Yuan, A. Gordon, M. Gymrek, Y. Erlich, Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods*, doi: 10.1038/nmeth.4267 (2017).
 9. A. J. Hannan, Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **19**, 286–298 (2018).
 10. R. E. Mukamel, R. E. Handsaker, M. A. Sherman, A. R. Barton, Y. Zheng, S. A. McCarroll, P.-R. Loh, Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science* **373**, 1499–1505 (2021).
 11. E. Dolzhenko, M. F. Bennett, P. A. Richmond, B. Trost, S. Chen, J. J. F. A. van Vugt, C. Nguyen, G. Narzisi, V. G. Gainullin, A. M. Gross, B. R. Lajoie, R. J. Taft, W. W. Wasserman, S. W. Scherer, J. H. Veldink, D. R. Bentley, R. K. C. Yuen, M. Bahlo, M. A. Eberle, ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol.* **21**, 102 (2020).
 12. E. Dolzhenko, J. J. F. A. van Vugt, R. J. Shaw, M. A. Bekritsky, M. van Blitterswijk, G. Narzisi, S. S. Ajay, V. Rajan, B. R. Lajoie, N. H. Johnson, Z. Kingsbury, S. J. Humphray, R. D. Schellevis, W. J. Brands, M. Baker, R. Rademakers, M. Kooyman, G. H. P. Tazelaar, M. A. van Es, R. McLaughlin, W. Sproviero, A. Shatunov, A. Jones, A. Al Khleifat, A. Pittman,

- S. Morgan, O. Hardiman, A. Al-Chalabi, C. Shaw, B. Smith, E. J. Neo, K. Morrison, P. J. Shaw, C. Reeves, L. Winterkorn, N. S. Wexler, US–Venezuela Collaborative Research Group, D. E. Housman, C. W. Ng, A. L. Li, R. J. Taft, L. H. van den Berg, D. R. Bentley, J. H. Veldink, M. A. Eberle, Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).
13. N. Mousavi, S. Shleizer-Burko, R. Yanicky, M. Gymrek, Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 (2019).
 14. S. Kristmundsdóttir, B. D. Sigurpálsdóttir, B. Kehr, B. V. Halldórsson, popSTR: population-scale detection of STR variants. *Bioinformatics*, doi: 10.1093/bioinformatics/btw568 (2016).
 15. T. Willems, M. Gymrek, G. Highnam, 1000 Genomes Project Consortium, D. Mittelman, Y. Erlich, The landscape of human STR variation. *Genome Res.* **24**, 1894–1904 (2014).
 16. S. F. Fotsing, J. Margoliash, C. Wang, S. Saini, R. Yanicky, S. Shleizer-Burko, A. Goren, M. Gymrek, The impact of short tandem repeat variation on gene expression. *Nat. Genet.* **51**, 1652–1659 (2019).
 17. M. Bakhtiari, J. Park, Y.-C. Ding, S. Shleizer-Burko, S. L. Neuhausen, B. V. Halldórsson, K. Stefánsson, M. Gymrek, V. Bafna, Variable number tandem repeats mediate the expression of proximal genes. *Nat. Commun.* **12**, 2075 (2021).
 18. I. Mitra, B. Huang, N. Mousavi, N. Ma, M. Lamkin, R. Yanicky, S. Shleizer-Burko, K. E. Lohmueller, M. Gymrek, Patterns of de novo tandem repeat mutations and their role in autism. *Nature* **589**, 246–250 (2021).
 19. F. R. Wendt, G. A. Pathak, R. Polimanti, Phenome-wide association study of loci harboring de novo tandem repeat mutations in UK Biobank exomes. *Nat. Commun.* **13**, 7682 (2022).
 20. B. V. Halldórsson, H. P. Eggertsson, K. H. S. Moore, H. Hauswedell, O. Eiriksson, M. O. Ulfarsson, G. Palsson, M. T. Hardarson, A. Oddsson, B. O. Jensson, S. Kristmundsdóttir, B. D. Sigurpálsdóttir, O. A. Stefánsson, D. Beyter, G. Holley, V. Tragante, A. Gylfason, P. I. Olason, F. Zink, M. Asgeirsdóttir, S. T. Sverrisson, B. Sigurdsson, S. A. Gudjonsson, G. T. Sigurdsson, G. H. Halldórsson, G. Sveinbjörnsson, K. Norland, U. Styrkarsdóttir, D. N. Magnúsdóttir, S. Snorraddóttir, K. Kristinsson, E. Sobeck, H. Jonsson, A. J. Geirsson, I. Olafsson, P. Jonsson, O. B. Pedersen, C. Erikstrup, S. Brunak, S. R. Ostrowski, DBDS Genetic Consortium, G. Thorleifsson, F. Jonsson, P. Melsted, I. Jonsdóttir, T. Rafnar, H. Holm, H. Stefánsson, J. Saemundsdóttir, D. F. Gudbjartsson, O. T. Magnússon, G. Masson, U. Thorsteinsdóttir, A. Helgason, H. Jonsson, P. Sulem, K. Stefánsson, The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
 21. A. Choudhury, S. Aron, L. R. Botigué, D. Sengupta, G. Botha, T. Bensellak, G. Wells, J. Kumuthini, D. Shriner, Y. J. Fakim, A. W. Ghoorah, E. Dareng, T. Odia, O. Falola, E. Adebisi, S. Hazelhurst, G. Mazandu, O. A. Nyangiri, M. Mbiyavanga, A. Benkahla, S. K. Kassim, N. Mulder, S. N. Adebamowo, E. R. Chimusa, D. Muzny, G. Metcalf, R. A. Gibbs, TrypanoGEN Research Group, C. Rotimi, M. Ramsay, H3Africa Consortium, A. A. Adeyemo, Z. Lombard, N. A. Hanchard, High-depth African genomes inform human

- migration and health. *Nature* **586**, 741–748 (2020).
22. M. Gymrek, PCR-free library preparation greatly reduces stutter noise at short tandem repeats. <https://doi.org/10.1101/043448>.
 23. S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, N.-C. Chen, H. Cheng, C.-S. Chin, W. Chow, L. G. de Lima, P. C. Dishuck, R. Durbin, T. Dvorkina, I. T. Fiddes, G. Formenti, R. S. Fulton, A. Functammasan, E. Garrison, P. G. S. Grady, T. A. Graves-Lindsay, I. M. Hall, N. F. Hansen, G. A. Hartley, M. Haukness, K. Howe, M. W. Hunkapiller, C. Jain, M. Jain, E. D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, M. Kremitzki, H. Li, V. V. Maduro, T. Marschall, A. M. McCartney, J. McDaniel, D. E. Miller, J. C. Mullikin, E. W. Myers, N. D. Olson, B. Paten, P. Peluso, P. A. Pevzner, D. Porubsky, T. Potapova, E. I. Rogaev, J. A. Rosenfeld, S. L. Salzberg, V. A. Schneider, F. J. Sedlazeck, K. Shafin, C. J. Shew, A. Shumate, Y. Sims, A. F. A. Smit, D. C. Soto, I. Sović, J. M. Storer, A. Streets, B. A. Sullivan, F. Thibaud-Nissen, J. Torrance, J. Wagner, B. P. Walenz, A. Wenger, J. M. D. Wood, C. Xiao, S. M. Yan, A. C. Young, S. Zarate, U. Surti, R. C. McCoy, M. Y. Dennis, I. A. Alexandrov, J. L. Gerton, R. J. O'Neill, W. Timp, J. M. Zook, M. C. Schatz, E. E. Eichler, K. H. Miga, A. M. Phillippy, The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
 24. H. Dashnow, M. Lek, B. Phipson, A. Halman, S. Sadedin, A. Lonsdale, M. Davis, P. Lamont, J. S. Clayton, N. G. Laing, D. G. MacArthur, A. Oshlack, STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* **19**, 121 (2018).
 25. T. Tsutsumi, S. E. Holmes, M. G. McInnis, A. Sawa, C. Callahan, J. R. DePaulo, C. A. Ross, L. E. DeLisi, R. L. Margolis, Novel CAG/CTG repeat expansion mutations do not contribute to the genetic risk for most cases of bipolar disorder or schizophrenia. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **124B**, 15–19 (2004).
 26. J. Park, E. Kaufman, P. N. Valdmantis, V. Bafna, TRviz: a Python library for decomposing and visualizing tandem repeat sequences. *Bioinform. Adv.* **3**, vbad058 (2023).
 27. J. X. Sun, A. Helgason, G. Masson, S. S. Ebenesersdóttir, H. Li, S. Mallick, S. Gnerre, N. Patterson, A. Kong, D. Reich, K. Stefansson, A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44**, 1161–1165 (2012).
 28. B. A. Payseur, P. Jing, R. J. Haasl, A genomic portrait of human microsatellite variation. *Mol. Biol. Evol.* **28**, 303–312 (2011).
 29. M. Molla, A. Delcher, S. Sunyaev, C. Cantor, S. Kasif, Triplet repeat length bias and variation in the human transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 17095–17100 (2009).
 30. S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, C. K. Glass, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–

589 (2010).

31. T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. C. 't Hoen, J. Monlong, M. A. Rivas, M. González-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padiouleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, Geuvadis Consortium, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Häsler, A.-C. Syvänen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, E. T. Dermitzakis, Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
32. M. Gymrek, T. Willems, A. Guilmatre, H. Zeng, B. Markus, S. Georgiev, M. J. Daly, A. L. Price, J. K. Pritchard, A. J. Sharp, Y. Erlich, Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).
33. C. Borel, E. Migliavacca, A. Letourneau, M. Gagnebin, F. Béna, M. R. Sailani, E. T. Dermitzakis, A. J. Sharp, S. E. Antonarakis, Tandem repeat sequence variation as causative cis-eQTLs for protein-coding gene expression variation: the case of CSTB. *Hum. Mutat.* **33**, 1302–1309 (2012).
34. M. D. Lalioti, H. S. Scott, C. Buresi, C. Rossier, A. Bottani, M. A. Morris, A. Malafosse, S. E. Antonarakis, Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature* **386**, 847–851 (1997).
35. S. Saini, I. Mitra, N. Mousavi, S. F. Fotsing, M. Gymrek, A reference haplotype panel for genome-wide imputation of short tandem repeats. *Nat. Commun.* **9**, 4397 (2018).
36. J. Margoliash, S. Fuchs, Y. Li, A. Massarat, A. Goren, M. Gymrek, Polymorphic short tandem repeats make widespread contributions to blood and serum traits. <https://doi.org/10.1101/2022.08.01.502370>.
37. S. Fazal, M. C. Danzi, V. P. Cintra, D. M. Bis-Brewer, E. Dolzhenko, M. A. Eberle, S. Zuchner, Large scale in silico characterization of repeat expansion variation in human genomes. *Sci Data* **7**, 294 (2020).
38. J. Quilez, A. Guilmatre, P. Garg, G. Highnam, M. Gymrek, Y. Erlich, R. S. Joshi, D. Mittelman, A. J. Sharp, Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res.* **44**, 3750–3762 (2016).
39. G. R. Verheyen, J. Del-Favero, J. Mendlewicz, K. Lindblad, K. Van Zand, M. Aalbrechtse, M. Schalling, D. Souery, C. Van Broeckhoven, Molecular interpretation of expanded RED products in bipolar disorder by CAG/CTG repeats located at chromosomes 17q and 18q. *Neurobiol. Dis.* **6**, 424–432 (1999).

40. J. B. Vincent, Unstable repeat expansion in major psychiatric disorders: two decades on, is dynamic DNA back on the menu? *Psychiatr. Genet.* **26**, 156–165 (2016).
41. D. Hassel, T. Dahme, J. Erdmann, B. Meder, A. Huge, M. Stoll, S. Just, A. Hess, P. Ehlermann, D. Weichenhan, M. Grimmeler, H. Liptau, R. Hetzer, V. Regitz-Zagrosek, C. Fischer, P. Nürnberg, H. Schunkert, H. A. Katus, W. Rottbauer, Nexilin mutations destabilize cardiac Z-disks and lead to dilated cardiomyopathy. *Nat. Med.* **15**, 1281–1288 (2009).
42. B. M. Mayosi, K. Somers, Cardiomyopathy in Africa: heredity versus environment. *Cardiovasc. J. Afr.* **18**, 175–179 (2007).
43. H. Rafehi, D. J. Szmulewicz, M. F. Bennett, N. L. M. Sobreira, K. Pope, K. R. Smith, G. Gillies, P. Diakumis, E. Dolzhenko, M. A. Eberle, M. G. Barcina, D. P. Breen, A. M. Chancellor, P. D. Cremer, M. B. Delatycki, B. L. Fogel, A. Hackett, G. M. Halmagyi, S. Kapetanovic, A. Lang, S. Mossman, W. Mu, P. Patrikios, S. L. Perlman, I. Rosemergy, E. Storey, S. R. D. Watson, M. A. Wilson, D. S. Zee, D. Valle, D. J. Amor, M. Bahlo, P. J. Lockhart, Bioinformatics-based identification of expanded repeats: A non-reference intronic pentamer expansion in RFC1 causes CANVAS. *Am. J. Hum. Genet.* **105**, 151–165 (2019).
44. E. Dolzhenko, A. English, H. Dashnow, G. De Sena Brandine, T. Mokveld, W. J. Rowell, C. Karniski, Z. Kronenberg, M. C. Danzi, W. Cheung, C. Bi, E. Farrow, A. Wenger, V. Martínez-Cerdeño, T. D. Bartley, P. Jin, D. Nelson, S. Zuchner, T. Pastinen, A. R. Quinlan, F. J. Sedlazeck, M. A. Eberle, Resolving the unsolved: Comprehensive assessment of tandem repeats at scale, *bioRxiv* (2023)p. 2023.05.12.540470.
45. W.-W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, S. Lu, J. K. Lucas, J. Monlong, H. J. Abel, S. Buonaiuto, X. H. Chang, H. Cheng, J. Chu, V. Colonna, J. M. Eizenga, X. Feng, C. Fischer, R. S. Fulton, S. Garg, C. Groza, A. Guarracino, W. T. Harvey, S. Heumos, K. Howe, M. Jain, T.-Y. Lu, C. Markello, F. J. Martin, M. W. Mitchell, K. M. Munson, M. N. Mwaniki, A. M. Novak, H. E. Olsen, T. Pesout, D. Porubsky, P. Prins, J. A. Sibbesen, C. Tomlinson, F. Villani, M. R. Vollger, G. Bourque, M. J. P. Chaisson, P. Flicek, A. M. Phillippy, J. M. Zook, E. E. Eichler, D. Haussler, E. D. Jarvis, K. H. Miga, T. Wang, E. Garrison, T. Marschall, I. Hall, H. Li, B. Paten, Human Pangenome Reference Consortium, A Draft Human Pangenome Reference, *bioRxiv* (2022).
<https://doi.org/10.1101/2022.07.09.499321>.
46. G. Benson, Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
47. N. Mousavi, J. Margoliash, N. Pusarla, S. Saini, R. Yanicky, M. Gymrek, TRTools: a toolkit for genome-wide analysis of tandem repeats. *Bioinformatics* **37**, 731–733 (2021).
48. D. Karolchik, A. S. Hinrichs, W. J. Kent, The UCSC Genome Browser. *Curr. Protoc. Bioinformatics* **Chapter 1**, Unit1.4 (2012).
49. Y. Wang, S. Kikuchi, H. Suzuki, S. Nagase, A. Koyama, Endothelial nitric oxide synthase gene polymorphism in intron 4 affects the progression of renal failure in non-diabetic renal diseases. *Nephrol. Dial. Transplant* **14**, 2898–2902 (1999).

50. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, D. Haussler, The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
51. M. Schuelke, An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* **18**, 233–234 (2000).
52. A. De Luca, A. Morella, F. Consoli, S. Fanelli, J. R. Thibert, S. Statt, G. J. Latham, F. Squitieri, A Novel Triplet-Primed PCR Assay to Detect the Full Range of Trinucleotide CAG Repeats in the Huntingtin Gene (). *Int. J. Mol. Sci.* **22** (2021).
53. E. Suh, K. Grando, V. M. Van Deerlin, Validation of a Long-Read PCR Assay for Sensitive Detection and Sizing of C9orf72 Hexanucleotide Repeat Expansions. *J. Mol. Diagn.* **20**, 871–882 (2018).
54. S. Filipovic-Sadic, S. Sah, L. Chen, J. Krosting, E. Sekinger, W. Zhang, P. J. Hagerman, T. T. Stenzel, A. G. Hadd, G. J. Latham, F. Tassone, A novel FMR1 PCR method for the routine detection of low abundance expanded alleles and full mutations in fragile X syndrome. *Clin. Chem.* **56**, 399–408 (2010).
55. P. Shrivastava, T. Jain, V. B. Trivedi, Genetic polymorphism study at 15 autosomal locus in central Indian population. *Springerplus* **4**, 566 (2015).
56. N. Sarkar, V. K. Kashyap, Genetic diversity at two pentanucleotide STR and thirteen tetranucleotide STR loci by multiplex PCR in four predominant population groups of central India. *Forensic Sci. Int.* **128**, 196–201 (2002).
57. N. Masuda, J. Goto, N. Murayama, M. Watanabe, I. Kondo, I. Kanazawa, Analysis of triplet repeats in the huntingtin gene in Japanese families affected with Huntington's disease. *J. Med. Genet.* **32**, 701–705 (1995).
58. F. K. Baine, C. Kay, M. E. Ketelaar, J. A. Collins, A. Semaka, C. N. Doty, A. Krause, L. J. Greenberg, M. R. Hayden, Huntington disease in the South African population occurs on diverse and ethnically distinct genetic haplotypes. *Eur. J. Hum. Genet.* **21**, 1120–1127 (2013).
59. Q. Saleem, S. Roy, U. Murgood, R. Saxena, I. C. Verma, A. Anand, U. Muthane, S. Jain, S. K. Brahmachari, Molecular analysis of Huntington's disease and linked polymorphisms in the Indian population. *Acta Neurol. Scand.* **108**, 281–286 (2003).
60. I. Paradisi, A. Hernández, S. Arias, Huntington disease mutation in Venezuela: age of onset, haplotype analyses and geographic aggregation. *J. Hum. Genet.* **53**, 127–135 (2008).
61. K. K. Ambrose, T. Ishak, L.-H. Lian, K.-J. Goh, K.-T. Wong, A. Ahmad-Annuar, M.-K. Thong, Analysis of CTG repeat length variation in the gene in the general population and the molecular diagnosis of myotonic dystrophy type 1 in Malaysia. *BMJ Open* **7**, e010711 (2017).
62. R. T. Acton, C. A. Rivers, B. Watson, S. J. Oh, DMPK-associated myotonic dystrophy and CTG repeats in Alabama African Americans. *Clin. Genet.* **72**, 448–453 (2007).

63. J. J. Magaña, P. Cortés-Reynosa, R. Escobar-Cedillo, R. Gómez, N. Leyva-García, B. Cisneros, Distribution of CTG repeats at the DMPK gene in myotonic distrophy patients and healthy individuals from the Mexican population. [Preprint] (2011). <https://doi.org/10.1007/s11033-010-0235-7>.
64. E. Majounie, M. Wardle, M. Muzaimi, W. C. Cross, N. P. Robertson, N. M. Williams, H. R. Morris, Case control analysis of repeat expansion size in ataxia. *Neurosci. Lett.* **429**, 28–32 (2007).
65. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Í. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
66. R. Garreta, G. Moncecchi, *Learning Scikit-Learn: Machine Learning in Python* (Packt Pub Limited, 2013).
67. T. Wang, L. Antonacci-Fulton, K. Howe, H. A. Lawson, J. K. Lucas, A. M. Phillippy, A. B. Popejoy, M. Asri, C. Carson, M. J. P. Chaisson, X. Chang, R. Cook-Deegan, A. L. Felsenfeld, R. S. Fulton, E. P. Garrison, N. A. Garrison, T. A. Graves-Lindsay, H. Ji, E. E. Kenny, B. A. Koenig, D. Li, T. Marschall, J. F. McMichael, A. M. Novak, D. Purushotham, V. A. Schneider, B. I. Schultz, M. W. Smith, H. J. Sofia, T. Weissman, P. Flicek, H. Li, K. H. Miga, B. Paten, E. D. Jarvis, I. M. Hall, E. E. Eichler, D. Haussler, the Human Pangenome Reference Consortium, The Human Pangenome Project: a global resource to map genomic diversity. [Preprint] (2022). <https://doi.org/10.1038/s41586-022-04601-8>.
68. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative genomics viewer. [Preprint] (2011). <https://doi.org/10.1038/nbt.1754>.
69. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
70. Z. Zhang, S. Schwartz, L. Wagner, W. Miller, A Greedy Algorithm for Aligning DNA Sequences. [Preprint] (2000). <https://doi.org/10.1089/10665270050081478>.
71. Website. doi:10.18129/B9.BIOC.ORGANISMDBI.
72. K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
73. H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, F. Petitjean, InceptionTime: Finding AlexNet for time series classification. *Data Min. Knowl. Discov.* **34**, 1936–1962 (2020).
74. M. Sundararajan, A. Taly, Q. Yan, Axiomatic Attribution for Deep Networks. doi:

10.48550/arXiv.1703.01365 (2017).

75. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
76. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
77. A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
78. O. Stegle, L. Parts, M. Piipari, J. Winn, R. Durbin, Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
79. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. [Preprint] (1995). <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
80. S. R. Browning, B. L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
81. B. L. Browning, X. Tian, Y. Zhou, S. R. Browning, Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
82. H. Ziaei Jam, Y. Li, R. DeVito, N. Mousavi, N. Ma, I. Lujumba, Y. Adam, M. Maksimov, B. Huang, E. Dolzhenko, Y. Qiu, F. E. Kakembo, H. Joseph, B. Onyido, J. Adeyemi, M. Bakhtiari, J. Park, S. Javadzadeh, D. Jjingo, E. Adebisi, V. Bafna, M. Gymrek, A deep population reference panel of tandem repeat variation, *EnsembleTR* (2023). <https://doi.org/10.5281/zenodo.8365672>.

2 Chapter 2: LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads

Abstract: Tandem repeats are frequent across the human genome, and variation in repeat length has been linked to a variety of traits. Recent improvements in long read sequencing technologies have the potential to greatly improve tandem repeat analysis, especially for long or complex repeats. Here we introduce LongTR, which accurately genotypes tandem repeats from high fidelity long reads available from both PacBio and Oxford Nanopore Technologies. LongTR is freely available at <https://github.com/gymrek-lab/longtr> and <https://zenodo.org/doi/10.5281/zenodo.11403979>.

2.1 Introduction

Tandem repeats (TRs), including short tandem repeats (STRs; repeat unit 1-6bp) and variable number tandem repeats (VNTRs; repeat unit 7+bp), refer to regions of the genome that consist of adjacent repeated units. TRs are a large source of genetic variation in humans¹ and are implicated in a growing list of Mendelian and complex traits². In the last decade, multiple tools have been developed to estimate the repeat length and/or sequence of TRs using short reads (e.g.[3–6]) but certain repeats such as highly complex TRs have remained intractable. Long-read sequencing technologies offer a promising solution. However, tools designed for short reads are ineffective on long reads given the considerable differences including in read length, base calling accuracy, error profiles at STRs (Fig. 2.1) and paired-end vs. single-end format.

2.2 Results and Discussion

Here, we introduce LongTR, which extends the HipSTR³ method originally developed for short read STR analysis in order to genotype STRs and VNTRs from accurate long reads available from both PacBio⁷ and Oxford Nanopore Technologies⁸ (ONT). LongTR takes as input sequence

alignments for one or more samples and a reference set of TRs and outputs the inferred sequence and length of each allele at each locus. It uses a clustering strategy combined with partial order alignment to infer consensus haplotypes from error-prone reads, leveraging read phase information when available, followed by sequence realignment using a Hidden Markov Model which is used to score each possible diploid genotype at each locus (Methods; Fig. 2.2). Unlike other existing long read TR genotypers, LongTR supports multi-sample calling, employs a technology-specific homopolymer error model, and outputs genotype quality scores.

We ran LongTR to genotype repeats in a reference set of 937,122 human TRs from Project Adotto⁹ using 30x PacBio HiFi reads from HG002 (Data Availability). To evaluate the accuracy of genotype calls, we extracted alleles for genotyped TRs from the haplotype-resolved genome assembly of HG002¹⁰, which was generated using multiple technologies and orthogonal computational methods, and is thus treated as a ground truth here. LongTR outputs inferred allele sequences in addition to repeat length, enabling sequence level comparisons of the alleles to the assembly. Of 814,319 repeats for which exactly two alleles were extracted from the assembly, LongTR showed 84.7% sequence concordance and 98.5% length concordance allowing for 1 bp differences. To further evaluate the accuracy of LongTR, we performed TR genotyping in an Ashkenazi trio (HG002, HG003, and HG004) using the same reference set and determined the Mendelian inheritance (MI) of TR genotypes. Overall, LongTR showed 86% MI at sites where at least one trio member was not homozygous for the reference allele. Mendelian consistency monotonically increases with genotype quality scores computed by LongTR, which can be used to filter low quality calls (Fig. 2.3a). Homopolymer repeats showed the lowest consistency with MI of 78.3%. We trained a PacBio HiFi homopolymer error model at 840,248 homopolymer repeats from the HipSTR reference, which contains homopolymers with more precise boundaries

than the Adotto set, by comparing observed repeat lengths at HiFi reads to those obtained from the HG002 assembly (Table 2.1). In the Ashkenazi trio we observed an MI of 83% and 81% for LongTR with and without error modeling (Fig. 2.4). Further, we observed a 13% increase in concordance of LongTR alleles with the HG002 assembly when using the homopolymer error model.

Multiple methods for genotyping TRs from long reads have been developed over the last several years^{11–14}. We first focused on benchmarking LongTR against TRGT¹⁵, a recently developed TR genotyper that outperforms previous methods on PacBio HiFi reads. Similar to LongTR, TRGT outputs inferred sequences in addition to estimated repeat length. We ran TRGT on HG002 using the same sequencing data and reference set as LongTR. Both tools had similar running time, with TRGT taking 436 minutes and LongTR taking 428 minutes to finish. TRGT and LongTR genotyped 99.83% and 99.18% of the repeats respectively. At sites called by both methods, 86.0% of alleles have identical lengths and 98.5% differ in length by at most a single copy number. TRGT calls in the Ashkenazi trio showed reduced Mendelian inheritance rates compared to LongTR (79% for TRGT vs. 86% for LongTR).

LongTR and TRGT showed similar length concordance with the HG002 assembly, with LongTR performing slightly better (97.8% for TRGT and 98.5% for LongTR allowing for 1 bp differences). LongTR showed further gains when evaluating sequence concordance. In both cases, the advantage of LongTR over TRGT was highest at long (>500bp) repeats (Fig. 2.3b). We identified multiple scenarios in which LongTR calls match the assembly and TRGT does not, including regions with a high number of truncated reads or regions called as large insertions by TRGT that had low read support. Further, LongTR detected 514 TRs with large structural deletions that remove the entire repeat, resulting in a null allele, whereas these cases are not reported by

TRGT (Fig. 2.4). Overall, these results suggest both tools perform similarly at TRs in regions that are easier to genotype, whereas LongTR obtains higher quality genotypes at longer or structurally complex regions. Notably, we observed multiple instances where both LongTR and TRGT reported identical genotypes that differed from the assembly. These cases usually occurred in complex genomic regions such as large insertions or segmental duplications or regions of high homozygosity. While complex regions tend to be more accurate with assembly-based approaches, regions of high homozygosity are known to pose challenges to diploid assemblies¹⁶ and thus likely represent assembly errors rather than errors from LongTR or TRGT (Fig. 2.5).

We next benchmarked LongTR against two alternative methods for genotyping TRs from long reads, Straglr¹⁴ and Tandem-genotypes¹³. Both comparisons were done using 30x PacBio HiFi reads for HG002. Unlike LongTR and TRGT, these methods output estimated repeat copy number but not inferred allele sequences at each locus. We ran Straglr and LongTR to genotype repeats from HipSTR reference set, which contains repeats with simpler motif structure, as Straglr automatically infers the motif sequence and more complex repeats caused large discrepancies between LongTR and Straglr calls. Out of 20,592 repeats on chromosome 21, LongTR genotyped 20,158 and Straglr genotyped 13,705. From 13,655 repeats genotyped by both, they matched on both alleles for 47% of the repeats and matched for one allele for another 5% allowing for 1bp off. For benchmarking LongTR against Tandem-genotypes, we used the Adotto reference set as input to both tools. Out of 12,802 repeats on chromosome 21, Tandem-genotypes and LongTR genotyped 97% and 96% of regions respectively. Regions skipped by LongTR lacked a sufficient number of high quality reads and were filtered. Since Tandem-genotypes only reports the copy number, we checked if the difference between copy number of LongTR alleles and Tandem-genotypes was less than 1. Out of 12,232 repeats genotyped by both, LongTR and Tandem-

genotypes agreed on both alleles for 75% of TRs and agreed on one allele in 24% of repeats. Overall these comparisons suggest modest concordance between Straglr and Tandem-genotypes vs. LongTR which may be driven in part by differences in how alleles are reported (imprecise copy number count vs. allele sequences).

We sought to further assess the ability of LongTR to genotype longer repeats. First, we compared LongTR genotypes in HG002 to those from adVNTR¹⁷, a tool specifically designed for genotyping VNTRs, on a reference set of 10,186 autosomal gene-proximal VNTRs (Data Availability). When allowing for differences in up to a single copy number (Methods), LongTR and adVNTR showed 96% concordance. adVNTR required approximately 23 hours to genotype HG002, compared to 1.5 hours for LongTR on this repeat set. Second, we determined whether LongTR could identify large expansions in HiFi reads obtained from patients harboring long pathogenic alleles implicated in Huntington's disease (n=4) and Fragile X Syndrome (n=3). LongTR correctly identified expansions in HTT and FMR1, including alleles consisting of up to several thousand bp for both loci (Table 2.2). Allele sequences reported by LongTR match repeat unit copy number counts for these reads reported on the dataset website for the tested samples (Data Availability, Fig. 2.7, visualized using TRviz¹⁸).

We next evaluated the ability of LongTR to genotype TRs in a separate long read technology, using ONT's recently released Duplex reads (average length 27kb compared to 15-20kb for PacBio HiFi reads) available for HG002¹⁹. Overall, we observed high concordance between genotypes obtained from phased ONT duplex and PacBio HiFi (90% allowing for 1bp off; Fig. 2.8a) with the latter showing higher concordance with the assembly (99% for PacBio HiFi vs 88% for ONT, allowing for 1bp off; n=798,291 TRs called for both ONT and PacBio HiFi). We found that phasing ONT reads improved performance, with 87% and 89% concordance against the

assembly for unphased and phased reads, respectively, across 12,503 TRs on Chromosome 21. We further evaluated ONT using haplotagged simplex SUP reads on Chromosome 21 and found that SUP and duplex reads showed similar concordance with the assembly (88.9% for duplex vs. 88.7% for SUP).

We identified several large repeat expansions uniquely detected by the ONT and not PacBio HiFi data, which were consistent with the alleles in the assembly. These repeats are enriched in the GIAB set of hard-to-map regions²⁰ (Fisher's Exact Test two-sided $p=1.75e-292$) (Data Availability). Examining these expansions showed that discrepancies often occurred in regions where few or no PacBio HiFi reads aligned to or spanned the insertion, leading LongTR to inaccurately genotype the locus (Fig. 2.8b). Our observations suggest that the longer read lengths of ONT enhance the detection accuracy particularly for regions with large insertions compared to the reference.

Finally, we compared TR genotypes obtained from HipSTR on Illumina short reads and LongTR on PacBio HiFi reads for HG002. For this analysis, we used the hg38 reference set of 1,638,945 STRs available from HipSTR's website, of which only 1.6% are longer than 100bp. Of 1,556,278 STRs that were genotyped by both methods, 88% were concordant, increasing to 97% if allowing for 1bp length difference (Fig. 2.9a). HipSTR reported homozygous reference for all repeats with length above 250bp (the Illumina read length) and no read distribution information, indicating genotypes for longer repeats are not reliable. Concordance (by allowing 1bp off) decreases with increasing length of the repeat (Fig. 2.9b).

2.3 Conclusions

Overall, LongTR provides accurate sequence-resolved TR genotyping from long reads for nearly 99% of TRs in mappable regions of the genome and outperforms existing methods for this

task. While the majority of TRs can be resolved, we identified multiple TRs with large, complex insertions relative to the reference that are challenging to span even with long reads and may in some cases be misrepresented by reference assemblies. These cases may represent the limits of mapping-based approaches. Future work is needed to incorporate alternative approaches, such as pangenome or assembly-based methods, that do not suffer from these limitations. We envision these improvements will enable systematic incorporation of TRs into genome-wide analyses for a range of applications.

2.4 Methods

2.4.1 Overview of the LongTR method

LongTR extends HipSTR³, which was originally developed to analyze short reads, to genotype both VNTRs and STRs using accurate long reads. Here, we use accurate long reads to refer to PacBio HiFi and ONT duplex reads, each of which have been shown to have per-base error rates comparable to those of Illumina reads^{22,23}. Like HipSTR, LongTR begins with aligned reads for one or more samples and the reference coordinates for a predefined set of TRs. Predefined TRs may be simple repeats (e.g. [AC]*n*) but can also comprise more complex TRs with multiple distinct repeat units (e.g. [AC]*n*[GT]*k*[T]*l*). Then for each TR, it extracts reads encompassing the repeat, infers candidate TR haplotypes, and uses a Hidden Markov Model to realign all the reads overlapping a repeat region to the candidate haplotypes. Finally, it outputs a VCF file containing each individual's TR genotypes and corresponding quality scores (Fig. 2.2). Below we discuss key steps where LongTR differs from the HipSTR method to enable analysis of TRs from long reads.

Haplotype identification: LongTR uses a new method for the identification of candidate haplotypes from input reads of samples at each TR. It starts by trimming all reads aligned to the target TR of interest to include the repeat plus a user-defined window of context sequence. It then

iterates over all trimmed sequences and includes any sequence supported by a sufficient number of reads by one or more samples (at least two reads and more than 20% of reads in a single sample, or more than 5% of all reads across all samples) as a candidate haplotype.

In some scenarios, typically when the repeat is long or the region is complex with multiple insertions and deletions, reads from a haplotype fail to meet the criteria set in the first stage, resulting in their exclusion from the set of haplotypes. To address this, we perform a second iteration during which we identify samples for which over 25% of aligned reads lack a corresponding representative haplotype and form additional candidate haplotypes using excluded reads for each of these samples. These previously excluded reads for each sample are then sorted by length of the sequence aligned to the repeat region, after which a greedy clustering algorithm is applied to form the initial sequence clusters. In this method, the first cluster is formed by designating the first sequence as its centroid. Starting with the second sequence S we evaluate whether there exists a centroid C within the centroids set for which the edit distance between S and C is below a given threshold T . T is initially set to a small number (10). We then refine the initial sequence clusters through the following steps: (1) A consensus sequence for each cluster is generated using partial order alignment²⁴ (POA) on read sequences within each cluster. The consensus is used to update the cluster centroid. (2) After updating the centroids for all clusters, we again sort the clusters in order of the length of their revised centroids. (3) We iterate through the clusters, merging two clusters whenever the edit distance between their updated centroids is below T . (4) This process is repeated until no further clusters can be merged. (5) After cluster refinement, we only include clusters with a number of reads above $\min(10, 0.1 \times n_s)$ where n_s is the number of excluded reads per sample s . Finally, we check if the total number of reads in all remaining clusters are more than $0.8 \times n_s$, otherwise we will increase T (attempting in order the

following values: 20, 50, 80, 100, 150, 200, 300, 400, 500, 600, 700) to relax the constraints on sequence similarity for both adding sequences to existing clusters and merging cluster and repeat the steps above. Centroid sequences of final clusters are added to the set of potential haplotypes. Edit distance is calculated using Needleman-Wunsch algorithm with parameters $\text{gap_score} = 1$, $\text{match_score} = 0$, $\text{mismatch_score} = 1$. We optimized the algorithm by computing a lower bound of total edit distance at each row of the dynamic programming table $\text{dp}[n][m]$, where n and m are length of sequences. For each cell in a row i , we iterate over columns j and compute $\min(\text{dp}[i, j] + \text{abs}((n - m) - (i - j)))$. This value shows the minimum number of insertions or deletions needed to reach cell $\text{dp}[n, m]$. If the minimum value per row exceeds the threshold T , computation stops.

Alignment of reads to candidate haplotypes: The original HipSTR approach employs a rigid alignment technique for aligning repeat sequences to candidate haplotypes, operating under the assumption that errors occur in multiples of the repeat unit length and happen only once within the repeat. However, these assumptions do not hold true for long reads (Fig. 2.1), likely due to the fact that errors are driven by other processes than PCR. Instead, numerous errors can be present at different positions of the repeat rather than being strictly tied to the repeat unit size (Fig. 2.1), with single base insertions or deletions most prevalent. To allow for a more flexible error model, we used a Hidden Markov Model approach based on that of Dindel²⁵ to align reads to each candidate haplotype.

LongTR models errors in homopolymer repeats using a geometric distribution as follows:

$$p(\text{error size} = l) = \begin{cases} 1 - i - d, & l = 0 \\ i \times \rho(1 - \rho)^{l-1}, & l > 0 \\ d \times \rho(1 - \rho)^{-l-1}, & l < 0 \end{cases}$$

Where ρ controls the size of error, i and d are the probability of error increasing or decreasing the length of repeat respectively. To obtain the values for i , d , and ρ , we used the ALLREADS format field in LongTR output to extract the observed read lengths at each locus and use assembly genotypes as ground truth to compute the base pair differences s with the actual genotype. Then $\rho = \frac{\#InDels}{\sum abs(s)}$, $i = \frac{\#insertion}{\#reads}$, $d = \frac{\#deletions}{\#reads}$ were computed for homopolymers falling in each specific length range.

Phasing information: We leverage HipSTR's existing option to use read phase information from haplotagged reads (and renamed the option from `--10x-bams` to `--phased-bams`) to accurately identify sample haplotypes. To consider phasing information, LongTR requires that at least one read from each haplotype be present, and unphased reads constitute no more than 20% of the total sample's reads.

Genotyping: LongTR iterates over all pairs of haplotypes (H_i, H_j) (including homozygous pairs) and calculates a score for each possible genotype $G(H_i, H_j)$ based on all observed reads R using the following formula:

$$S(G(H_i, H_j)) = \prod_{r \in R} P(h_r = 1) * S_{rH_i} + P(h_r = 2) * S_{rH_j}$$

where S_{rH_i} is the alignment score from aligning read r to haplotype H_i and $P(h_r = 1)$ is the probability that read r is generated from the first haplotype. When read phase information is available, $P(h_r = k)$ is set to 0 or 1 based on the haplotag field. Otherwise we set the probability to come from haplotype 1 vs. 2 as equally likely. The score for each haplotype pair is then normalized by the sum of scores for all possible genotypes and the result is reported as the quality score of the genotype in LongTR output, defined as Q .

Implementation: LongTR is implemented in C++. It leverages the HTSlib²⁶ library to read directly from cloud addresses or URLs, which can avoid the costly step of downloading large sequence alignment files.

2.4.2 Evaluating LongTR

LongTR v1.0 was run using non-default parameters `--min-reads 4`, `--haploid-chrs chrX,chrY`, `--max-tr-len 10000`, `--skip-assembly`, `--indel-flank-len 25`, and `--phased-bam` to genotype repeats with at least 4 overlapping reads, consider chromosome X and chromosome Y as haploid chromosomes, genotype repeats with length up to 10,000bp, skipping assembly of flanking regions, considering InDels up to 25bp around the repeat as InDels affecting the repeat size, and leveraging haplotag information when genotyping, respectively. This set of parameters was used for all the following experiments, unless otherwise stated. To evaluate repeat allele sequences returned by LongTR, the HG002 assembly v1.0.1¹⁰ (Data Availability), with higher accuracy at homopolymers, was mapped to the GRCh38 reference genome using minimap²⁷ v2.26-r1175. For each repeat on autosomal chromosomes, allele sequences from the maternal and paternal haplotypes that completely span the repeat were extracted and compared to allele sequences reported in the output VCF files of LongTR. For Mendelian Consistency analysis on autosomal chromosomes, LongTR was used to perform joint genotyping in all three samples. We only considered a repeat in the Mendelian consistency analysis if 1) all samples were successfully genotyped at that repeat and 2) the genotype for at least one sample was not homozygous for the reference allele. The minimum quality score reported by LongTR among all three samples is considered the assigned score for that repeat.

2.4.3 Benchmarking against TRGT

TRGT v0.5.0 was used to genotype TRs from PacBio HiFi reads using non-default parameters `--karyotype XY`, `--flank-len 25`, and `--max-depth 10000`. For Mendelian Consistency analysis on autosomal chromosomes, TRGT was run separately on each sample. We ran both LongTR and TRGT on the same computer clusters running Rocky Linux 9. HPC hardware specification is as follows: CPU: (Intel(R) Xeon(R)) Platinum 8358 CPU @ 2.60GHz) with 1 TB RAM. All analyses were performed on a single core with 4GB memory. Each experiment was run 5 times and the mean value was reported. Timing was performed with the UNIX time command and the sum of the sys and user times was reported.

2.4.4 Benchmarking against Straglr

We ran Straglr v1.5.0 using the non-default parameters `--min_support 4`, `--genotype_in_size`, `--min_str_len 1`, and `--max_str_len 1000` to genotype repeats with at least 4 overlapping reads, report genotype in terms of allele sizes instead of copy numbers, genotype homopolymers, and genotype repeats with motif length up to 1000bp.

2.4.5 Benchmarking against Tandem-genotypes

Input sequencing data should be aligned using LAST²⁸ prior to TR genotyping with Tandem-genotypes. The aligned BAM file for HG002 was converted to fastq format using samtools²⁶ and then realigned using LAST v1542 to the GRCh38 reference genome as described in the Tandem-genotypes documentation. Tandem-genotypes v1.9.1 was then run using non-default parameters `-o2` and `-u1` to report 2 alleles per repeat and genotype homopolymers.

2.4.6 Benchmarking against adVNTR

For evaluation of adVNTR and LongTR on VNTRs, adVNTR v1.5.0 was run with non-default parameters `--accuracy-filter`, `--pacbio`, and `--log-pacbio-reads`. The reference set of VNTRs

was downloaded from the adVNTR GitHub (see Data Availability). Since adVNTR represents alternative alleles as integer multiples of the consensus repeat unit, direct allele length comparison was not possible. Therefore, we computed the concordance between LongTR and adVNTR copy number estimates allowing for one copy number difference to accommodate the complex VNTRs consisting of multiple motifs with different sequences. This analysis was done on autosomal chromosomes.

2.4.7 Evaluation on ONT reads

Oxford Nanopore Duplex data for HG002 (Data Availability) was aligned to GRCh38 using minimap2 v2.26-r1175. We used WhatsHap²⁹ v2.2 to haplotag sequencing data with small variants called by DeepVariant³⁰ as input (Data Availability). The simplex SUP dataset was already aligned and phased (Data Availability). LongTR was then run on the Adotto reference restricting to autosomal chromosomes.

2.4.8 Comparison to short read STR calls

HipSTR v0.7 was used to genotype STRs from 250bp paired-end PCR-free Illumina reads for HG002 (Data Availability) with non-default parameters --min-reads 4, --def-stutter-model, and --haploid-chrs chrY,chrX to genotype repeats with at least 4 overlapping reads, use default values for the stutter error model, and to consider chromosome Y and X as haploid chromosomes.

2.5 Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Samples HG002, HG003, and HG004 are part of the Personal Genome Project collection and have been consented for commercial use and for public posting of Personally Identifying Genetic Information (PIGI), which allows open-access, public posting of extensive genetic data.

2.6 Availability of data and materials:

LongTR is implemented as a C++ package and is accessible on GitHub (<https://github.com/gymrek-lab/LongTR>³¹) under GNU General Public License v2.0. An archived version of LongTR is available on Zenodo (<https://zenodo.org/doi/10.5281/zenodo.11403979>)³².

Project Adotto Tandem-Repeat Regions and Annotations³³:

https://zenodo.org/record/8329210/files/adotto_repeats.hg38.bed.gz?download=1

PacBio HiFi reads aligned to GRCh38 for HG002, HG003, and HG004³⁴:

<https://downloads.pacbcloud.com/public/revio/2022Q4/>

PacBio HiFi reads for patients with pathogenic expansions and metadata³⁴:

https://downloads.pacbcloud.com/public/dataset/RepeatExpansionDisorders_NoAmp.

HG002 assembly¹⁰:

<https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/HG002/assemblies/hg002v1.0.1.fasta.gz>

adVNTR reference set of repeats¹⁷:

<https://drive.google.com/file/d/1DetpBQySPNe2YAJa4FsJHn9qiRNS3wEV/view>

Oxford Nanopore Duplex data for HG002¹⁹:

<https://human-pangenomics.s3.amazonaws.com/index.html?prefix=submissions/0CB931D5-AE0C-4187-8BD8-B3A9C9BFDAD5-->

[UCSC_HG002_R1041_Duplex_Dorado/Dorado_v0.1.1/stereo_duplex/](https://human-pangenomics.s3.amazonaws.com/index.html?prefix=submissions/0CB931D5-AE0C-4187-8BD8-B3A9C9BFDAD5--UCSC_HG002_R1041_Duplex_Dorado/Dorado_v0.1.1/stereo_duplex/)

Oxford Nanopore aligned and haplotagged Simplex SUP data for HG002³⁵:

<https://42basepairs.com/download/s3/ont-open->

[data/giab_2023.05/analysis/variant_calling/hg002_sup_all/hg002.haplotagged.bam](https://42basepairs.com/download/s3/ont-open-data/giab_2023.05/analysis/variant_calling/hg002_sup_all/hg002.haplotagged.bam)

DeepVariant SNP calls used by WhatsHap³⁴:

<https://downloads.pacbcloud.com/public/revio/2022Q4/HG002->

[rep3/analysis/HG002.m84005_220827_014912_s1.GRCh38.deepvariant.phased.vcf.gz](https://downloads.pacbcloud.com/public/revio/2022Q4/HG002-rep3/analysis/HG002.m84005_220827_014912_s1.GRCh38.deepvariant.phased.vcf.gz)

GIAB set of difficult to map regions²⁰:

<https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome->

[stratifications/v3.3/GRCh38@all/Union/GRCh38_allallowmapandsegdupregions.bed.gz](https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.3/GRCh38@all/Union/GRCh38_allallowmapandsegdupregions.bed.gz)

Illumina reads for HG002³⁶:

<https://ftp->

[trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_Illumina_2x250bps/novoalign_bams/](https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_Illumina_2x250bps/novoalign_bams/)

[ST_Illumina_2x250bps/novoalign_bams/](https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_Illumina_2x250bps/novoalign_bams/)

HipSTR hg38 reference TR set³:

<https://github.com/HipSTR-Tool/HipSTR->

[references/raw/master/human/hg38.hipstr_reference.bed.gz](https://github.com/HipSTR-Tool/HipSTR-references/raw/master/human/hg38.hipstr_reference.bed.gz)

2.7 Acknowledgements

Chapter 2, in full, is a reprint of the material in “LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads” by Helyaneh Ziaei Jam, Justin M. Zook, Sara Javadzadeh, Jonghun Park, Aarushi Sehgal and Melissa Gymrek as it appeared in *Genome Biology* 2024. This work was supported in part by NIH grants 1R01HG010149 (M.G.) and U01DA051234 (M.G.). Author contributions: H.Z.J. developed LongTR, led analyses, and wrote the manuscript.

J.M.Z. helped perform comparison of LongTR and TRGT genotypes to the HG002 assembly and helped supervise ONT analyses. S.J. and J.P. helped with adVNTR analyses. A.S. helped with testing and development of LongTR. M.G. supervised development of LongTR and analyses and wrote the manuscript. Certain commercial equipment, instruments, or materials are identified to specify adequately experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose. We thank Eric Mendenhall for helpful comments on the manuscript.

2.8 Appendix

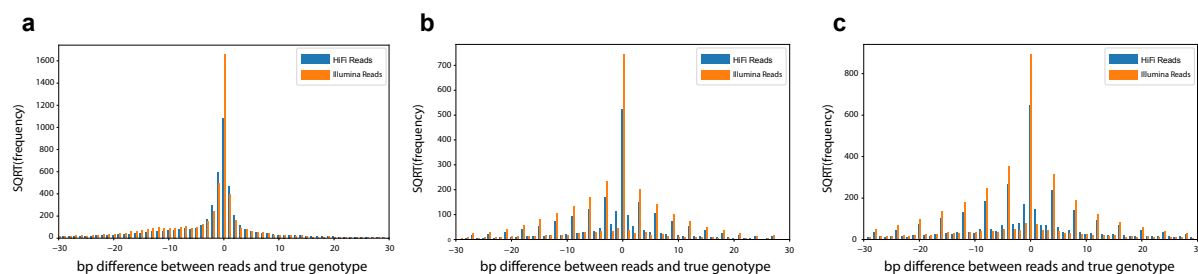


Figure 2.1: Base pair error distribution for PacBio HiFi and Illumina reads. Bars show the

Bars show the square root of the frequency of error sizes for PacBio HiFi (blue) and Illumina (orange) reads from sample HG002. **a.** bp difference with the maximum likelihood genotype for homopolymers with total length between 20-60bp; **b.** bp difference with the maximum likelihood genotype for trinucleotides with total length between 30-100bp; **c.** bp difference with the maximum likelihood genotype for tetranucleotides with total length between 40-100bp. Notably, for **b-c**, PacBio reads show a higher rate of errors that are not a multiple of the repeat unit length.

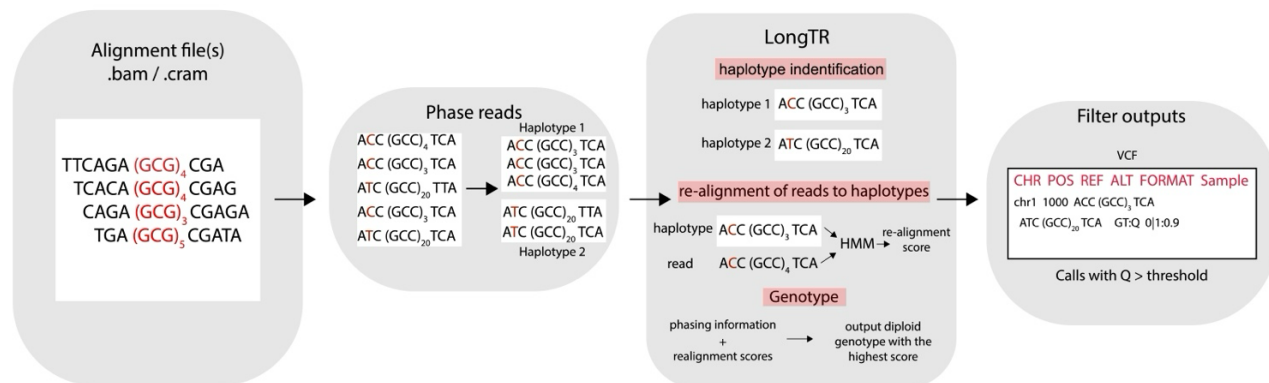


Figure 2.2: LongTR workflow

Users may first optionally haplotag (phase) the input reads. The aligned reads, along with a reference set of TRs, are input to LongTR for genotyping. LongTR uses a clustering strategy combined with partial order alignment to infer consensus haplotypes from error-prone reads, followed by sequence realignment using a Hidden Markov Model to infer the highest scoring genotypes. Finally, it outputs a VCF file with inferred genotypes, quality scores, and other fields that can be used to further filter low quality calls.

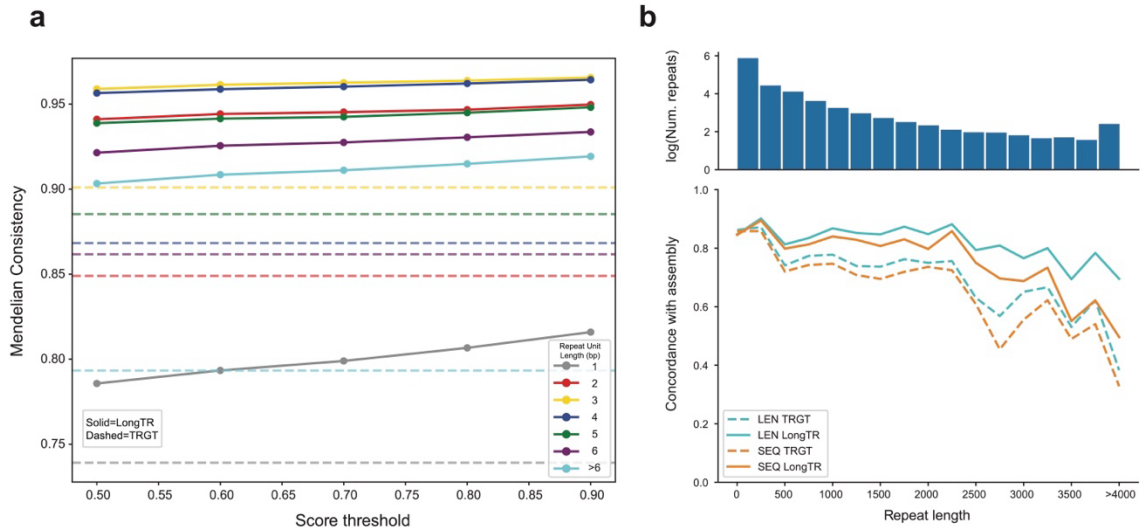


Figure 2.3: Benchmarking LongTR against TRGT

a. Assessing Mendelian consistency of TR calls in an Ashkenazi trio using PacBio HiFi reads. The x-axis gives the LongTR score threshold to include calls and the y-axis gives the percentage of TRs for which genotypes in the trio follow Mendelian consistency. Trio-TR pairs for which all members were called as homozygous for the reference allele were excluded. Dashed=TRGT; solid=LongTR. Note TRGT does not report a quality score and thus a single horizontal line is shown. Color indicates the size of the repeat unit (in bp) considered. **b.** Concordance of TR genotypes obtained from PacBio HiFi with assembly alleles in HG002. TRs were binned by length of the repeat (in bp, bin size = 250bp) in GRCh38. The x-axis shows the TR length and the y-axis shows the percent of alleles that match the assembly. Blue lines show when only length is considered. Orange lines show when both length and sequence are considered. Dashed=TRGT; solid=LongTR. The top panel shows the number of repeats in each bin, on a logarithmic scale.

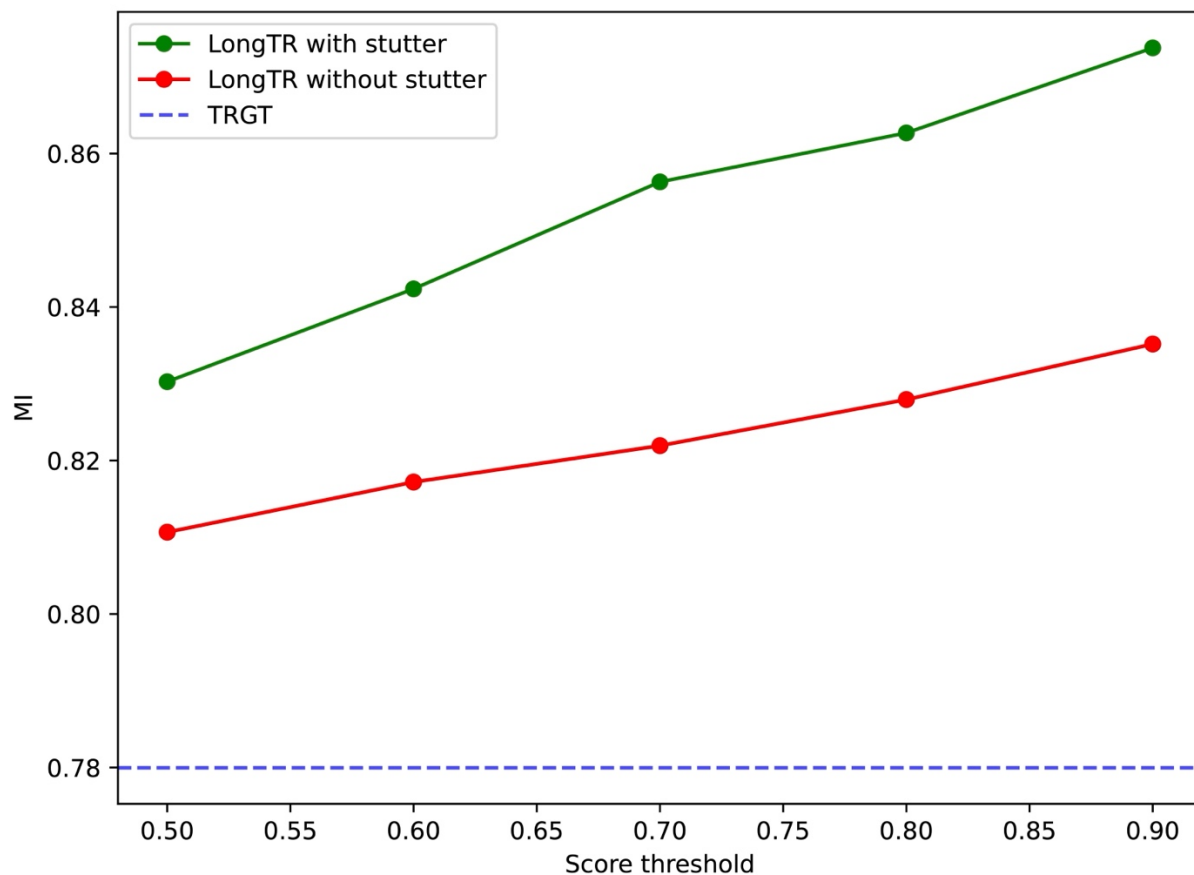


Figure 2.4: Mendelian Inheritance for an Ashkenazi Trio at homopolymer TRs as a function of LongTR quality score.

The x-axis shows LongTR quality score threshold, and the y-axis shows the percent of genotyped loci that follow Mendelian Inheritance. The green line shows LongTR with stutter error modeling, the red line shows LongTR without using stutter error modeling, and the blue dashed line shows TRGT. Each locus was included if all 3 samples passed the score threshold. Loci where all three samples were homozygous for the reference allele were excluded from the analysis.

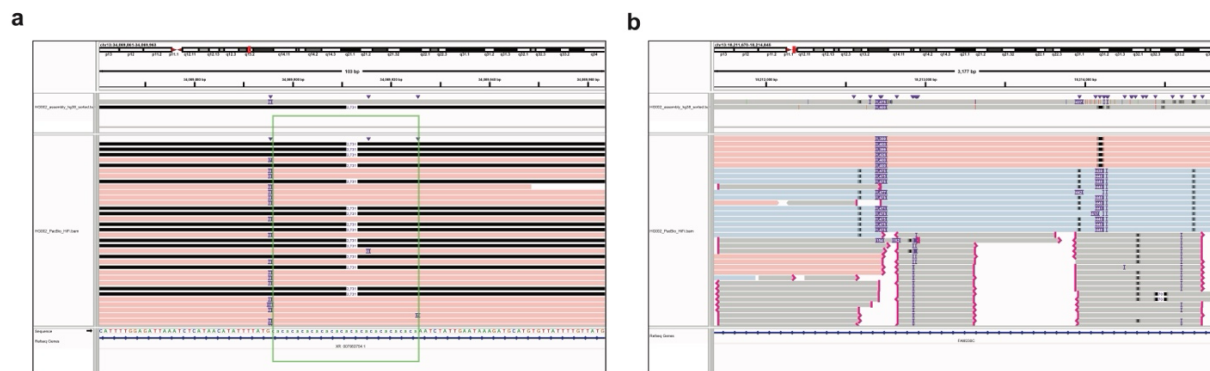


Figure 2.5: IGV screenshots of examples where LongTR matches the assembly but TRGT does not.

The top window shows alignments of the maternal and paternal assemblies and the bottom window shows aligned HiFi reads to GRCh38. Red and blue denote PacBio HiFi reads from the two haplotypes of HG002 based on haplotag information. Gray reads have no haplotag information. The screenshot was generated using the Integrative Genomics Viewer²¹. **a.** LongTR detects structural deletions resulting in null TR alleles. The repeat boundary is denoted by the green box. Note no reads from the haplotype labeled in blue are visible since those reads harbor a 2,731bp deletion at this locus (denoted by the black lines). **b.** The window shows the repeat boundary. TRGT reports deletion in regions with high number of truncated reads.

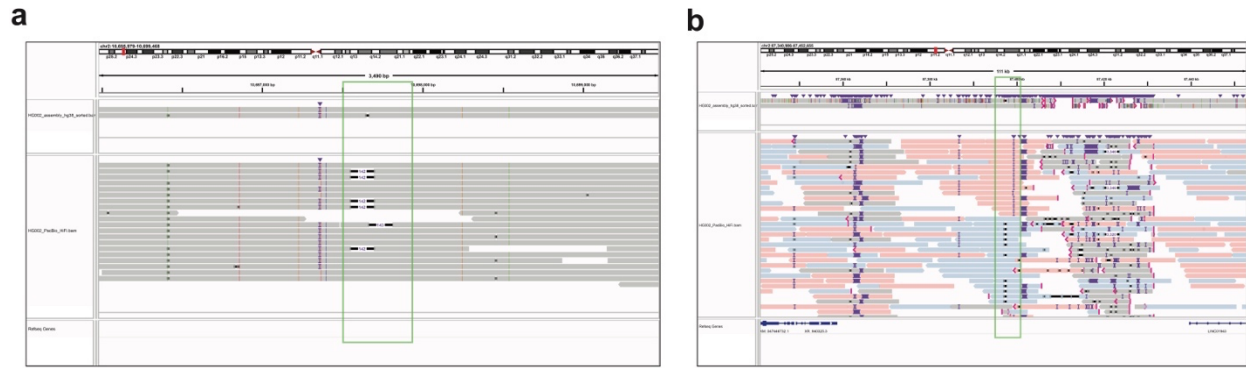


Figure 2.6: IGV screenshots comparing HG002 assembly alleles with PacBio HiFi reads aligned to GRCh38 at loci where TRGT and LongTR agree but do not match the assembly.

The top window shows alignment of the maternal and paternal assemblies and the bottom window shows aligned HiFi reads. Red and blue denote PacBio HiFi reads from the two haplotypes of HG002 based on haplotag information. Gray reads have no haplotag information. The repeat boundary is denoted by the green box. **a.** shows a VNTR with a 141bp repeat unit and total length of 474bp on GRCh38. Example of a highly homozygous region where assembly alleles are most likely incorrect as neither contain the 142bp deletion supported by multiple reads. **b.** shows a VNTR with repeat unit of CCAAGCCAG and total length of 5,180bp on GRCh38 where the insertion on the red haplotype is missing from the assembly due to a nearby highly divergent tandem repeat that causes a break in assembly alignment.

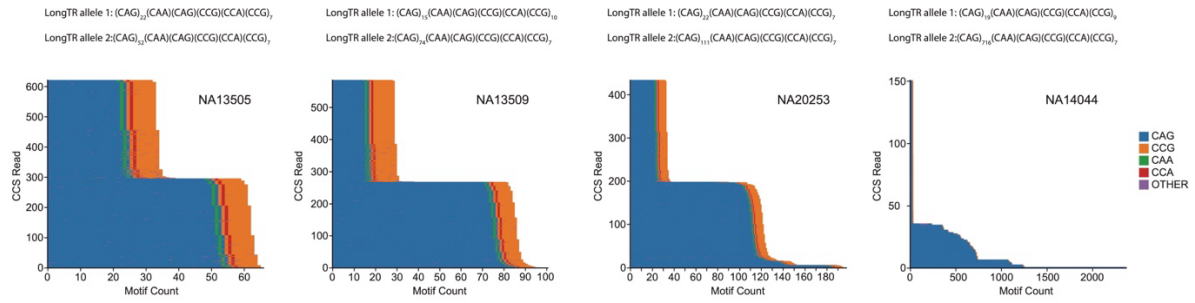


Figure 2.7: Waterfall plots for reads from samples with an HTT expansion genotyped by LongTR.

Each row in each plot represents a single HiFi read. The x-axis shows the count of each motif identified in each read. LongTR precisely identified heterozygous allele sequences for which the repeat unit with the expansion can be inferred. From the left, plots show reads for samples NA13505, NA13509, NA20253, and NA14044. Plots were generated using TRviz¹⁸.

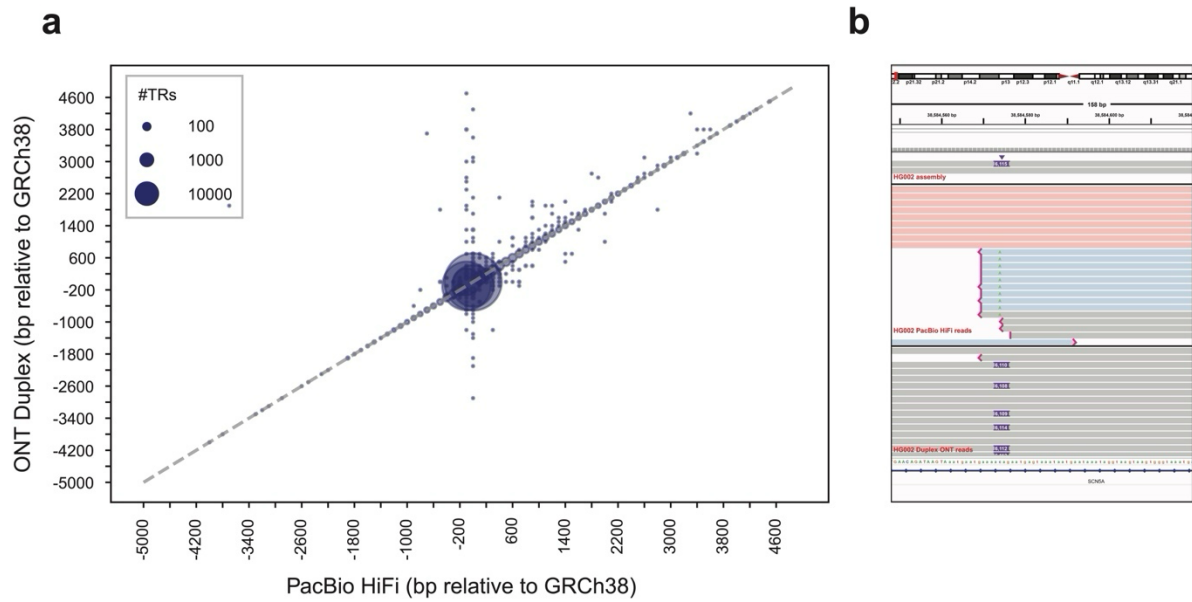


Figure 2.8: ONT vs PacBio

a. Comparison of LongTR genotypes on ONT Duplex vs. PacBio HiFi data. For each call, we computed the average of the length of each allele relative to the GRCh38 reference. The x-axis gives the calls using PacBio data and the y-axis gives the calls using ONT Duplex data. Bubble size scales with the number of calls at each coordinate. **b.** IGV²¹ screenshot comparing PacBio HiFi reads vs. ONT duplex reads at a TR genotyped by LongTR. The top window shows the assembly alignment, the middle window shows aligned HiFi reads, and the bottom window shows aligned ONT duplex reads at a (AGTAAATAATG)_n VNTR. All data is aligned to GRCh38. Red and blue denote PacBio HiFi reads from the two haplotypes of HG002 based on haplotag information. Gray reads have no haplotag information. HiFi reads were clipped at the large repeat insertion, resulting in an incorrect genotype call.

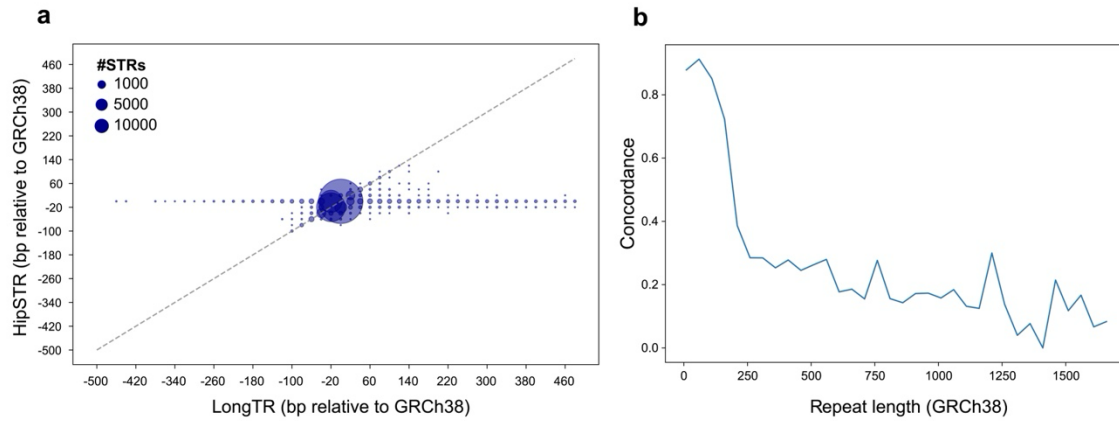


Figure 2.9: Comparison between STR calls by HipSTR using Illumina short reads vs. LongTR on PacBio HiFi reads from sample HG002

a. Length concordance between HipSTR (short reads) vs. LongTR (long reads) genotypes. The x-axis and y-axis show the average base pair difference from the GRCh38 reference genome for the two alleles at each TR. Bubble size shows the number of points at each coordinate. **b. Most discordances are at longer TRs.** The x-axis shows the repeat length on hg38. The y-axis shows the fraction of concordant calls between HipSTR and LongTR.

Table 2-1: Inferred error model parameters at homopolymer TRs in PacBio HiFi reads from HG002.

Parameters u and d indicate the probability to see an expansion or deletion error in each read. The step size of errors (length difference from the true allele in bp) is characterized by a geometric distribution with a parameter 'p'. We computed separate error models for different length ranges of homopolymers. As expected, errors become larger and more frequent with the increase in homopolymer repeat length.

Homopolymer repeat length (bp)	Step-size (p)	Deletion coefficient (d)	Insertion coefficient (u)
10-19	0.92	0.15	0.10
20-29	0.86	0.23	0.13
30-39	0.80	0.28	0.14
40-49	0.75	0.31	0.14
≥ 50	0.71	0.34	0.15

Table 2-2: Benchmarking LongTR against TRGT in patients with pathogenic expansions

Base pair differences from the reference genome at four TR loci were measured by LongTR using PacBio HiFi data. The first four patients have expansions in *HTT*. The next three have expansions in *FMR1*. HEK293 has no known expansion at these loci. None of the samples have expansions in *ATXN10* or *C9orf72*.

Sample	HTT		ATXN10		C9orf72		FMR1	
	TRGT	LongTR	TRGT	LongTR	TRGT	LongTR	TRGT	LongTR
NA12505	9,96	9,96	0,20	0,20	24,42	24,42	30,33	30,33
NA13509	-3,168	-3,168	-10,30	-10,30	12,18	12,18	30,33	30,33
NA20253	9,273	9,276	-5,10	-5,10	-6,-6	-6,-6	0,0	0,0
NA14044	6,1876	6,2115	-10,-10	-10,-10	-6,-6	-6,-6	30,33	30,33
NA13664	-6,0	-6,0	0,0	0,0	-6,-6	-6,-6	30,99	30,99
NA06896	-18,3	-18,3	-5,20	-5,20	-6,30	-6,30	9,509	9,426
NA07537	-18,-6	-18,-6	-10,10	-10,10	-6,-6	-6,-6	27,954	27,1000
HEK293	-6,-3	-6,-3	10,10	10,10	-6,6	-6,6	30,33	30,30

2.9 References

1. H. Ziaei Jam, Y. Li, R. DeVito, N. Mousavi, N. Ma, I. Lujumba, Y. Adam, M. Maksimov, B. Huang, E. Dolzhenko, Y. Qiu, F. E. Kakembo, H. Joseph, B. Onyido, J. Adeyemi, M. Bakhtiari, J. Park, S. Javadzadeh, D. Jjingo, E. Adebiyi, V. Bafna, M. Gymrek, A deep population reference panel of tandem repeat variation. *Nat. Commun.* **14**, 6711 (2023).
2. A. J. Hannan, Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **19**, 286–298 (2018).
3. T. Willems, D. Zielinski, J. Yuan, A. Gordon, M. Gymrek, Y. Erlich, Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods*, doi: 10.1038/nmeth.4267 (2017).
4. S. Kristmundsdottir, H. P. Eggertsson, G. A. Arnadottir, B. V. Halldorsson, popSTR2 enables clinical and population-scale genotyping of microsatellites. *Bioinformatics* **36**, 2269–2271 (2020).
5. N. Mousavi, S. Shleizer-Burko, R. Yanicky, M. Gymrek, Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 (2019).
6. E. Dolzhenko, J. J. F. A. van Vugt, R. J. Shaw, M. A. Bekritsky, M. van Blitterswijk, G. Narzisi, S. S. Ajay, V. Rajan, B. R. Lajoie, N. H. Johnson, Z. Kingsbury, S. J. Humphray, R. D. Schellevis, W. J. Brands, M. Baker, R. Rademakers, M. Kooyman, G. H. P. Tazelaar, M. A. van Es, R. McLaughlin, W. Sproviero, A. Shatunov, A. Jones, A. Al Khleifat, A. Pittman, S. Morgan, O. Hardiman, A. Al-Chalabi, C. Shaw, B. Smith, E. J. Neo, K. Morrison, P. J. Shaw, C. Reeves, L. Winterkorn, N. S. Wexler, US–Venezuela Collaborative Research Group, D. E. Housman, C. W. Ng, A. L. Li, R. J. Taft, L. H. van den Berg, D. R. Bentley, J. H. Veldink, M. A. Eberle, Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).
7. A. M. Wenger, P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Functammasan, A. Kolesnikov, N. D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C.-S. Chin, A. M. Phillippy, M. C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F. J. Sedlazeck, J. M. Zook, H. Li, S. Koren, A. Carroll, D. R. Rank, M. W. Hunkapiller, Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
8. M. Jain, H. E. Olsen, B. Paten, M. Akeson, The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).
9. A. C. English, E. Dolzhenko, H. Ziaei Jam, S. K. McKenzie, N. D. Olson, W. De Coster, J. Park, B. Gu, J. Wagner, M. A. Eberle, M. Gymrek, M. J. P. Chaisson, J. M. Zook, F. J. Sedlazeck, Analysis and benchmarking of small and large genomic variants across tandem repeats. *Nat. Biotechnol.*, doi: 10.1038/s41587-024-02225-z (2024).
10. A. Rhie, S. Nurk, M. Cechova, S. J. Hoyt, D. J. Taylor, N. Altemose, P. W. Hook, S. Koren, M. Rautiainen, I. A. Alexandrov, J. Allen, M. Asri, A. V. Bzikadze, N.-C. Chen, C.-S. Chin,

- M. Diekhans, P. Flicek, G. Formenti, A. Functammasan, C. Garcia Giron, E. Garrison, A. Gershman, J. L. Gerton, P. G. S. Grady, A. Guarracino, L. Haggerty, R. Halabian, N. F. Hansen, R. Harris, G. A. Hartley, W. T. Harvey, M. Haukness, J. Heinz, T. Hourlier, R. M. Hubley, S. E. Hunt, S. Hwang, M. Jain, R. K. Kesharwani, A. P. Lewis, H. Li, G. A. Logsdon, J. K. Lucas, W. Makalowski, C. Markovic, F. J. Martin, A. M. Mc Cartney, R. C. McCoy, J. McDaniel, B. M. McNulty, P. Medvedev, A. Mikheenko, K. M. Munson, T. D. Murphy, H. E. Olsen, N. D. Olson, L. F. Paulin, D. Porubsky, T. Potapova, F. Ryabov, S. L. Salzberg, M. E. G. Sauria, F. J. Sedlazeck, K. Shafin, V. A. Shepelev, A. Shumate, J. M. Storer, L. Surapaneni, A. M. Taravella Oill, F. Thibaud-Nissen, W. Timp, M. Tomaszewicz, M. R. Vollger, B. P. Walenz, A. C. Watwood, M. H. Weissensteiner, A. M. Wenger, M. A. Wilson, S. Zarate, Y. Zhu, J. M. Zook, E. E. Eichler, R. J. O'Neill, M. C. Schatz, K. H. Miga, K. D. Makova, A. M. Phillippy, The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
11. J. Ren, B. Gu, M. J. P. Chaisson, VAMOS: Variable-number tandem repeats annotation using efficient motif sets. *Genome Biol.* **24**, 175 (2023).
 12. D. Bolognini, A. Magi, V. Benes, J. O. Korbel, T. Rausch, TRiCoLoR: tandem repeat profiling using whole-genome long-read sequencing data. *Gigascience* **9** (2020).
 13. S. Mitsuhashi, M. C. Frith, T. Mizuguchi, S. Miyatake, T. Toyota, H. Adachi, Y. Oma, Y. Kino, H. Mitsuhashi, N. Matsumoto, Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol.* **20**, 58 (2019).
 14. C. Readman, R.-B. Indhu-Shree, F. Jan M, B. Inanc, Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol.* **22**, 224 (2021).
 15. E. Dolzhenko, A. English, H. Dashnow, G. De Sena Brandine, T. Mokveld, W. J. Rowell, C. Karniski, Z. Kronenberg, M. C. Danzi, W. A. Cheung, C. Bi, E. Farrow, A. Wenger, K. P. Chua, V. Martínez-Cerdeño, T. D. Bartley, P. Jin, D. L. Nelson, S. Zuchner, T. Pastinen, A. R. Quinlan, F. J. Sedlazeck, M. A. Eberle, Characterization and visualization of tandem repeats at genome scale. *Nat. Biotechnol.*, doi: 10.1038/s41587-023-02057-3 (2024).
 16. S. Nurk, B. P. Walenz, A. Rhie, M. R. Vollger, G. A. Logsdon, R. Grothe, K. H. Miga, E. E. Eichler, A. M. Phillippy, S. Koren, HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
 17. M. Bakhtiari, J. Park, Y.-C. Ding, S. Shleizer-Burko, S. L. Neuhausen, B. V. Halldórsson, K. Stefánsson, M. Gymrek, V. Bafna, Variable number tandem repeats mediate the expression of proximal genes. *Nat. Commun.* **12**, 2075 (2021).
 18. J. Park, E. Kaufman, P. N. Valdmanis, V. Bafna, TRviz: a Python library for decomposing and visualizing tandem repeat sequences. *Bioinform. Adv.* **3**, vbad058 (2023).
 19. S. Koren, Z. Bao, A. Guarracino, S. Ou, S. Goodwin, K. M. Jenike, J. Lucas, B. McNulty, J. Park, M. Rautiainen, A. Rhie, D. Roelofs, H. Schneiders, I. Vrijenhoek, K. Nijbroek, D.

- Ware, M. C. Schatz, E. Garrison, S. Huang, W. R. McCombie, K. H. Miga, A. H. J. Wittenberg, A. M. Phillippy, Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. *bioRxiv*org, doi: 10.1101/2024.03.15.585294 (2024).
20. N. D. Olson, J. Wagner, J. McDaniel, S. H. Stephens, S. T. Westreich, A. G. Prasanna, E. Johanson, E. Boja, E. J. Maier, O. Serang, D. Jáspez, J. M. Lorenzo-Salazar, A. Muñoz-Barrera, L. A. Rubio-Rodríguez, C. Flores, K. Kyriakidis, A. Malousi, K. Shafin, T. Pesout, M. Jain, B. Paten, P.-C. Chang, A. Kolesnikov, M. Nattestad, G. Baid, S. Goel, H. Yang, A. Carroll, R. Eveleigh, M. Bourgey, G. Bourque, G. Li, C. Ma, L. Tang, Y. Du, S. Zhang, J. Morata, R. Tonda, G. Parra, J.-R. Trotta, C. Brueffer, S. Demirkaya-Budak, D. Kabakci-Zorlu, D. Turgut, Ö. Kalay, G. Budak, K. Narıcı, E. Arslan, R. Brown, I. J. Johnson, A. Dolgoborodov, V. Semenyuk, A. Jain, H. S. Tetikol, V. Jain, M. Ruehle, B. Lajoie, C. Roddey, S. Catreux, R. Mehio, M. U. Ahsan, Q. Liu, K. Wang, S. M. E. Sahraeian, L. T. Fang, M. Mohiyuddin, C. Hung, C. Jain, H. Feng, Z. Li, L. Chen, F. J. Sedlazeck, J. M. Zook, PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell Genom.* **2** (2022).
 21. IGV: Integrative genomics viewer. <https://www.igv.org/>.
 22. Oxford Nanopore technologies, *Oxford Nanopore Technologies*. <https://nanoporetech.com/platform/accuracy>.
 23. PacBio revio, *PacBio* (2022). <https://www.pacb.com/revio/>.
 24. C. Lee, Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics* **19**, 999–1008 (2003).
 25. C. A. Albers, G. Lunter, D. G. MacArthur, G. McVean, W. H. Ouwehand, R. Durbin, Dindel: accurate indel calls from short-read data. *Genome Res.* **21**, 961–973 (2011).
 26. J. K. Bonfield, J. Marshall, P. Danecek, H. Li, V. Ohan, A. Whitwham, T. Keane, R. M. Davies, HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience* **10** (2021).
 27. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
 28. S. M. Kielbasa, R. Wan, K. Sato, P. Horton, M. C. Frith, Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
 29. M. Martin, P. Ebert, T. Marschall, Read-based phasing and analysis of phased variants with WhatsHap. *Methods Mol. Biol.* **2590**, 127–138 (2023).
 30. R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, S. S. Gross, L. Dorfman, C. Y. McLean, M. A. DePristo, A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).

31. Ziaei Jam H, Zook JM, Javadzadeh S, Park J, Sehgal A, Gymrek M, LongTR, *GitHub* (2023). <https://github.com/gymrek-lab/LongTR>.
32. Ziaei Jam H, Zook JM, Javadzadeh S, Park J, Sehgal A, Gymrek M, LongTR, *Zenodo* (2024). <https://zenodo.org/doi/10.5281/zenodo.11403979>.
33. A. English, Project adotto tandem-repeat regions and annotations, *Zenodo* (2022); <https://doi.org/10.5281/ZENODO.6930201>.
34. Datasets - PacBio - Highly accurate long-read sequencing, *PacBio* (2020). <https://www.pacb.com/connect/datasets/>.
35. Oxford Nanopore Technologies, Sequencing Genome in a Bottle samples, *Zenodo* (2023); <https://doi.org/10.5281/ZENODO.8363974>.
36. J. M. Zook, D. Catoe, J. McDaniel, L. Vang, N. Spies, A. Sidow, Z. Weng, Y. Liu, C. E. Mason, N. Alexander, E. Henaff, A. B. R. McIntyre, D. Chandramohan, F. Chen, E. Jaeger, A. Moshrefi, K. Pham, W. Stedman, T. Liang, M. Saghbini, Z. Dzakula, A. Hastie, H. Cao, G. Deikus, E. Schadt, R. Sebra, A. Bashir, R. M. Truty, C. C. Chang, N. Gulbahce, K. Zhao, S. Ghosh, F. Hyland, Y. Fu, M. Chaisson, C. Xiao, J. Trow, S. T. Sherry, A. W. Zaranek, M. Ball, J. Bobe, P. Estep, G. M. Church, P. Marks, S. Kyriazopoulou-Panagiotopoulou, G. X. Y. Zheng, M. Schnall-Levin, H. S. Ordonez, P. A. Mudivarti, K. Giorda, Y. Sheng, K. B. Rypdal, M. Salit, Extensive sequencing of seven human genomes to characterize benchmark reference materials, *Scientific data*. **3** (2016)p. 160025.

3 Chapter 3: Improving polygenic risk score prediction in multi-ancestry cohorts

Abstract: Polygenic risk score portability across populations has long been a major challenge, largely due to differences in linkage disequilibrium structure, allele frequencies, and genetic architecture across ancestries. We propose two main frameworks for polygenic risk score (PRS) prediction in multi-ancestry settings: a novel multi-ancestry approach using our new method, imPRS, and a pooled ancestry approach using PRS-CS on a combined GWAS of multiple ancestries. imPRS extends existing Bayesian models by introducing a correlation parameter for shared variants across ancestries and jointly updates the posterior effect sizes simultaneously for all ancestries. We compared imPRS, along with other multi-ancestry methods (PRS-CSx, SDPRX), against the pooled PRS-CS approach. In simulations, we found that the pooled GWAS approach yielded comparable PRS performance for the majority European ancestry samples when causal variants were shared across populations, but its performance declined for the underrepresented African ancestry samples, particularly under population-specific genetic architectures. For real data, we trained PRS models on All of Us dataset and tested on the UK Biobank across 6 quantitative biomarker traits in individuals of African and European ancestries. Our results show that for most traits examined, pooled PRS improves prediction in the target populations by 5–46% relative to multi-ancestry methods, with the largest gains observed in samples of African ancestry. White blood cell count is an exception, where ancestry-specific approaches achieve higher accuracy by 4-9%. This likely reflects its highly ancestry-specific genetic architecture, driven in large part by a single locus with a large effect.

3.1 Introduction

The use of PRS represents a shift in risk prediction and disease classification, offering the potential for more personalized medical care in which genetic risk information can inform clinical decision-making. However, this potential is limited by substantial differences in PRS predictive performance across populations. Most Genome-Wide Association Studies (GWAS), which precede PRS construction, have been conducted primarily in individuals of European ancestry. As a result, PRS developed from these data tend to perform less well in individuals from other ancestries, reducing their generalizability and risking the reinforcement of existing health disparities.^{1,2}

To address this challenge, the field has increasingly focused on developing methods that can leverage GWAS summary statistics from multiple, diverse populations. Initial efforts to simply meta-analyze multi-ancestry GWAS often failed to account for heterogeneity in linkage disequilibrium (LD) structure and differences in allele frequencies and causal effect sizes between populations^{3,4}. More sophisticated PRS approaches have emerged, such as PRS-CSx⁵ and SDPRX⁶, BridgePRS⁷, and JointPRS⁸, which jointly model GWAS summary statistics from multiple ancestries while using ancestry-specific LD reference panels. These methods estimate population-specific posterior effect sizes, borrowing strength across populations to improve statistical power.

Importantly, many complex traits are governed by largely shared causal variants across populations, with effect sizes that are correlated but not identical due to differences in LD structure, allele frequencies, and environmental context.^{9,10} Explicitly modeling this cross-population correlation is therefore critical for efficient information sharing, particularly when GWAS sample sizes are highly imbalanced across ancestries. PRS-CSx has demonstrated strong empirical

performance by borrowing strength across populations through shared variant-specific shrinkage parameters, but it does not explicitly model correlations in effect sizes among shared variants across ancestries. While other methods, such as SDPRX and XPASS¹¹, do attempt to explicitly model cross-population effect-size correlations, these approaches have shown variable or reduced performance across traits and are restricted to modeling only two populations⁸, motivating the development of multi-ancestry models that directly incorporate cross-population effect-size correlations within a unified and flexible Bayesian framework.

Moreover, despite the improvements achieved through these multi-ancestry methods, a shared limitation is their reliance on discrete, predefined ancestry labels. These labels, whether based on self-report or genetic clustering, often fail to adequately represent the complex and continuous nature of human genetic variation. The reliance on discrete ancestry classifications is particularly problematic for admixed populations, where individuals possess a mosaic of ancestries, and for underrepresented populations, where small GWAS sample sizes can lead to less power in finding significant genetic associations within that cohort. Furthermore, cohorts labeled as a single superpopulation for tuning and evaluation often encompass substantial genetic heterogeneity, contributing to variability in predictive accuracy.^{12,13} Recent work has shown that, despite limitations related to population heterogeneity, pooling GWAS data across ancestries can increase statistical power for association discovery, particularly for traits with similar genetic architectures¹⁴.

Building on these observations, we investigate two approaches to build PRS in this manuscript. First, we evaluate a label-free framework in which GWAS summary statistics are generated from pooled multi-ancestry cohorts and used for label-free PRS construction. Second, we propose a multi-ancestry extension of PRS-CS, imPRS (integrated multi-population PRS),

which incorporates an inter-population correlation parameter and a shared variant-level shrinkage parameter to jointly infer posterior effect sizes across populations within a single Bayesian model. We evaluate these two approaches using simulated data for traits with a range of shared and population specific genetic architectures. Finally, we construct PRSs for six quantitative traits trained in the multi-ancestry All of Us (AoU) cohort and tested in individuals of European and African ancestry from UK Biobank. Overall, our results demonstrate that the pooled approach achieves similar or better performance across most scenarios tested compared to PRS methods that explicitly consider ancestry.

3.2 Results

3.2.1 Overview of PRS analysis

In this study, we introduce a novel multi-ancestry PRS method (imPRS) and evaluate two general frameworks for computing PRS (Figure. 3.1a):

(1) In the multi-ancestry framework, we applied three multi-ancestry PRS methods—imPRS, PRS-CS⁵, and SDPRX⁶—to GWAS results obtained from separate cohorts for each ancestry considered. Each method outputs posterior effect sizes per ancestry, which were further used to compute ancestry-specific PRS.

(2) In the pooled framework, samples from all ancestries in the training dataset were combined to perform a single GWAS. The resulting summary statistics were used as input to PRS-CS¹⁵, a single-ancestry PRS method, to generate one set of posterior effect sizes, which were applied to all ancestries in the test dataset to compute the ancestry-specific PRS.

All PRS methods evaluated in this study operated on summary statistics and did not require individual-level genotype or phenotype data. For analyses on real data, linkage disequilibrium (LD) reference panels were derived from Phase 3 of the 1000 Genomes Project (1KG) for each PRS method.

3.2.2 Overview of imPRS

imPRS employs a Bayesian framework similar to the one incorporated in PRS-CS to model and predict polygenic risk scores in multi-ancestry settings. To share information across populations, imPRS applies a common local shrinkage factor to variants shared among ancestries and incorporates a variance-covariance matrix that encourages similar posterior effect size estimates across populations through correlation parameter. imPRS relies only on the GWAS summary statistics and LD information to predict the posterior effect sizes. Notably, in each iteration of the posterior inference algorithm (Gibbs sampling), imPRS jointly updates model parameters across all ancestries within a single integrated model, rather than fitting ancestry-specific models independently. Posterior beta estimates are then produced for each ancestry (Figure 3.1b) (see Methods for detailed model description and comparison to other methods).

3.2.3 Simulations

We used simulated genotype data generated by [6], comprising 3,000 HapMap3¹⁶ variants (minor allele frequency > 0.05 in East Asians (EAS), Europeans (EUR), and Africans (AFR)) on chromosome 10. The training dataset in the multi-ancestry setting included 40,000 samples of European ancestry and 10,000 samples of African ancestry, reflecting real-world imbalances in population representation. For the pooled setting, we pooled all 50,000 samples from both ancestries into one single cohort. The validation dataset contained 5,000 samples per ancestry and was used to generate LD reference panels and to tune hyperparameters. For the pooled analysis,

we combined all 10,000 samples in the validation dataset to generate the pooled LD reference panel. The test dataset contained 5,000 samples per ancestry. Samples in the training, validation, and test datasets were non-overlapping.

We randomly selected 0.5% of variants as causal, explaining either 10% or 30% of the phenotypic variance. Three architectures were simulated: (1) all causal variants shared across ancestries, (2) 20% of causal variants specific to the European population, and (3) 20% of the causal variants specific to the African population. In the latter two settings, causal variants were either absent in the other population (defined here as ancestry-specific variants, ASVs) or present but non-causal (defined here as ancestry-specific effects, ASEs). Per-population effect sizes were modeled jointly under a multivariate normal distribution with a specified cross-population correlation structure (Methods). We ran imPRS with $s \in \{0, 0.2, 0.4, 0.6, 0.8\}$ and selected the optimal value for each scenario using the validation dataset.

We next compared the performance of all methods tested here as measured by the variance explained R^2 between the predicted phenotype and the simulated phenotype. Across all scenarios, imPRS either matched or exceeded the performance of other methods. In all scenarios, PRS-CS performed worse in the African samples, where the performance drop is the most visible in the presence of ASE variants in either population. This is probably because 1) in the pooled GWAS, marginal effect sizes are highly influenced by the values in European samples, due to the larger sample size, and 2) LD matrix in the pooled setting was calculated using LD boundaries from European population (Figure 3.2, Supplementary Figure 3.4).

We next tested 3 populations scenario with 40,000, 10,000, and 10,000 samples from European, African, and East Asian populations, respectively. Similar to the two-populations

setting, we randomly selected 0.5% of variants as causal, explaining either 10% or 30% of the phenotypic variance. Causal variants are shared across all populations, and their effect sizes were modeled using a multivariate normal distribution with cross populations effect sizes. PRS-CS, imPRS, and PRS-CSx were used for comparison, as SDPRX is only designed for the two-populations setting. Once again, we observed superior performance from imPRS for all populations, and the performance drop of the pooled method was most visible for African samples. (Figure 3.5).

3.2.4 Real Data analysis

We used the All of Us cohort as our discovery dataset, and we performed GWAS analysis for 6 traits: HDL cholesterol, platelet count, white blood cell count, red blood cell count, glycated haemoglobin, and triglycerides. Sample sizes per trait and per ancestry are provided in Supplementary Table 1. For each trait, we conducted three GWAS analyses using individuals who passed quality control: (1) European ancestry samples, (2) African ancestry samples, and (3) all samples combined across ancestries (see Methods, Supplementary Table 1). Manhattan and Q-Q plots for all GWAS results are shown in Figures 3.6-3.11.

Next, we evaluated the predictive performance of different polygenic prediction methods using the same 6 traits from UKBB (Figure 3.3). Samples in the UKBB were filtered as described in Margoliash et. al.¹⁷, leading to 452,453 unrelated, QCed samples in total. We assigned each target sample to one of the six self-reported ethnicities summarized by UKB showcase data field 21,000, White British, Black, Chinese, Irish, South Asian, and Other White (Methods). We used the samples from White British and Black populations to assess the prediction accuracy, measured by variance explained R^2 in linear regression, separately, adjusting for age, sex, and top 20 principal components (PCs) of the genotypes.

For 5 out of 6 traits, per-trait predictive performance of all methods was consistently higher in European samples, likely reflecting the substantially larger sample sizes available for EUR-dominated GWAS, which improve effect-size estimation and downstream PRS accuracy. We observed that PRS-CS matched or outperformed other multi-ancestry PRS methods across both AFR and EUR populations in most traits. A notable exception to both observations was white blood cell count, where pooled approach showed degraded performance in both EUR- and AFR-ancestry individuals, and in general, we had better prediction power in African samples. This pattern is consistent with the well-established African-specific effect at the Duffy (ACKR1) locus on chromosome 1, where the near-fixation of the Duffy-null allele in African populations drives large ancestry-specific effects on leukocyte counts; pooling across ancestries or enforcing shared effects can potentially dilute this signal and harm prediction. imPRS achieved the best performance for white blood cell count. This pattern suggests that while pooled approach is generally well-suited for most traits, explicitly modeling population-specific signals while allowing for shared effect sizes—as done by imPRS—provides a clear advantage in scenarios with strong ancestry-specific effects.

3.3 Methods

3.3.1 Overview of imPRS

We introduce a joint Bayesian framework for polygenic risk score (PRS) construction across multiple populations that extends PRS-CS by modeling genetic effect sizes across populations within a single hierarchical model. This approach explicitly captures cross-population correlation in effect sizes while allowing population-specific variation due to differences in linkage disequilibrium (LD), allele frequencies, sample sizes, and environmental effects.

We consider k populations indexed by $i = 1, \dots, k$. For each population, phenotypes are modeled using a linear regression framework. Stacking all populations yields:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_k \end{bmatrix}$$

Where $y_i \in \mathbb{R}^{N_i}$ is the phenotype vector, $X_i \in \mathbb{R}^{N_i \times M_i}$ is the standardized genotype matrix, and $\beta_i \in \mathbb{R}^{M_i}$ denotes population-specific genetic effect sizes. Residual errors are assumed independent across populations:

$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2 I_{N_i})$$

Stacking all populations yields a block-diagonal regression with:

$$\epsilon \sim \mathcal{N}(0, \bar{\Sigma}), \quad \bar{\Sigma} = \text{diag}(\sigma_1^2 I_{N_1}, \dots, \sigma_k^2 I_{N_k})$$

We assume no residual correlation across populations. Inference is performed using GWAS summary statistics:

$$\hat{\beta}_i = \frac{1}{N_i} X_i^\top y_i, \quad D_i = \frac{1}{N_i} X_i^\top X_i$$

Where D_i is the LD matrix for population i , estimated from an ancestry-matched reference panel and $\hat{\beta}_i$ is the marginal effect estimates for population i . We define the stacked parameters as:

$$\hat{\beta} = \bar{N}^{-1} X^\top y, \quad D = \bar{N}^{-1} X^\top X$$

With $\bar{N} = \text{diag}(N_1 I_{M_1}, \dots, N_k I_{M_k})$.

For each variant j shared across populations, we model the vector of population-specific effects

$$\beta_j = (\beta_{1j}, \dots, \beta_{kj})^\top$$

using a multivariate normal prior $\beta_j \sim \mathcal{N}(0, \psi_j \Sigma_N)$. Where ψ_j is a variant-specific shrinkage parameter, and Σ_N captures cross-population covariance:

$$\Sigma_N(i, i) = \frac{\sigma_i^2}{N_i}, \quad \Sigma_N(i, \ell) = \frac{\sigma_i \sigma_\ell s_{i\ell}}{\sqrt{N_i N_\ell}}, \quad i \neq \ell$$

Here, $s_{i\ell} \in [0, 1]$ is a sharing parameter controlling genetic correlation between populations i and ℓ for the shared variants j . This structure allows effect sizes to be correlated but not identical across populations, reflecting shared causal variants with population-specific modulation. Variants present in only a subset of populations are naturally accommodated by omitting the corresponding entries in β_j .

Following PRS-CS, we impose continuous shrinkage priors:

$$\psi_j \sim \text{Gamma}(a, \delta_j), \quad \delta_j \sim \text{Gamma}(b, \phi)$$

Where ψ_j controls local (variant-specific) shrinkage, ϕ controls global sparsity, and a, b are fixed hyperparameters. Under this formulation, the posterior distribution of the full effect-size vector β is multivariate normal, $\beta \mid \hat{\beta}, D \sim \mathcal{N}(\mu_n, \Sigma_n)$, with

$$\Sigma_n = \bar{N}^{-1} \bar{\Sigma} (D + T^{-1})^{-1}, \mu_n = \Sigma_n \bar{\Sigma}^{-1} \bar{N} \hat{\beta}$$

where T is a block-diagonal matrix encoding the local shrinkage parameters and cross-population covariance for each variant. Posterior inference is performed using Gibbs sampling.

3.3.2 Overview of Existing PRS Methods

3.3.2.1 *PRS-CS*

PRS-CS is a single-ancestry Bayesian polygenic prediction method that models SNP effect sizes under continuous shrinkage priors. It assumes that true SNP effect sizes follow a hierarchical prior that adaptively shrinks small, likely non-causal effects toward zero while allowing large effects to remain relatively unshrunk. The model uses GWAS summary statistics together with an external linkage disequilibrium (LD) reference panel to jointly model SNP correlations and infer posterior effect sizes. PRS-CS employs a Gibbs sampling algorithm for posterior inference, which iteratively updates SNP effect estimates while accounting for LD.

3.3.2.2 *PRS-CSx*

PRS-CSx extends PRS-CS to a multi-ancestry setting by inferring posterior effect sizes for each population separately while sharing the same local shrinkage parameter across variants present in multiple populations. This design enables information sharing across ancestries during inference while preserving population-specific signals. PRS-CSx produces one set of posterior effect sizes per population, which can subsequently be linearly combined using weights derived from a validation dataset to generate ancestry-specific PRS for the trait of interest.

3.3.2.3 *SDPRX*

SDPRX is a multi-ancestry PRS tool that employs a hierarchical Bayesian model to jointly integrate GWAS summary statistics and LD information from two populations. The model partitions SNPs into four categories to represent different genetic architectures: null (no effect) in both populations, causal in only one of the populations, or causal and shared across populations. The shared component is central to the framework and is modeled using a mixture of bivariate

Gaussian distributions parameterized by a pre-estimated genetic correlation, allowing the model to capture the shared genetic architecture between populations.

3.3.3 Simulation setting

We sampled effect sizes from a multivariate normal distribution as follows:

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1k}\sigma_1\sigma_k \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2k}\sigma_2\sigma_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1}\sigma_k\sigma_1 & \rho_{k2}\sigma_k\sigma_2 & \cdots & \sigma_k^2 \end{pmatrix} \right]$$

Where $\mu_i = 0$ and $\sigma_i = \frac{h^2}{m_i}$. h^2 is the heritability of the trait and m_i is the number of causal variants in the i 'th population. We set the correlation parameter $\rho = 0.98$ to reflect high similarity between genetic settings across populations.

Phenotypes were simulated using GCTA¹⁸, setting --simu-hsq according to the heritability defined for each scenario. GWAS was conducted using PLINK2¹⁹ on the simulated genotypes and phenotypes, with the top 20 PCs included as covariates to adjust for population structure. In the pooled GWAS setting, PCs were derived from the pooled genotype dataset from all ancestries included in the analysis. PCs were computed using PLINK²⁰.

PRS-CSx and SDPRX were run using the LD blocks available by SDPRX. For imPRS, LD matrices for African and East Asian samples were recalculated based on European LD block boundaries. For PRS-CS, LD was recalculated on the pooled samples using European boundaries. LD calculation was performed using 5,000 validation samples from each population.

3.3.4 Building PRS models

PRS-CS, PRS-CSx, and imPRS were all run with global shrinkage factor ϕ set to 1×10^{-05} . Number of iterations for the MCMC was set to 1000, with 400 iterations as burn-in. SDPRX was run using default values.

3.3.5 Real Data Analysis

GWAS analysis in the All of Us cohort

Samples in the All of Us cohort were filtered and assigned ancestry as described in the original study²¹, resulting in 119,184 samples from European White ancestry, 44,386 samples from African black ancestry, and 225,116 samples in the pooled cohort. Phenotypes were preprocessed as described in Rivera et. al.²² Association analyses were conducted using Hail 0.2²³. Each model included age, sex, and the top ten ancestry principal components computed on all samples combined (provided by the All of Us Research Program) as covariates. Phenotypes were rank-based inverse normal transformed prior to the GWAS analysis.

Before performing association testing, the ACAF dataset was filtered to retain only high-quality genotypes. We excluded genotypes that did not have the filter field set to “PASS” or had a genotype quality below 20. Samples with a call rate below 0.9 were removed, as were variants that (1) had more than two alleles, (2) showed a call rate < 0.9 , (3) had a minor allele frequency (MAF) ≤ 0.01 , or (4) showed extreme deviation from Hardy–Weinberg equilibrium ($p \leq 1 \times 10^{-100}$). Association testing was performed using Hail’s `linear_regression_rows` function.

For the pooled GWAS, we first ran the analysis with `--MAF 0.001` and `--HWE -1`, and then filtered the variants to those that passed minor allele frequency and Hardy-Weinberg equilibrium filters in at least one of the African or European cohorts.

Building PRS on AoU data

imPRS, PRS-CSx, and SDPRX were applied to ancestry-specific GWAS summary statistics derived from European White and African Black samples in the AoU cohort. For imPRS, we tested three values of the hyperparameter $s \in (0.0, 0.2, 0.4)$, as informed by simulation results.

PRS-CS was applied to the pooled GWAS using LD reference panels and allele frequencies constructed from EUR, AFR, and AMR superpopulations in the 1000 Genomes Project, reflecting the primary ancestries represented in the All of Us dataset. For imPRS, PRS-CS, and PRS-CSx, the global shrinkage parameter was set to 10^{-5} . Each model was run for 1,000 MCMC iterations, with the first 400 iterations as burn-in.

Testing on UKB samples

We used the preprocessed UK Biobank genotype phenotype data as described in the previous study¹⁷. For each trait, additional trait-specific filters—such as restricting to unrelated individuals with non-missing phenotype values—were applied, which resulted in smaller sample sizes per trait. Phenotype values were further normalized as described in the original study. The top 40 PCs used for adjusting R^2 were provided by the UKB team (field 22009).

3.4 Discussion

In this study, we systematically evaluated label-free pooled GWAS and PRS construction alongside multi-ancestry methods across both simulated and real-world data. Our results demonstrate that, in most of tested phenotypes, pooling samples across ancestries and applying PRS-CS yields predictive performance that matches or exceeds that of more complex multi-ancestry approaches. In real-data analyses using All of Us (AoU) for training and UK Biobank

(UKB) for evaluation, the pooled PRS-CS framework consistently performed well across both African- and European-ancestry target populations for most traits, highlighting the benefits of increased effective sample size for polygenic prediction.

Beyond predictive accuracy, a key advantage of the pooled, label-free approach lies in its practical simplicity. Conducting GWAS and PRS analyses in admixed and ancestrally diverse cohorts presents substantial methodological and computational challenges, including the need for population assignment, ancestry-specific principal component analysis, LD estimation, and repeated model fitting across groups. Population-stratified analyses further require duplicated GWAS, LD reference construction, and PRS computation, leading to increased computational cost and complexity. By avoiding population labels altogether, the pooled framework circumvents these challenges, offering a streamlined alternative that reduces analytical overhead and improves scalability—an advantage that becomes increasingly important as biobanks grow in size and diversity.

However, our simulation and real-world analysis also demonstrate that when ancestry-specific causal variants or effect sizes are present, the performance of pooled models deteriorates, indicating an inherent limitation of ignoring cross-population heterogeneity. As an example, our real-data analyses reveal that traits such as white blood cell count with strong population-specific genetic effects can challenge pooled approaches, motivating the need for methods that explicitly model ancestry-dependent architectures in such cases.

Together, these findings suggest that pooled, label-free GWAS and PRS construction provide a powerful and pragmatic baseline for multi-ancestry genetic prediction, particularly in large, diverse cohorts where population assignment is ambiguous and computational efficiency is paramount. At the same time, they highlight the continued importance of developing flexible

models that can accommodate ancestry-specific effects when the genetic architecture demands it. Balancing these considerations will be essential for the equitable and effective application of polygenic risk scores in increasingly diverse populations.

3.5 Code Availability

imPRS is available on GitHub:

https://github.com/heliziii/PRS_analysis_new/tree/master/imPRS

AoU GWAS workflow for this project is available on GitHub: <https://github.com/cast-genomics/cast-workflows>

3.6 Author Contributions

H.Z.J. developed imPRS, led analyses, and wrote the manuscript. T.M developed the GWAS pipeline in AoU and helped with the analyses. M.G. supervised development of imPRS and analyses and wrote the manuscript.

3.7 Acknowledgements

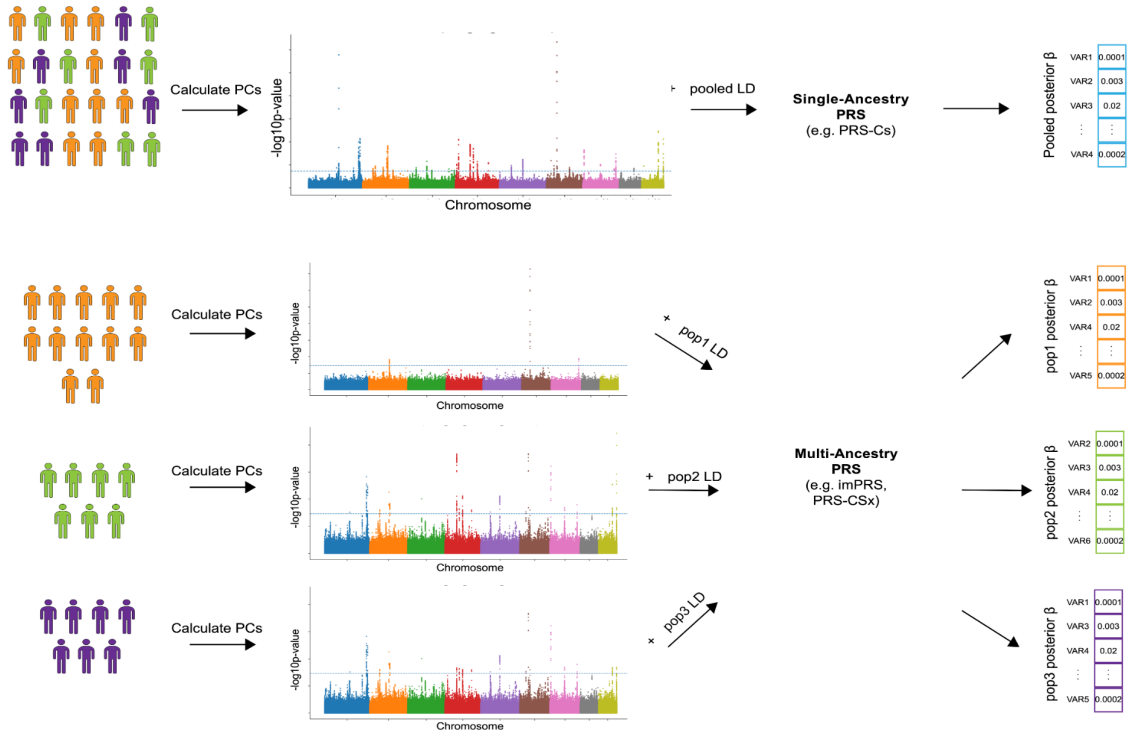
Chapter 3, is unpublished material by Helyaneh Ziaei Jam, Tara Mirmira, and Melissa Gymrek. The dissertation author was the primary investigator and author of this paper. This work was funded by NIH/NHGRI 1RM1HG011558 and NIH Common Fund U01HG013442.

3.8 Appendix

Supplementary Data 1 can be found here:

<https://docs.google.com/spreadsheets/d/1fKepy7kGMM0fRUm9Jsw2w4aXC1wq-aXh0AIzRCx4YAk/edit?usp=sharing>

a



b

$$E \left[\begin{bmatrix} \beta_{1,1} \\ \beta_{1,2} \\ \vdots \\ \beta_{1,k_1} \\ \beta_{2,1} \\ \beta_{2,2} \\ \vdots \\ \beta_{2,k_2} \end{bmatrix} \middle| \begin{bmatrix} \hat{\beta}_{1,1} \\ \hat{\beta}_{1,2} \\ \vdots \\ \hat{\beta}_{1,k_1} \\ \hat{\beta}_{2,1} \\ \hat{\beta}_{2,2} \\ \vdots \\ \hat{\beta}_{2,k_2} \end{bmatrix} \right] = \left(\begin{bmatrix} \text{Diagonal} & 0 \\ 0 & \text{Diagonal} \end{bmatrix} + \begin{bmatrix} \psi_{1,1} & 0 & \dots & \psi_{1,k_1} \\ \psi_{1,2} & \dots & \dots & \psi_{1,k_1} \\ \vdots & \vdots & \vdots & \vdots \\ \psi_{2,1} & \psi_{2,2} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \psi_{2,k_2} & \vdots \end{bmatrix} \right)^{-1} \begin{bmatrix} \hat{\beta}_{1,1} \\ \hat{\beta}_{1,2} \\ \vdots \\ \hat{\beta}_{1,k_1} \\ \hat{\beta}_{2,1} \\ \hat{\beta}_{2,2} \\ \vdots \\ \hat{\beta}_{2,k_2} \end{bmatrix}$$

Figure 3.1: Overview of PRS construction

a. Two approaches for PRS construction. Top: A pooled framework in which GWAS is performed on a combined multi-ancestry cohort, and the resulting summary statistics together with pooled linkage disequilibrium (LD) are used as input to a single-ancestry PRS method (e.g., PRS-CS), yielding a single set of posterior effect sizes for the union of variants across populations. Bottom: A multi-ancestry framework in which GWAS summary statistics and LD are estimated separately for each population and jointly analyzed using a multi-ancestry PRS method (e.g., PRS-CSx or imPRS) to produce population-specific posterior effect sizes. **b. Overview of the imPRS model.** A shared Bayesian framework explicitly incorporates a cross-population correlation parameter to enable information sharing across ancestries at shared variants. Posterior effect sizes for all populations are simultaneously inferred using population-specific GWAS summary statistics.

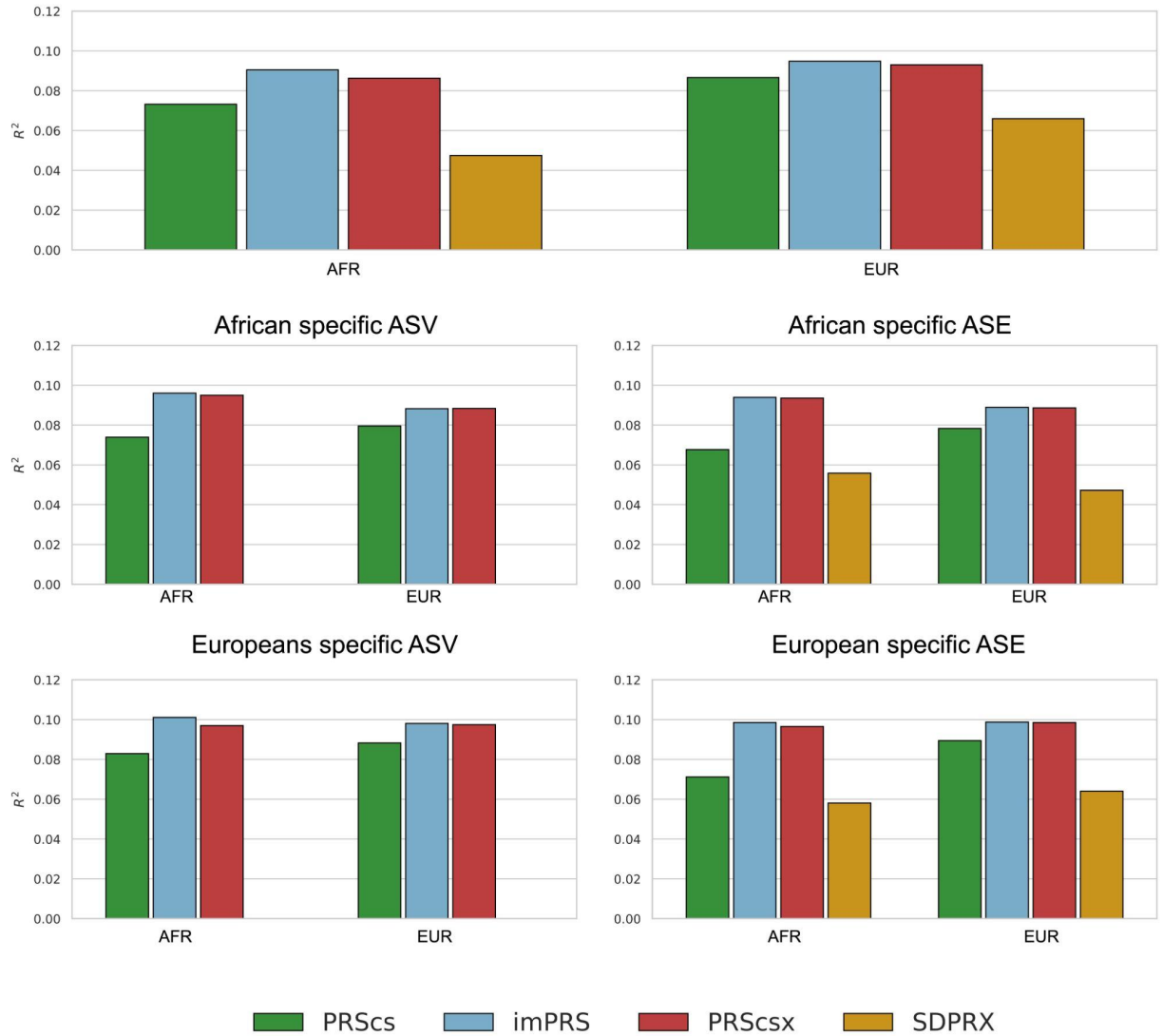


Figure 3.2: Benchmarking PRS prediction accuracy across ancestral architectures.

Prediction accuracy is compared for PRScs, imPRS, PRScsx, and SDPRX across African (AFR) and European (EUR) populations. Heritability is fixed at 0.1. The top row illustrates a baseline where causal architecture is shared between ancestries. Lower panels display performance under ancestry-specific causal configurations (Causal AFR and Causal EUR), further stratified by Ancestry Specific Variation (ASV) and Ancestry Specific Effect (ASE). SDPRX performance is reported for the baseline and ASE evaluations only, consistent with its methodological design in simulation setting.

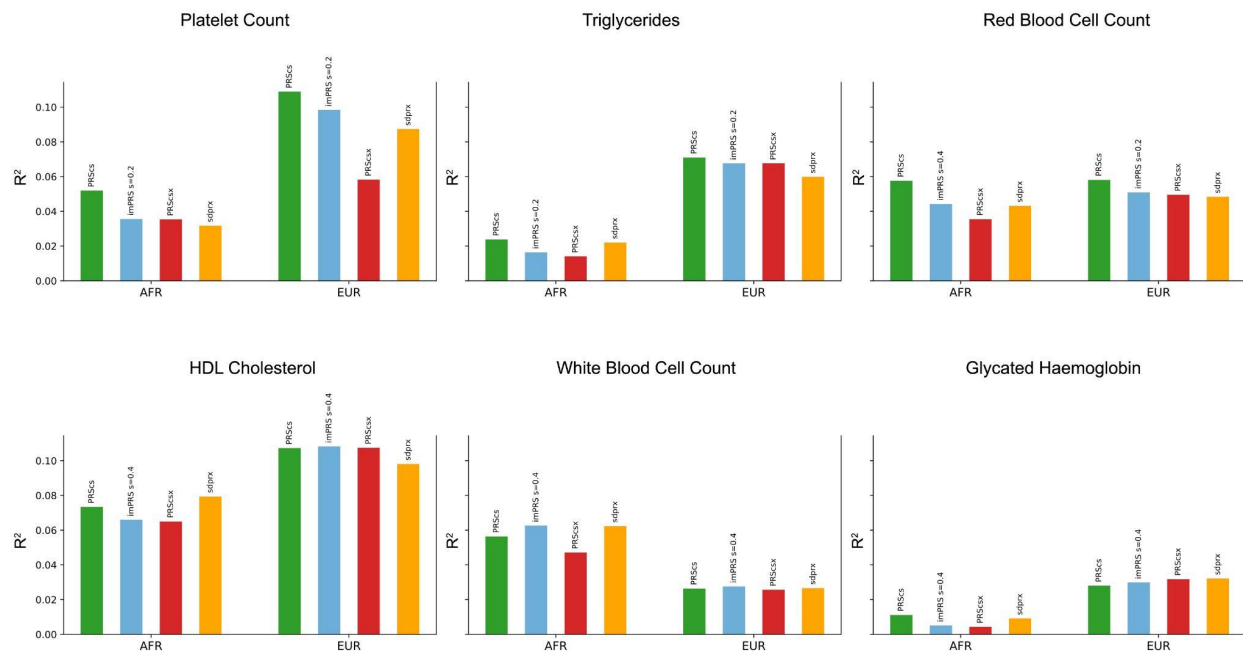


Figure 3.3: Benchmarking PRS prediction accuracy across 6 traits in UKB samples

For each trait, variance explained (R^2) for PRS-Cs, imPRS, PRS-CSx, and SDPRX are shown for African and European ancestry samples. The imPRS model presented here has the s value that demonstrated optimal predictive performance.

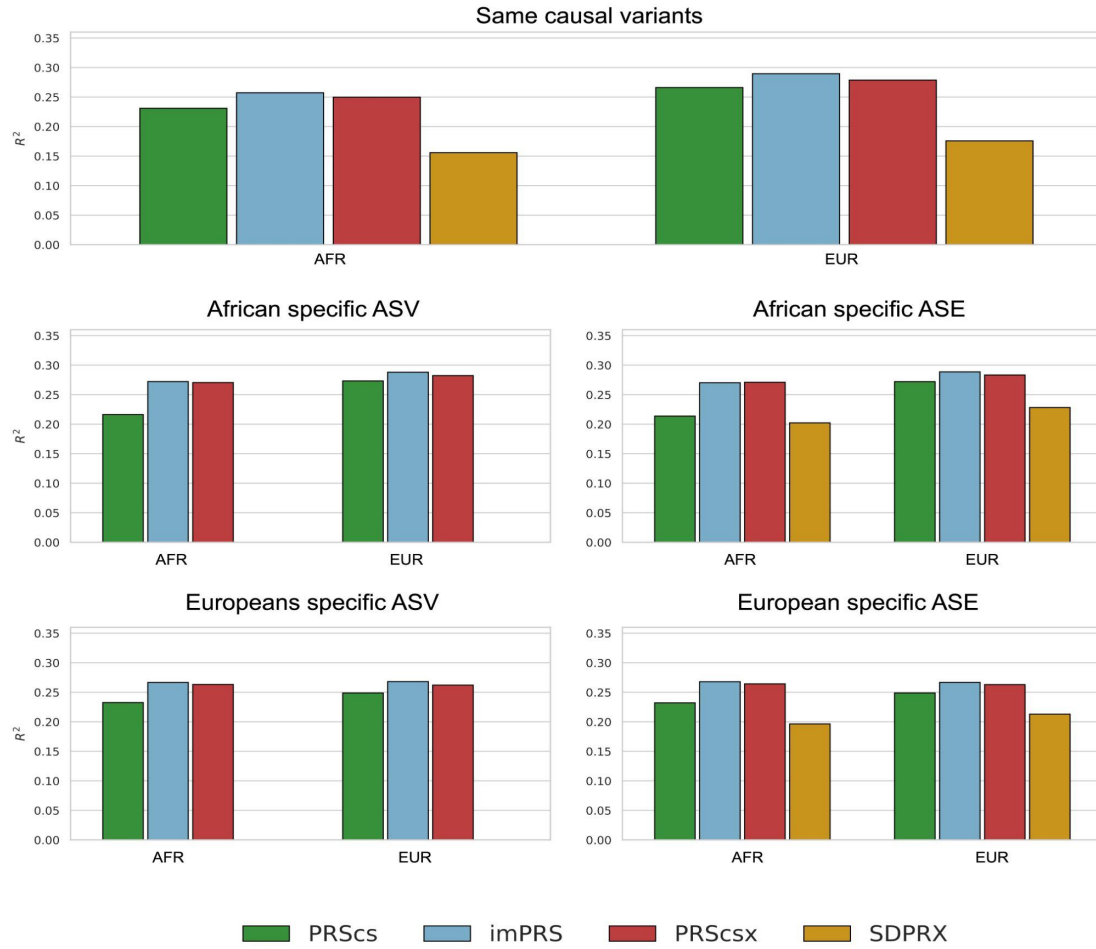


Figure 3.4: Benchmarking PRS prediction accuracy across ancestral architectures.

Prediction accuracy is compared for PRScs, imPRS, PRScsx, and SDPRX across African (AFR) and European (EUR) populations. Heritability is fixed at 0.3. The top row illustrates a baseline where causal architecture is shared between ancestries. Lower panels display performance under ancestry-specific causal configurations (Causal AFR and Causal EUR), further stratified by Ancestry Specific Variation (ASV) and Ancestry Specific Effect (ASE). SDPRX performance is reported for the baseline and ASE evaluations only, consistent with its methodological design in simulation setting.

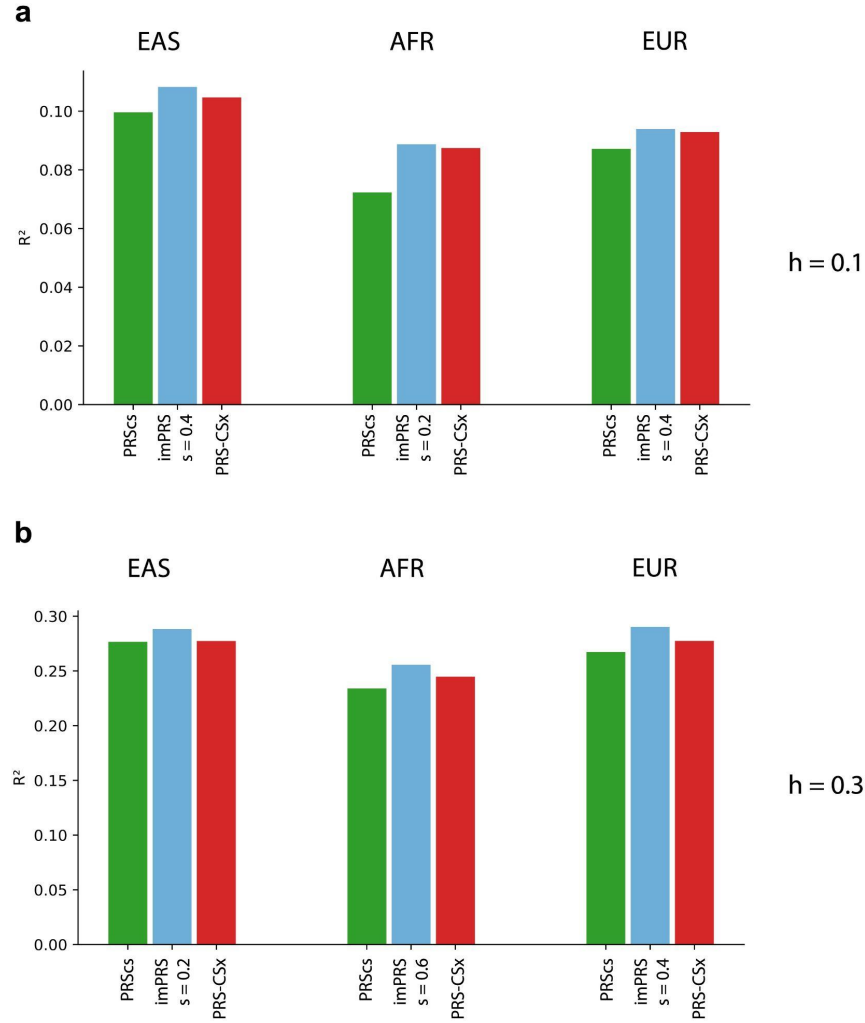


Figure 3.5: Benchmarking PRS prediction accuracy in EAS, AFR, and EUR populations

The top panel presents results with heritability fixed at 0.1, bottom panel shows for heritability fixed at 0.3. From left to right, prediction R^2 is shown for East Asian, African, and European samples. Within each population, the green bars denote PRSs predictions using pooled samples across ancestries, the blue bars represent imPRS performance at the optimal s value selected on the validation set, and the red bars correspond to PRS-CSx prediction accuracy.

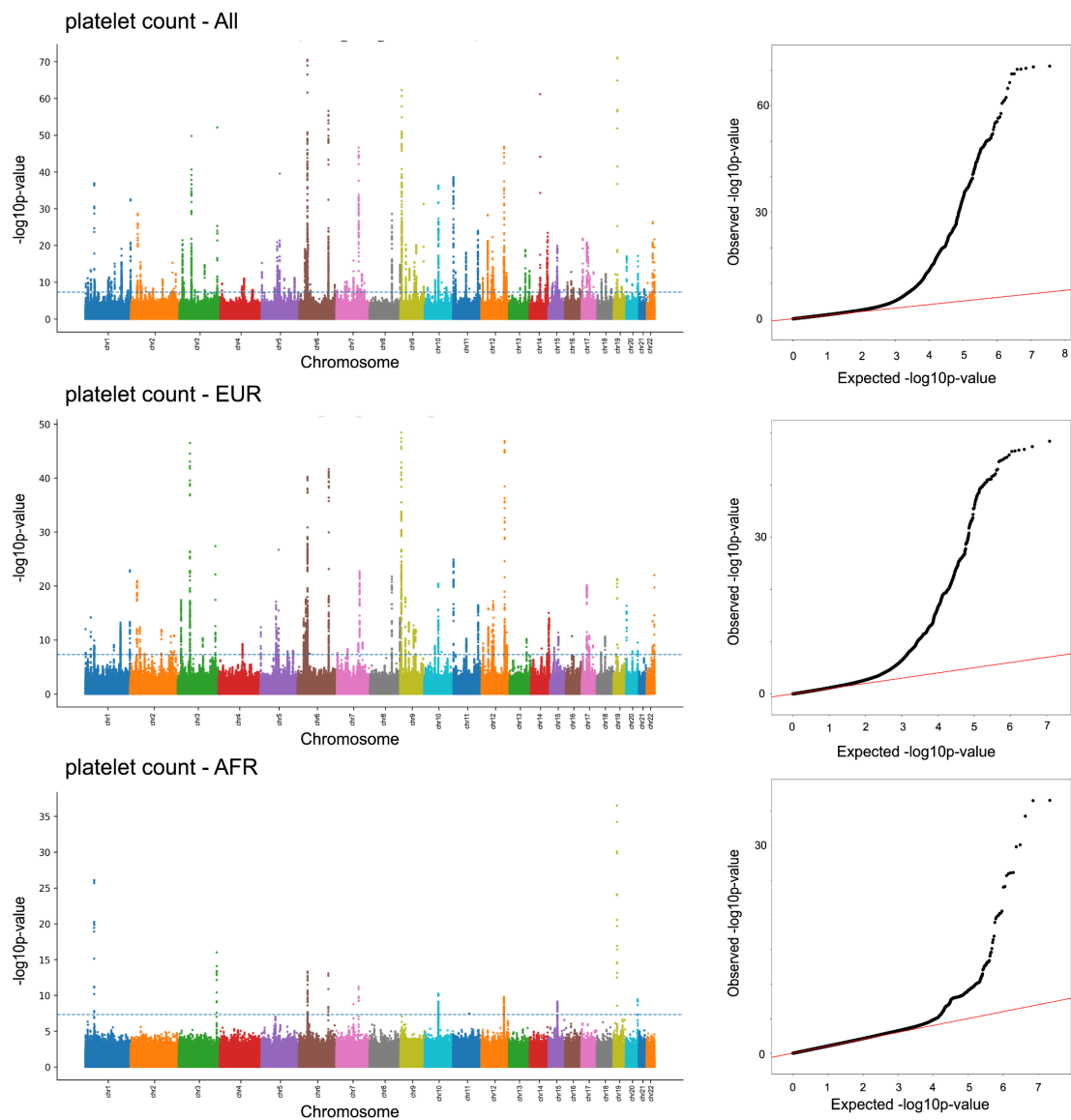


Figure 3.6: GWAS for platelet count.

Left panels show Manhattan plots for GWASs run separately in each of the 3 cohorts of AoU ALL, AoU EUR, AoU AFR. Right panels show quantile-quantile plots visualizing the distribution of p-values for the corresponding GWAS.

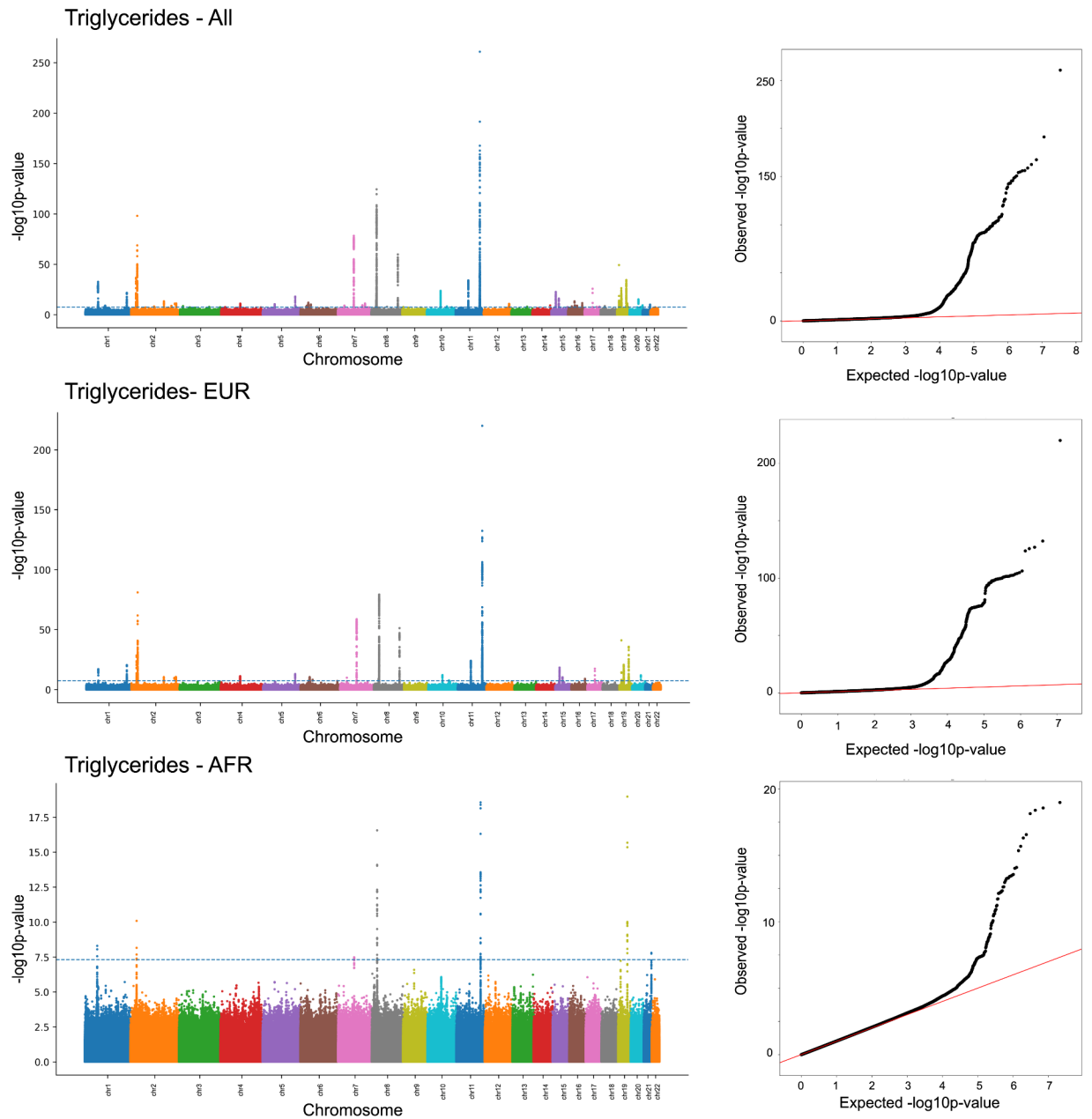


Figure 3.7: GWAS for triglycerides.

Left panels show Manhattan plots for GWASs run separately in each of the 3 cohorts of AoU ALL, AoU EUR, AoU AFR. Right panels show quantile-quantile plots visualizing the distribution of p-values for the corresponding GWAS.

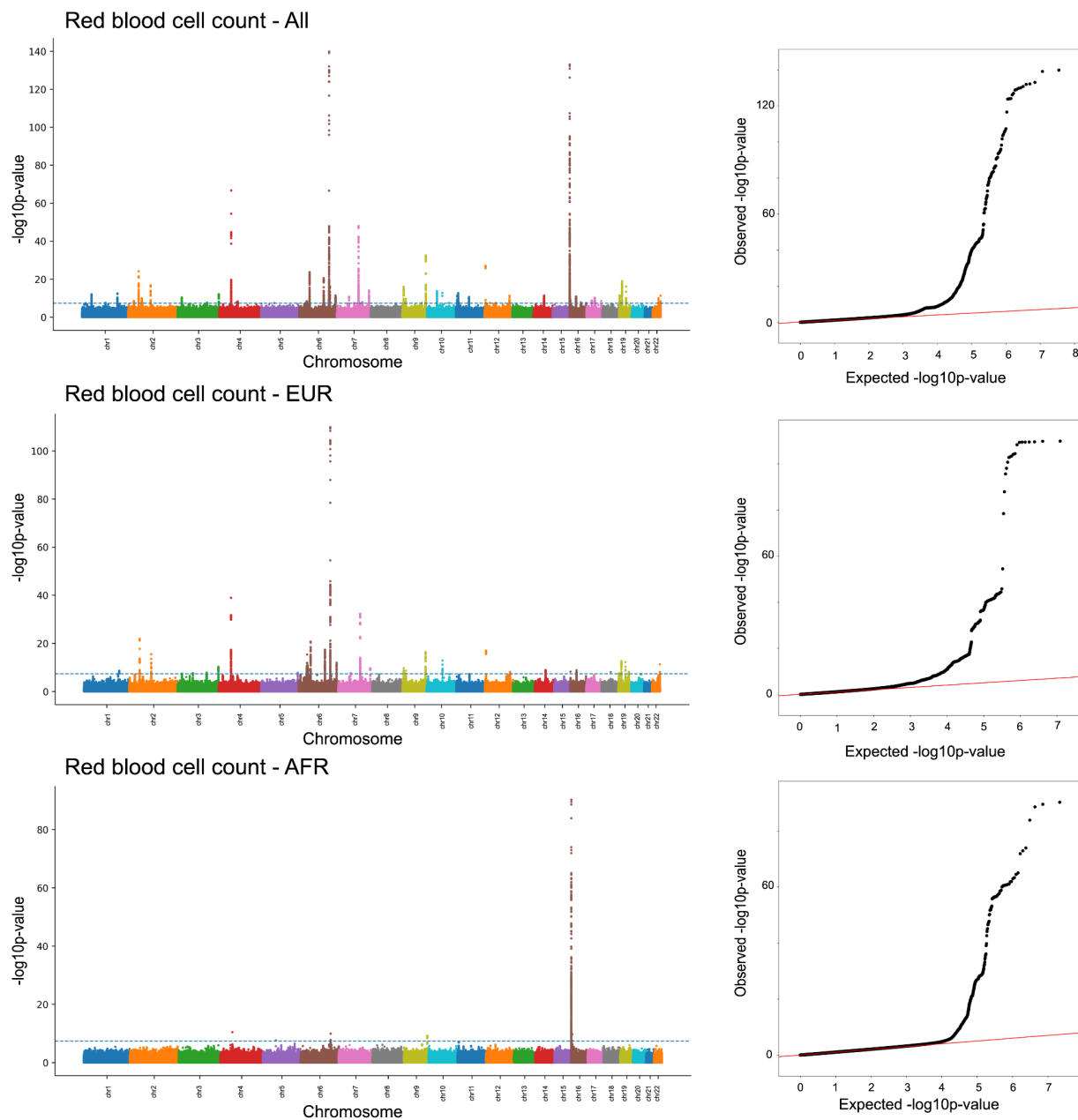


Figure 3.8: GWAS for red blood cell count.

Left panels show Manhattan plots for GWASs run separately in each of the 3 cohorts of AoU ALL, AoU EUR, AoU AFR. Right panels show quantile-quantile plots visualizing the distribution of p-values for the corresponding GWAS.

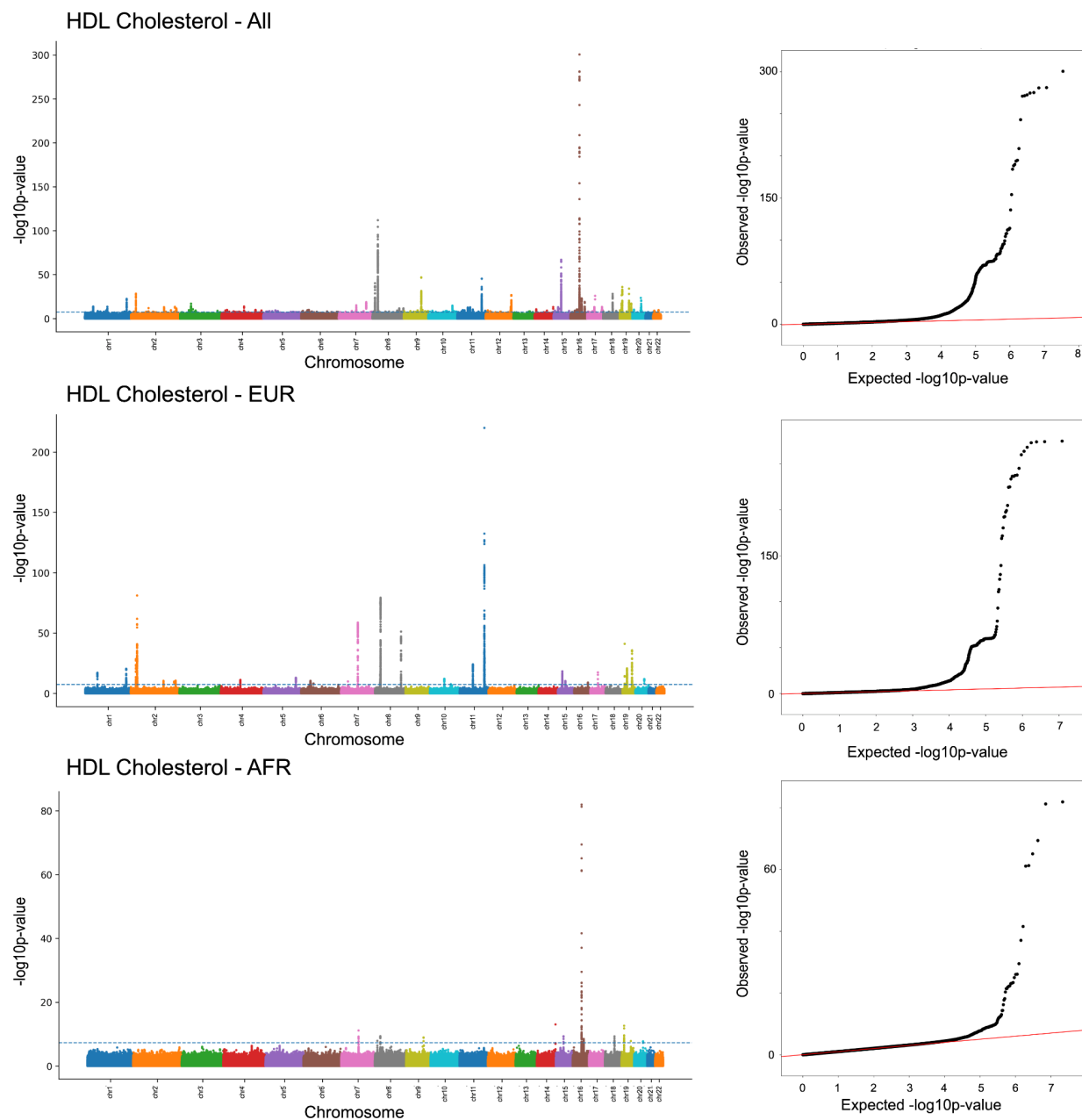


Figure 3.9: GWAS for HDL cholesterol.

Left panels show Manhattan plots for GWASs run separately in each of the 3 cohorts of AoU ALL, AoU EUR, AoU AFR. Right panels show quantile-quantile plots visualizing the distribution of p-values for the corresponding GWAS.

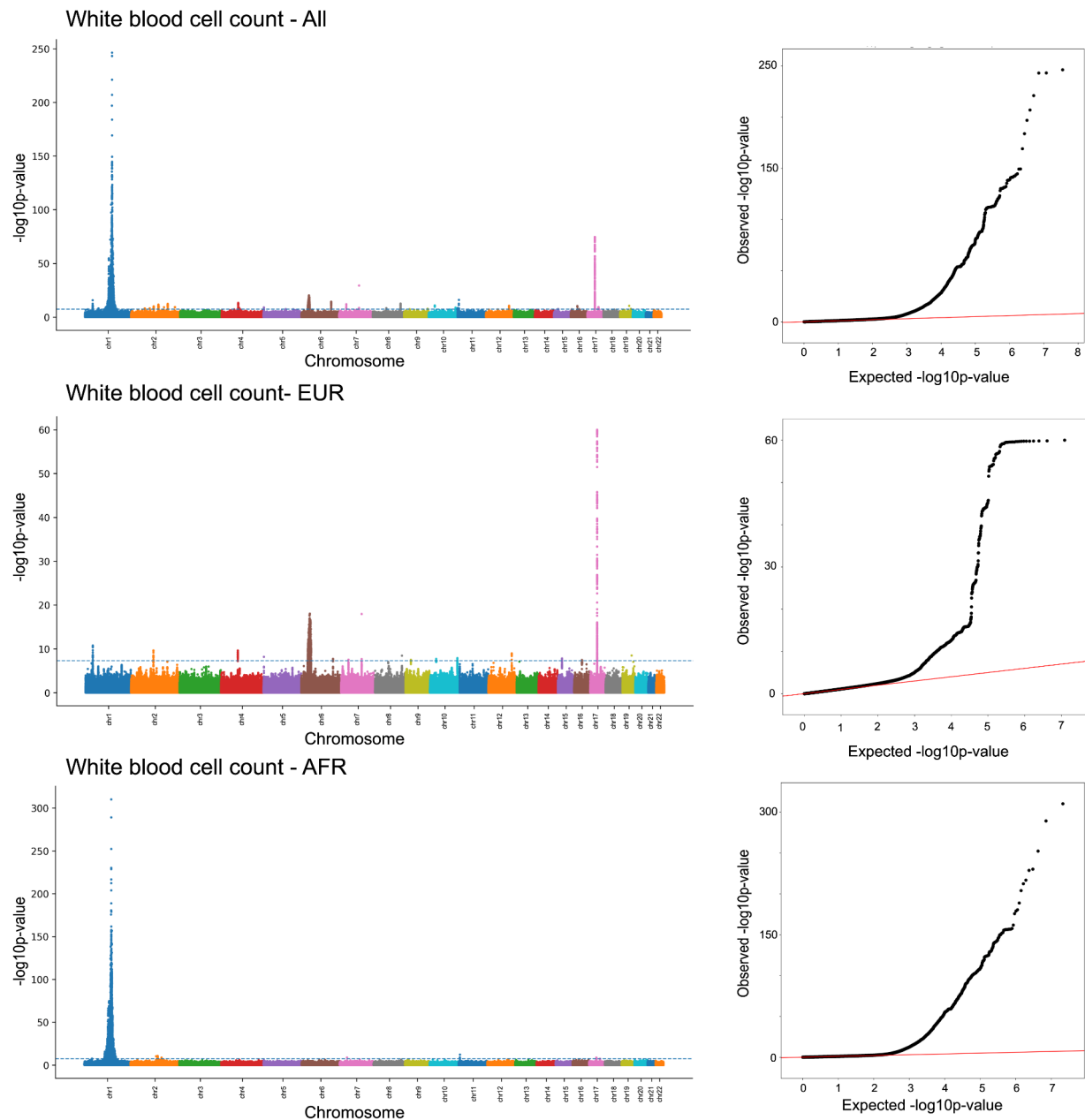


Figure 3.10: GWAS for white blood cell count.

Left panels show Manhattan plots for GWASs run separately in each of the 3 cohorts of AoU ALL, AoU EUR, AoU AFR. Right panels show quantile-quantile plots visualizing the distribution of p-values for the corresponding GWAS.

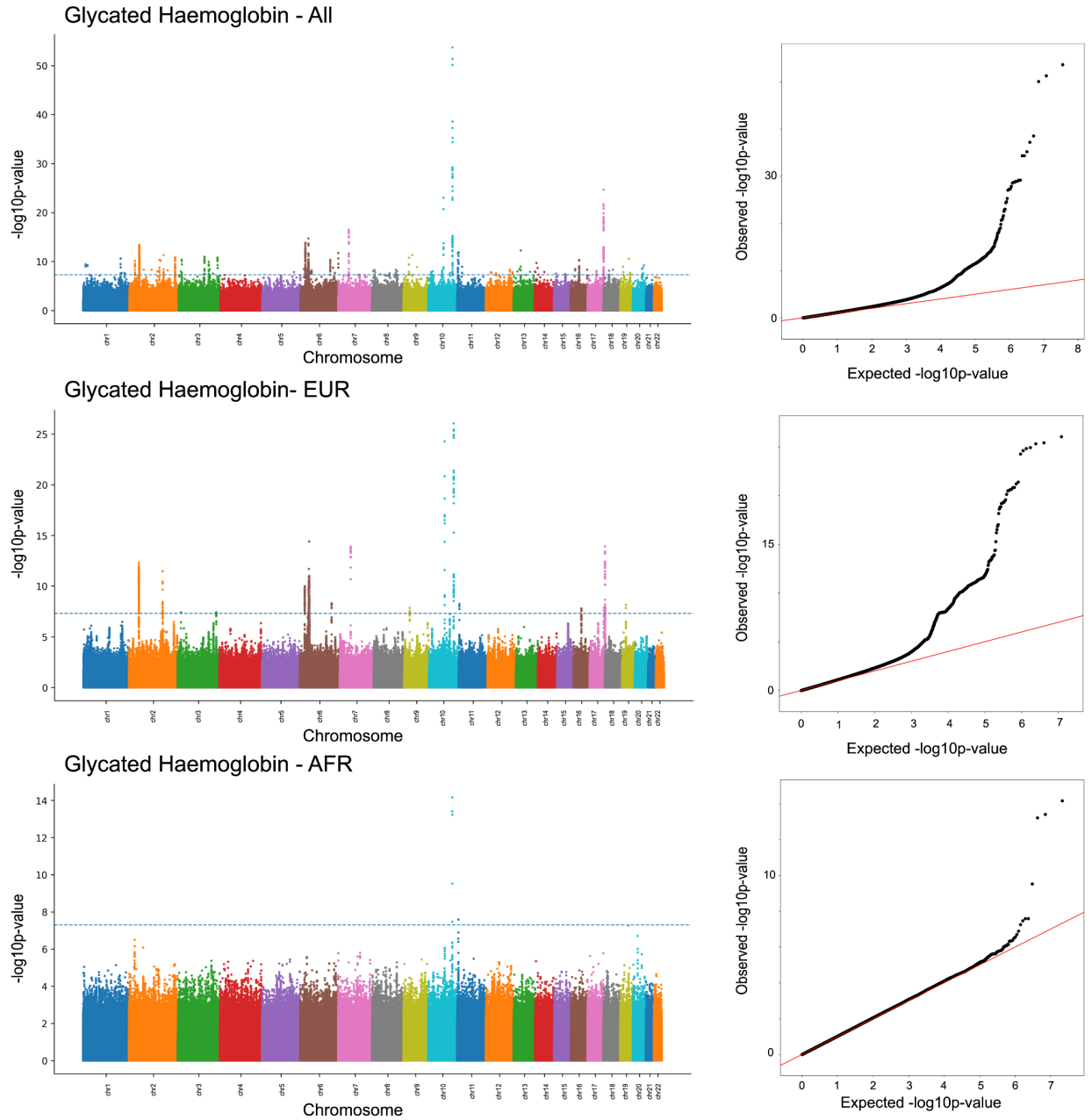


Figure 3.11: GWAS for glycated haemoglobin.

Left panels show Manhattan plots for GWASs run separately in each of the 3 cohorts of AoU ALL, AoU EUR, AoU AFR. Right panels show quantile-quantile plots visualizing the distribution of p-values for the corresponding GWAS.

3.9 References

1. A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, M. J. Daly, Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 51, 584–591 (2019).
2. S. Fatumo, D. Sathan, C. Samtal, I. Isewon, T. Tamuhla, C. Soremekun, J. Jafali, S. Panji, N. Tiffin, Y. J. Fakim, Polygenic risk scores for disease risk prediction in Africa: current challenges and future directions. *Genome Med* 15, 87 (2023).
3. M. A. Coram, H. Fang, S. I. Candille, T. L. Assimes, H. Tang, Leveraging Multi-ethnic Evidence for Risk Assessment of Quantitative Traits in Minority Populations. *Am J Hum Genet* 101, 638 (2017).
4. K. E. Grinde, Q. Qi, T. A. Thornton, S. Liu, A. H. Shadyab, K. H. K. Chan, A. P. Reiner, T. Sofer, Generalizing polygenic risk scores from Europeans to Hispanics/Latinos. *Genet Epidemiol* 43, 50–62 (2019).
5. Y. Ruan, Y.-F. Lin, Y.-C. A. Feng, C.-Y. Chen, M. Lam, Z. Guo, L. He, A. Sawa, A. R. Martin, S. Qin, H. Huang, T. Ge, Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics* 54, 573–580 (2022).
6. G. Zhou, T. Chen, H. Zhao, SDPRX: A statistical method for cross-population prediction of complex traits. *Am J Hum Genet* 110, 13–22 (2023).
7. C. J. Hoggart, S. W. Choi, J. García-González, T. Souaiaia, M. Preuss, P. F. O'Reilly, BridgePRS leverages shared genetic effects across ancestries to increase polygenic risk score portability. *Nat Genet* 56, 180–186 (2024).
8. L. Xu, G. Zhou, W. Jiang, H. Zhang, Y. Dong, L. Guan, H. Zhao, JointPRS: A data-adaptive framework for multi-population genetic risk prediction incorporating genetic correlation. *Nat Commun* 16, 3841 (2025).
9. K. Hou, Y. Ding, Z. Xu, Y. Wu, A. Bhattacharya, R. Mester, G. M. Belbin, S. Buyske, D. V. Conti, B. F. Darst, M. Fornage, C. Gignoux, X. Guo, C. Haiman, E. E. Kenny, M. Kim, C.

- Kooperberg, L. Lange, A. Manichaikul, K. E. North, U. Peters, L. J. Rasmussen-Torvik, S. S. Rich, J. I. Rotter, H. E. Wheeler, G. L. Wojcik, Y. Zhou, S. Sankararaman, B. Pasaniuc, Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nat Genet* 55, 549–558 (2023).
10. H. Shi, K. S. Burch, R. Johnson, M. K. Freund, G. Kichaev, N. Mancuso, A. M. Manuel, N. Dong, B. Pasaniuc, Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data. *Am J Hum Genet* 106, 805–817 (2020).
 11. M. Cai, J. Xiao, S. Zhang, X. Wan, H. Zhao, G. Chen, C. Yang, A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. *Am J Hum Genet* 108, 632–655 (2021).
 12. Y. Ding, K. Hou, Z. Xu, A. Pimplaskar, E. Petter, K. Boulier, F. Privé, B. J. Vilhjálmsdóttir, L. M. Olde Loohuis, B. Pasaniuc, Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* 618, 774–781 (2023).
 13. P. Anagnostou, V. Dominici, C. Battaggia, A. Lisi, S. Sarno, A. Boattini, C. Calò, P. Francalacci, G. Vona, S. Tofanelli, M. G. Vilar, V. Colonna, L. Pagani, G. Destro Bisol, Inter-individual genomic heterogeneity within European population isolates. *PLoS One* 14, e0214564 (2019).
 14. J.-A. Dias, T. Chen, H. Xing, X. Wang, A. A. Rodriguez, R. K. Madduri, P. Kraft, H. Zhang, Evaluating multi-ancestry genome-wide association methods: Statistical power, population structure, and practical implications. *Am J Hum Genet* 112, 2493–2508 (2025).
 15. T. Ge, C.-Y. Chen, Y. Ni, Y.-C. A. Feng, J. W. Smoller, Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun* 10, 1776 (2019).
 16. The International HapMap Project. *Nature* 426, 789–796 (2003).
 17. J. Margoliash, S. Fuchs, Y. Li, X. Zhang, A. Massarat, A. Goren, M. Gymrek, Polymorphic short tandem repeats make widespread contributions to blood and serum traits. *Cell Genom* 3, 100458 (2023).

18. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88, 76–82 (2011).
19. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7 (2015).
20. PLINK 1.9. www.cog-genomics.org/plink/1.9/.
21. T. Mirmira, N. Ma, J. Margoliash, W. G. Gonzalez Rivera, T. Amariuta, K. A. Frazer, A. Goren, M. Gymrek, Characterization of shared and ancestry-specific signals driving complex traits using multi-ancestry fine-mapping, *medRxiv* (2025)p. 2025.11.13.25339614.
22. W. G. Gonzalez Rivera, Y. Liu, T. Mirmira, N. Ma, J. Kim, M. D. Antonio, U. Dube, K. A. Frazer, M. Gymrek, Admixture mapping identifies complex trait associations with local ancestry in the All of Us Research Program, *medRxiv* (2025)p. 2025.12.29.25343152.
23. Website. <https://github.com/hail-is/hail>.