UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Haplotype-based fine-mapping of variant associations for complex traits

Permalink

https://escholarship.org/uc/item/5v35p6r8

ISBN

9798173721372

Author

Massarat, Arya

Publication Date

2026-11-02

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Haplotype-based fine-mapping of variant associations for complex traits

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

in

Bioinformatics and Systems Biology

by

Arya Massarat

Committee in charge:

Professor Melissa Gymrek, Chair
Professor Abraham Palmer, Co-Chair
Professor Tiffany Amariuta
Professor Graham McVicker
Professor Armin Schwartzman

2026

Copyright

Arya Massarat, 2026

All rights reserved.

The Dissertation of Arya Massarat is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2026

DEDICATION

Love is greater than peace, for peace is founded upon love. -'Abdu'l-Bahá

This dissertation is dedicated to the family that lovingly, without the merest hesitation, gave everything to me.

To my parents, Barzin and Shirin Massarat, for all that you have shown and taught me. I would not be a shred of the person that I am today without you.

To my younger brother, Navid Massarat, my first and forever friend, confidant, and companion. May we never stop growing together.

To my grandparents, cousins, aunts, and uncles. I am always learning love from you.

TABLE OF CONTENTS

DISSERTATION APPROVAL PAGE iii
DEDICATION iv
TABLE OF CONTENTS v
LIST OF FIGURES vi
LIST OF TABLES vii
LIST OF ABBREVIATIONS viii
ACKNOWLEDGEMENTS ix
VITA xi
ABSTRACT OF THE DISSERTATION xii
INTRODUCTION 1
0.1 Summary 1
0.2 Genome-wide association studies 1
0.3 Fine-mapping and its limitations 2
0.4 Haplotypes can serve as proxies for missing genetic contributions 3
0.5 Overview of subsequent chapters 4
Chapter 1 Haptools, a simulation framework for benchmarking fine-mapping methods 6
1.1 Introduction 6
1.2 Commands 7
1.2.1 .hap file format 7
1.2.2 haptools transform 7
1.2.3 haptools simphenotype 8
1.3 Discussion 9
1.4 Data availability 9
1.5 Acknowledgements 9
Chapter 2 Happler, a haplotype-based fine-mapping approach 11
2.1 Introduction 11
2.2 Methods 13
2.3 Results 24
2.4 Discussion 32
2.5 Code availability 34
2.6 Acknowledgements 34
REFERENCES 35

LIST OF FIGURES

Figure 1: Single nucleotide polymorphisms (SNPs) and tandem repeat (TRs) 4

Figure 2: Example analyses performed using haptools 8

Figure 3: The happler method 14

Figure 4: Benchmarking happler's haplotype construction step using simulations 27

Figure 5: Benchmarking haplotype-based fine-mapping with simulations 28

Figure 6: An application of happler to fine-map eQTLs in the Geuvadis cohort 31

Figure 7: Tagging rare causal variants in All of Us 33

LIST OF TABLES

Table 1: Summary of current haptools utilities 7

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

PC – (genetic) principal component

SNP/SNV – single nucleotide polymorphism/variant

TR – tandem repeat

AoU – All of Us cohort

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Professor Melissa Gymrek for her incredibly thoughtful mentorship throughout the course of my PhD. None of this would have been possible without her. She is caring, thoughtful, approachable, and patient. She helped me enter a field in which I had very little background, and yet she helped me become an expert, encouraging me and giving me hours of her time. I count myself incredibly fortunate to have been her mentee.

Thank you also to all of the family, friends, and lab members who supported me and encouraged me during my PhD. Thanks especially to my parents and my brother for all of the meals, phone calls, and love that you showered upon me. Thanks to the Gymrek lab for giving me a reason to come to campus everyday. My PhD has been so much fun because of you. Also, many thanks to all of the friends along the way. You were there through climbing, surfing, running, dog-sitting, camping, and many other adventures.

Formal acknowledgements for the papers in my thesis:

We thank Jonathan Margoliash for helpful discussions and Tara Mirmira and Wilfredo Gonzalez-Rivera for feedback on haptools utilities. The work in chapter 1 was supported in part by the National Institutes of Health [1RM1HG011558 to M.G. and R35GM133805 to A.L.W.]. A.R.M. was also supported by the National Science Foundation Graduate Research Fellowship [DGE-2038238] and grants from the National Institutes of Health [T32GM008806 and T32GM139790].

Chapter 1, in full, is a reformatted reprint of the material as it appears in "Haptools: a toolkit for admixture and haplotype analysis.” Arya R Massarat, Michael Lamkin, Ciara Reeve, Amy L Williams, Matteo D’Antonio, Melissa Gymrek. *Bioinformatics*, Volume 39, Issue 3,

March 2023. The dissertation author was one of two primary investigators and authors of this paper.

Chapter 2 was partially supported by NIH/NHGRI grant RM1HG011558 (M.G.). A.R.M. was also supported by the National Science Foundation Graduate Research Fellowship [DGE-2038238] and grants from the National Institutes of Health [T32GM008806 and T32GM139790].

Chapter 2, in part is currently being prepared for submission for publication of the material. "Haplotype-based fine-mapping of variant associations for complex traits." Arya R Massarat, Michael Lamkin, Tara Mirmira, Utkarsh Jain, Jonathan Margoliash, Yang Li, Melissa Gymrek. The dissertation author was the primary researcher and author of this material.

VITA

2020 Bachelor of Science in Mathematical and Computational Biology, Harvey Mudd College

2026 Doctor of Philosophy in Bioinformatics and Systems Biology, University of California San Diego

FIELD OF STUDY

Major Field: Bioinformatics and Systems Biology
Studies in statistical fine-mapping and population genetics
Professor Melissa Gymrek

ABSTRACT OF THE DISSERTATION

Haplotype-based fine-mapping of variant associations for complex traits

by

Arya Massarat

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2026

Professor Melissa Gymrek, Chair
Professor Abraham Palmer, Co-Chair

Genome-wide association studies (GWAS) have identified tens of thousands of associations between genomic regions and complex traits in humans, but identifying the causal variants underlying these associations is challenging. Statistical fine-mapping techniques attempt to quantify the probability of causality of individual variants while accounting for the correlation

induced by linkage disequilibrium (LD) between nearby variants. Existing fine-mapping solutions are limited to considering additive and independent contributions of individual variants, and also assume the causal variant is included in the analysis. Here, we introduce a novel fine-mapping method, happler, which models the effects of variant haplotypes (sets of alleles of two or more variants in cis) in addition to individual variants. By identifying haplotype proxies for causal variants, happler can infer additional effects including those driven by combinations of variants in cis and ungenotyped causal variants, such as complex structural variants, which are in some cases better tagged by one or more haplotypes than by individual SNPs. We benchmarked happler using simulated data from our custom toolkit for haplotype-based analyses, haptools. Simulation experiments show that whereas happler can correctly identify causal haplotype effects, traditional methods based on individual variants can produce inconclusive results in such cases. We apply happler to fine-map expression quantitative trait loci (eQTLs) using 359 individuals from the Geuvadis dataset. We identified a total of 19 signals that are fine-mapped to a candidate causal haplotype rather than an individual genotyped variant, 73.7% of which replicate in the Genotype Tissue Expression project. We also identify haplotypes associated with platelet count in the All of Us dataset and demonstrate that these are correlated with ungenotyped rare, causal variants. Future studies will be needed to comprehensively investigate the underlying mechanisms driving these haplotype-level effects.

INTRODUCTION

0.1 Summary

Many diseases or traits are attributed to whole-genome genetic variation. Researchers use *genome-wide association studies* (GWAS) of populations to look for patterns between common genetic variants and traits, and then identify trait-associated regions (*loci*) across a genome. Subsequent *fine-mapping* pinpoints putative *causal* variants in each GWAS region which are associated with the trait. Yet estimates of the genetic contribution of traits from these causal variants still fail to satisfy corresponding estimates from studies of twins. Geneticists hypothesize that part of the gap between the two estimates is due to either genetic variation that might be missed or interactions between the variants which may be ignored by fine-mapping. Both kinds of missingness may be captured by *haplotypes*, combinations of genetic variants that are inherited together. In this dissertation, I describe two methods for working with haplotypes: *haptools*, a toolkit for admixture and haplotype analysis, and *happler*, a haplotype-based fine-mapping approach.

0.2 Genome-wide association studies

Genome-wide association studies (GWASs) are a popular approach for exploring the genetic components of *complex traits*, which are affected by genes throughout the genome. GWASs to date have uncovered thousands of genomic loci associated with phenotypes. Each locus can serve as a starting point for understanding a trait's molecular mechanism and predicting an individual's proclivity for a trait.

Before performing a GWAS, all of the genetic variation in a population is genotyped to catalog differing alleles at each genomic position within every individual. Most differences

between people are single nucleotide variants known as single nucleotide polymorphisms (or SNPs) but there may also be other kinds of variants, including insertions/deletions (indels) or tandem repeats (TRs)[1]. GWAS have traditionally been limited to studying simple biallelic SNPs for the sake of computational simplicity[2]. Genotyping assigns to each individual in a population a number (0, 1, or 2), denoting the number of chromosomal copies of an allele of the SNP. In a genome-wide association study, the genotypes of each SNP are tested for an association with a target trait. For each SNP, we are essentially fitting a line on a scatter plot where the x-axis are the genotypes of the SNP (0, 1, or 2) and the y-axis is the target quantitative trait. The result is a set of p-values for each SNP. Regions in which there are strong p-values are flagged for downstream analysis[1].

0.3 Fine-mapping and its limitations

While GWAS hits can reliably point to important regions of the genome contributing to particular traits, GWAS often fail to identify the exact variant in the DNA that leads to the trait. Instead, they tend to identify a variant that is simply correlated with the true causal variant[1]. These correlations (known as *linkage disequilibrium* or LD) vary between populations, so they may differ between the study and target populations. Statistical *fine-mapping methods* have been developed to more rigorously probe the associations of each SNP in a GWAS-significant locus and take into account LD, but these too suffer from their own limitations, described below.

Until recently, most fine-mapping methods have only considered common simple single nucleotide polymorphisms (SNPs), as candidate causal variants, but increasing evidence may suggest other variants that might not be directly genotyped make important contributions as well. For example, it has been long hypothesized that many of the most important SNPs for disease are too rare to appear in many GWAS datasets[3]. In addition, there are many other kinds of *complex*

variants, like tandem repeats (TRs) and structural variants, which are often excluded from GWAS but are just as critical for evaluating disease risk. For example, studies have shown that the number of repeats that an individual may have in a repetitive region of the DNA (ex: ATATAT) can predict the occurrence of many traits, including epilepsy and muscular dystrophy[4]. In one study, five tandem repeats (TRs) in protein-coding exons were found to be causal drivers of the strongest GWAS hits for traits like height and kidney function[4]. Lastly, fine-mapping methods ignore interactions between variants that may non-additively associate with the trait, such as variants in epistasis. Properly accounting for all of these limitations can be difficult and potentially expensive.

These limitations may also contribute to the so-called missing heritability problem of genetics. The *heritability* of a trait measures the proportion of phenotypic variation which may be attributed to genetic variation. It can be estimated, for example, by considering all of the effects of putative causal variants from genome-wide association studies and fine-mapping. However, it can also be estimated by looking at trait differences between people who are genetically identical, such as twins. Broadly defined, the *missing heritability problem* refers to the fact that heritability estimates from twin studies typically overestimate the genetic contribution to traits, relative to heritability estimates from GWAS and fine-mapping. This might imply that genetic association studies are unable to account for certain genetic contributions, perhaps because they fail to account for ungenotyped variants described above including rare or complex variants and epistatic interactions.

0.4 Haplotypes can serve as proxies for missing genetic contributions

Fortunately, rare ungenotyped variants or complex variants may be inferred (or tagged) by a combination of correlated SNP alleles[5]. For example, a TR with three possible lengths could

be inferred from two correlated SNPs (**Figure 1**). In theory, this would make it possible to test for the association between ungenotyped variants and a trait using only the SNPs that tag it. We refer to these collections of SNPs as *haplotypes*. Formally, we consider *a haplotype* as a set of alleles that tend to appear together on the same chromosomal copy (in cis). Then our goal becomes to identify haplotypes that help tag ungenotyped causal variants. Haplotypes may also help capture the effects of variants acting in combination.



Figure 1 (A) The alleles in SNPs 1 and 2 can be used to infer the length of a tandem repeat (TR).

0.5 Overview of subsequent chapters

In this thesis, I aimed to develop a collection of tools for the analysis, storage, and creation of haplotypes. In Chapter 1, I describe *haptools*, a simulation toolkit for admixture and haplotype analysis, which can evaluate the performance of mine and existing fine-mapping tools in difficult situations, such as those in which the causal variant is not present in the dataset or those in which a haplotype is causal instead. The haptools suite of commands also enables researchers to more easily and efficiently work with a variety of standard file formats for genotypes, phenotypes, haplotypes, covariates, and local ancestry. In Chapter 2, I describe *happler*, a haplotype-based fine-mapping approach based on haplotypes, which uses phased individual-level genotypes and phenotypes to produce trait-associated haplotypes. These haplotypes can be included along with SNPs when executing traditional fine-mapping approaches, such as SuSiE[6] or FINEMAP[7]. We benchmarked happler using simulated data from haptools and show that whereas happler can

correctly identify causal haplotype effects, SuSiE with only individual variants can produce inconclusive results in such cases. We apply happler to fine-map expression quantitative trait loci (eQTLs) from the Geuvadis dataset and identify signals that are fine-mapped to a candidate causal haplotype rather than an individual genotyped variant, which replicate in an orthogonal dataset, the Genotype Tissue Expression project. We also identify haplotypes associated with platelet count in the All of Us dataset and demonstrate that these are correlated with ungenotyped rare, causal variants.

Chapter 1 Haptools, a simulation framework for benchmarking fine-mapping methods

1.1 Introduction

Existing frameworks for complex trait analysis are typically based on variant-level analysis. However, phenotypic effects may also be mediated by haplotypes[8,9] (combinations of variants on the same chromosome) or by the local ancestry background on which a variant falls[10,11]. Incorporating these effects may improve the utility of genomic information for diverse and recently admixed individuals, but current tools have limited support for including these features. We present haptools, an open-source toolkit for facilitating local-ancestry aware and haplotype-based analysis of complex traits. Haptools supports fast simulation of admixed genomes, visualization of admixture tracks, simulating haplotype- and local ancestry-specific phenotype effects, and computing a variety of common file operations and statistics in a haplotype-aware manner. Overall, haptools provides a valuable set of utilities for developing and benchmarking methods for ancestry-aware analysis of complex traits.

Haptools consists of a suite of command-line utilities and a corresponding Python library for performing simulations and common file operations on haplotypes, local ancestry labels, and individual variants (Supplementary Figures 1-2, Table 1). Haptools is compatible with standard file formats as inputs and outputs, including VCF, PLINK, and the newer PGEN format which results in greatly improved computational performance (Supplementary Figure 3). In the following sections, we summarize the current core functionality available in haptools.

1.2 Commands

Table 1: Summary of current haptools utilities

Command	Description
simgenotype	Simulate admixed genomes
karyogram	Generate chromosome paintings for admixed individuals
simphenotype	Simulate phenotypes for complex traits with variant-, haplotype-, or local-ancestry specific effects
transform	Obtain a VCF of pseudo-genotypes from a set of haplotypes
ld	Compute linkage disequilibrium between haplotypes (or genotypes) and a specific target haplotype
index	Sort, compress, and index .hap files

1.2.1 .hap file format

Haptools implements a custom file format (*.hap) for flexible representation of haplotype-level and other information. These files consist of a collection of haplotypes. Each haplotype is defined by a set of one or more variants and their alleles, and optionally a local ancestry label, that tend to be inherited together on an individual chromosome (Supplementary Figure 2). Unlike previous haplotype representations, the format is compatible with tabix[12], and can be easily sorted and queried at the variant or haplotype level.

1.2.2 haptools transform

The *transform* command produces a VCF file of pseudo-genotypes, in which each haplotype is encoded as a bi-allelic variant record, for a set of haplotypes in .hap format. This operation can facilitate downstream tasks which require variant-level information as input. For example, to perform association testing based on haplotypes, one could first use *transform* to

generate a VCF encoding haplotypes as bi-allelic variants, and then use a standard framework such as PLINK[13] to perform association tests.

1.2.3 haptools simphenotype

The *simphenotype* utility simulates phenotypes for complex traits with variant-, haplotype- or local-ancestry specific effects. Causal haplotypes are loaded from a VCF file output by the *transform* command and their effect sizes are specified in a .hap file. By default, the *simphenotype* command simulates quantitative traits. Users may specify case-control traits by providing disease prevalence, which results in that percentage of individuals with the highest trait value being labeled as cases. We evaluated *simphenotype* by simulating quantitative traits under various scenarios, including individual causal variants (Supplementary Figure 6), local-ancestry effects (Figure 2b, Supplementary Figure 7), and haplotype-level effects (Figure 2c).

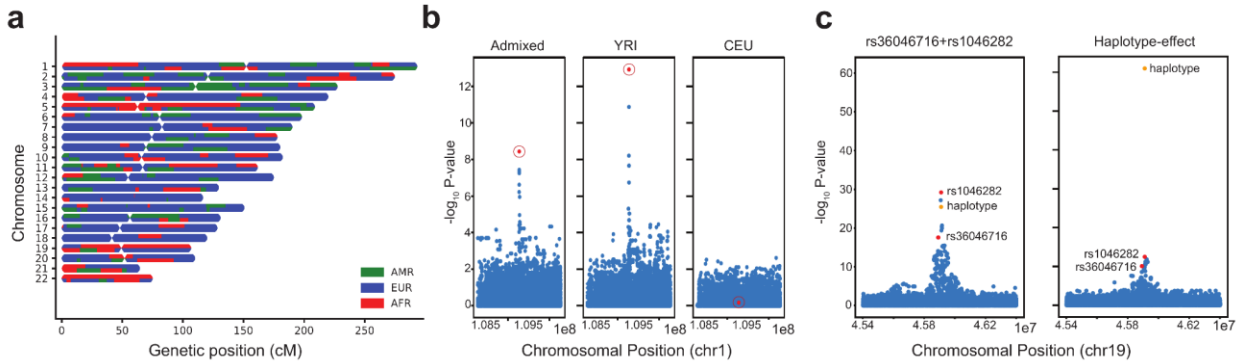


Figure 2: **Example analyses performed using haptools. (a)** An example karyogram depicting local ancestry tracts simulated by the simgenotype command. **(b)** Manhattan plot showing association summary statistics (-log10 p-values) for a trait with a single SNP (circled) simulated to be causal only when it occurs on an African haplotype. The SNP (rs12740374) is highly significant in simulated African but not European individuals. It has an intermediate p-value in a sample of simulated admixed individuals. **(c)** Manhattan plot showing association summary statistics for a trait simulated with either two causal SNPs (rs36046716 and rs1046282; left) or a single causal haplotype (composed of alleles from the two SNPs; right). Red=SNP-level p-values and orange=haplotype-level p-values for the variants of interest. When the haplotype is causal (right), it has a more significant p-value than the SNPs it is composed of. This large effect could be missed by variant-level association tests. Detailed methods underlying results shown in the figures are in the Supplementary Material.

1.3 Discussion

Accounting for ancestry and other more complex effects is becoming increasingly critical in association testing and downstream analysis pipelines. Haptools helps fulfill an unmet need to handle ancestry and haplotype level information in a standardized way that is compatible with existing file formats and workflows and is computationally efficient. It enables easily performing tasks such as genotype and phenotype simulation, local ancestry visualization, and power analyses which previously have been done primarily using a variety of custom scripts. Overall, haptools will help enable more systematic incorporation of ancestry and haplotype-level features in future workflows.

1.4 Data availability

The datasets used for validation and example generation are available from the haptools documentation page: https://haptools.readthedocs.io/en/stable/project_info/example_files.html, and the haptools-paper repository, https://github.com/CAST-genomics/haptools-paper.

1.5 Acknowledgements

We thank Jonathan Margoliash for helpful discussions and Tara Mirmira and Wilfredo Gonzalez-Rivera for feedback on haptools utilities. The work in chapter 1 was supported in part by the National Institutes of Health [1RM1HG011558 to M.G. and R35GM133805 to A.L.W.]. A.R.M. was also supported by the National Science Foundation Graduate Research Fellowship [DGE-2038238] and grants from the National Institutes of Health [T32GM008806 and T32GM139790].

Chapter 1, in full, is a reformatted reprint of the material as it appears in "Haptools: a toolkit for admixture and haplotype analysis." Arya R Massarat, Michael Lamkin, Ciara Reeve, Amy L Williams, Matteo D'Antonio, Melissa Gymrek. *Bioinformatics*, Volume 39, Issue 3,

March 2023. The dissertation author was one of two primary investigators and authors of this paper.

Chapter 2 Happler, a haplotype-based fine-mapping approach

2.1 Introduction

Genome-wide association studies (GWAS) have identified tens of thousands of genomic loci associated with a range of traits in humans[14]. However, identifying the underlying causal variant driving association signals is challenging. A typical GWAS signal may harbor hundreds of variants in strong linkage disequilibrium (LD) with each other, most of which are not causal for the trait. A variety of methods have been developed to perform statistical fine-mapping, with the goal of inferring the candidate causal variants at a particular GWAS locus, while accounting for LD[15]. However, existing methods face several important limitations: first, they assume the true causal variants exist within the set of candidate variants considered, which most often consists of common single nucleotide polymorphisms (SNPs) at a locus. Yet increasing evidence suggests that many of the strongest GWAS signals are driven by rare or complex variants that are in many cases not directly genotyped and may be in only modest LD with surrounding SNPs[2,16,17]. Second, they typically model independent contributions of individual variants rather than combinations of variants, which may either causally drive a signal or tag ungenotyped causal variants. Thus, existing fine-mapping solutions will not produce reliable results for such cases driven by ungenotyped variants or non-additive effects.

An alternative approach is to consider haplotypes, which we define here as sets of alleles in *cis*. Haplotypes can capture signals due to combinations of variants acting in *cis*[8,18], and are often better able than individual SNPs to tag ungenotyped variants such as rare variants[19], repeats[2], or structural variants[20]. Haplotype tests[21–24] were previously widely used when genotype data was limited to common SNPs and efficient phasing and imputation algorithms were not available, and thus the majority of variants were not directly tested. They declined in popularity after more

complete reference haplotype panels of genetic variation based on next-generation sequencing became available, enabling accurate imputation of most common variants.

Even with the advent of large reference panels or biobank-scale whole genome sequencing, many effects including those driven by rare or complex variants or variant interactions are still not directly captured. Newer haplotype-based methods have been designed to further close this gap. For example, Howey and Cordell proposed using haplotypes composed of two SNPs for association testing, chosen based on their pairwise correlation, and used simulations to show that their approach is well-powered to capture many signals driven by ungenotyped variants[25]. Zheng *et al.* demonstrated that performing finemapping on haplotypes inferred from an external reference panel can improve effect size estimation at low sample sizes or when signals are driven by rare variants[26]. Bkhetan *et al.* constructed haplotypes based on LD between phased SNPs and recovered known signals that could not be detected by single SNPs[27]. Zhang et al used ARGs (ancestral recombination graphs) to infer the associations of unobserved variants and found that ARG-derived variants, which in most cases capture individual rare variants, can reveal additional associations missed by single variant approaches[28]. Although these methods use different strategies to define haplotypes, they all construct haplotypes based on LD information alone. While this has the benefit of obtaining phenotype-agnostic sets of haplotypes, it limits the portability of those haplotypes across training datasets that might have different LD structures. Further, it may not be the optimal strategy to identify haplotypes most relevant to specific phenotypes.

Here, we present happler, a new fine-mapping method which considers effects of haplotypes in addition to individual variants. Happler takes as input individual-level phenotypes and phased genotypes. Unlike previous approaches, it directly uses phenotype data to construct candidate trait-associated haplotypes at each region of interest. Candidate haplotypes are then used

in addition to individual variants as input to an existing fine-mapping method to obtain a probability of causality for each variant and haplotype considered. We benchmarked happler using simulated data from haptools[29], our open-source toolkit for simulation of SNP and haplotype effects. We demonstrate that happler remains sensitive across a variety of simulated effect sizes and allele configurations, whereas existing fine-mapping methods based only on individual variants can often produce inconclusive results, especially when the causal haplotype or variant is untyped. Finally, we apply happler to fine-map expression quantitative trait loci (eQTLs) from the Geuvadis cohort[30] and GWAS loci for platelet count in the All of Us Research Program[31]. We identify specific signals at which haplotypes better explain variation in the trait than individual variants, suggesting that signals in those regions may instead be driven by complex variants, *cis* interaction effects, and rare variants not typically considered in GWAS and downstream analyses.

2.2 Methods

2.2.2 The happler method

We introduce a fine-mapping framework that takes as input a phased set of individual-level genotypes from a VCF or PGEN file of biallelic variants and phenotype values for a trait of interest and outputs posterior inclusion probabilities (PIPs, which quantify the probability of being causal) for each variant and a set of candidate haplotypes identified using the strategies described below (**Figure 3A-B**). Happler consists of two stages: (i) haplotype construction, followed by (ii) fine-mapping, in which haplotypes identified by the first stage are treated as additional variables in an existing fine-mapping tool, such as SuSiE[32] (**Figure 3B**). Happler operates on one locus at a time, considering only those genotypes in the locus surrounding each target lead variant. Typically, between 500kb-1Mb is chosen as the size of the locus, consistent with previous literature[7,33,34].

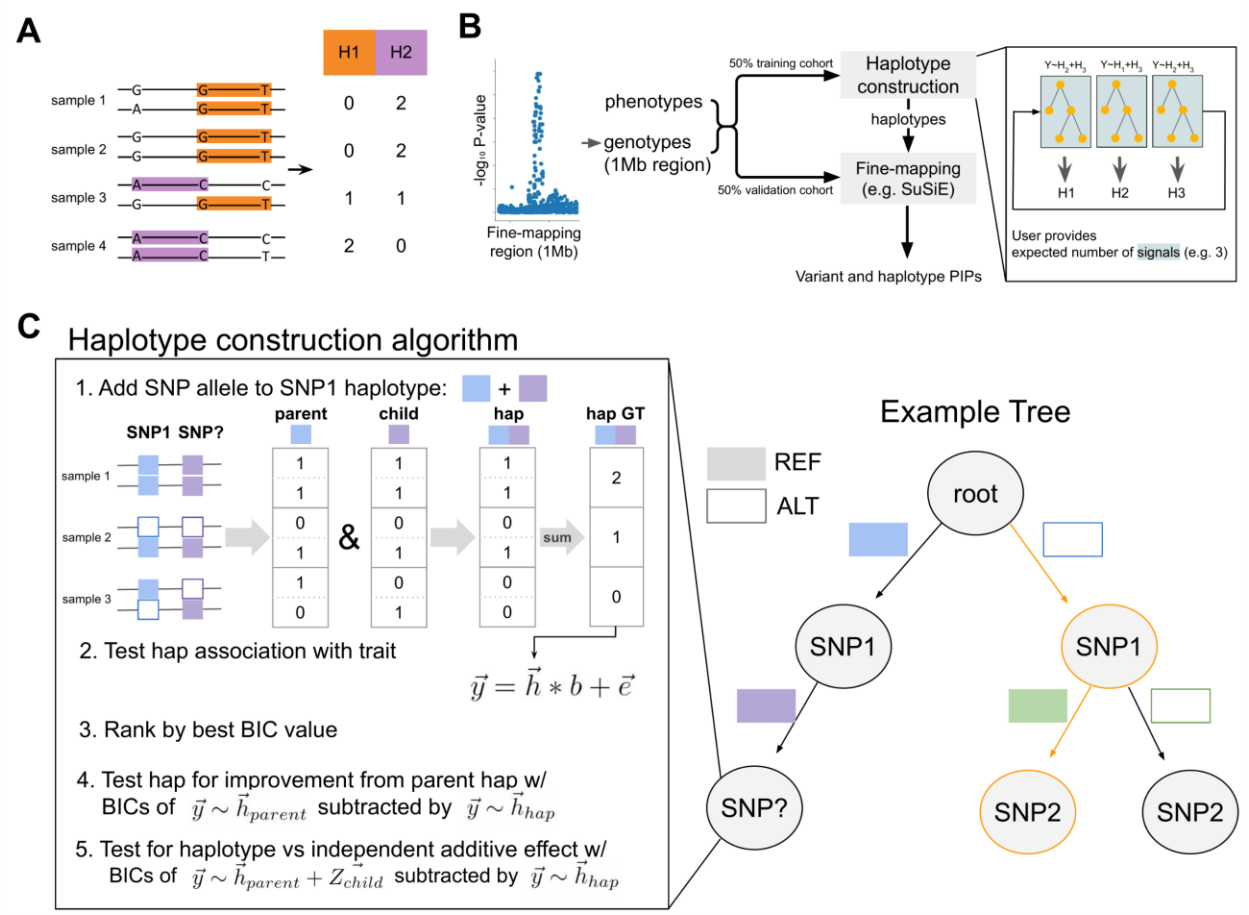


Figure 3. The happler method. A. Encoding haplotypes as individual-level genotypes. We define haplotypes as a set of two or more alleles in *cis*. These can be encoded as psuedo-genotypes by counting the number of times all alleles appear in cis across both chromosome copies. **B. Overview of the happler pipeline.** Happler consists of two stages performed in separate cohorts (50/50): *haplotype construction* identifies candidate trait-associated haplotypes, which are used alongside the input genotypes for individual variants as input to the *fine-mapping* stage. Inset: The method iteratively builds a tree for a user-specified maximum number of independent signals using residuals from regressing out haplotypes from other built trees as phenotypes. It repeats this process for a predefined number of iterations. **C. Overview of haplotype construction.** Within each iteration, the method builds a tree, where nodes of the tree are variants and branches are alleles. A path from the root of the tree to a leaf node (orange) represents a haplotype that would appear in the output of the method. At each node, we test whether to extend the current haplotype by each variant. The best variant is added to the tree if it is determined to improve the effect of the haplotype via a t-test and if it is not determined to be an independent additive signal based on an interaction test.

2.2.2.1 Haplotype construction

At each locus, happler uses phased input genotypes to identify a set of trait-associated *haplotypes*, where each haplotype is composed of a set of alleles at one or more variants that occur

in *cis*. The alleles are greedily chosen to maximize the strength of the association between the haplotype and the target trait. To identify candidate haplotypes, we construct a binary tree composed of linear regression models (**Figure 3B-C**). Each node in the tree represents a variant with branches of a node representing the variant's alleles: REF or ALT. A path from the root of the tree to a target node *V* passes through alleles of one or more variants, which defines the *haplotype* ending at node *V*. Note, a haplotype consisting of only a single node, arising from a tree with only one non-root node, is equivalent to an individual variant.

The algorithm begins by choosing the variant with the strongest trait association to place at the root node of the tree. Association tests are performed using a simple linear model for each variant *j*: $\vec{y} = \beta_j \overrightarrow{Z_J} + \overrightarrow{\epsilon_J}$, where $\vec{y}$ is a vector of individual-level phenotypes, $\overrightarrow{Z_J}$ is a vector of individual-level normalized genotypes, $\overrightarrow{\epsilon_J}$ is a vector of noise terms, and β_j is the unknown effect size of variant *j*. Phenotypes are assumed to already be adjusted for relevant covariates. Genotypes are Z-normalized to have mean 0 and variance 1. The strongest trait association is determined based on the Bayesian Information Criterion (BIC).

Non-root nodes are constructed according to a different strategy. For each node V_{parent} in the tree, we evaluate whether extending the haplotype of that node by each possible variant allele not already in the haplotype will improve the model fit. For each possible haplotype extension, we construct a *haplotype matrix* encoding whether each individual has 0, 1, or 2 copies of the candidate haplotype (**Figure 3C**). The haplotype matrix is analogous to the genotype matrix that is traditionally used for association testing in GWAS and fine-mapping methods. To determine the best allele for the next node V_{child} of the tree, each new haplotype is tested for an association with the phenotype using the model above, except the $\overrightarrow{Z_J}$ term encodes counts from the haplotype matrix rather than variant-level genotypes. Alleles that would reduce the MAF of the haplotype below a

user-defined threshold can optionally be excluded from consideration. The allele from the haplotype with the best p-value is chosen as the candidate allele for V_{child}.

2.2.2.2 Termination criteria

Before extending the haplotype by the best allele identified above, we perform two additional tests. Imagine we are considering whether to add a new allele at a node V_{child} in the tree. The standardized genotypes of the new allele are given by $\overrightarrow{Z_{child}}$, and those of the haplotype at the parent node are given by $\overrightarrow{h_{parent}}$. Then $\overrightarrow{h_{hap}}$ represents the haplotype at V_{child}, which is the extension (logical AND) of $\overrightarrow{h_{parent}}$ and $\overrightarrow{Z_{child}}$. First, to determine whether the addition of the allele to the haplotype improves the model fit compared to the parent haplotype, we compute the delta BIC of the extension, $\Delta BIC_{ext} = BIC(\, y \sim \overrightarrow{h_{parent}}) - BIC(\, y \sim \overrightarrow{h_{hap}})$ (**Figure 3C**). We also compute a separate delta BIC to determine whether the candidate extended haplotype outperforms a model in which the parent haplotype and candidate node have independent effects, $\Delta BIC_{ind} = BIC(\, y \sim \overrightarrow{h_{parent}} + \overrightarrow{Z_{child}}) - BIC(\, y \sim \overrightarrow{h_{hap}})$. If one or both tests do not pass then tree building is terminated. Default values were identified based on simulation analyses described below.

2.2.2.3 Handling multiple independent signals

Many trait-associated regions have been shown to be driven by multiple independent signals[35]. To allow for this case, we perform an iterative procedure, similar to that used in Algorithm 1 of the SuSiE method[32], in which we build multiple trees, each corresponding to a potential independent signal, where the user specifies a maximum number of causal signals (L) for the locus. For i in $1 \ldots L$, we build a separate haplotype tree, using phenotype values for which we have regressed out all candidate haplotypes already discovered by other trees (**Figure 3B**). We repeat this process for a user-specified maximum number of iterations (default 3) and return the union of haplotypes from each tree.

The final set of haplotypes output by each tree is filtered to exclude: those with association p-values below a user-specified threshold (default p=5*10-8), and those consisting of a single variant, since all individual-level variants are in any case provided as input to the fine-mapping stage.

2.2.2.5 Fine-mapping to computing haplotype and variant level PIPs

In the final step of the method, we encode the haplotypes at the leaves of the tree in a haplotype matrix. The resulting matrix is used in addition to variant-level genotypes as input to an existing fine-mapping method to compute a PIP for each haplotype and each variant. Although in theory any fine-mapper can be used, in practice we use SuSiE[32] over FINEMAP[7] because it explores more causal configurations and has demonstrated improvements in speed.

2.2.2.6 Implementation

Happler is implemented as an open-source Python package. All file IO is powered by the haptools[29] data library and all linear associations were performed with statsmodels.regression.linear_model.OLS[36].

2.2.3 Simulation analyses

To benchmark our method, we first use simulated phenotypes. To ensure our simulated data recapitulates realistic LD patterns, we constructed haplotypes from real genotypes in the 1000 Genomes Project Phase 3[37] dataset distributed on the PLINK2[38] website. Only bi-allelic SNPs with alternate allele frequencies between 0.05 and 0.95 were considered. We randomly selected pairs of alleles, each from a separate SNP, to generate haplotypes. The LD between the alleles of each haplotype was computed as the Pearson correlation coefficient. Then, the resulting haplotypes were filtered so that the final set has a roughly even distribution of LD values between the two alleles defining the haplotype, varying at approximate intervals of 0.1 between 0 and 1.

We used haptools[29] v0.6.2 to simulate phenotypes from these haplotypes according to the linear model $\vec{y} = \Sigma_j \beta_j \quad \overrightarrow{Z_j} + \vec{\epsilon}_j$ where $\overrightarrow{Z_j}$ corresponds to the standardized counts of haplotype j for all individuals and the errors are sampled randomly from a normal distribution and assumed to be independent for each sample: $\vec{\epsilon}_j \sim MVN(0, I(1 - \Sigma_j {\beta_j}^2))$. We repeated this process 10 times each for values of β_j of [0.05, 0.1, 0.15, 0.2, 0.25, 0.3]. We evaluated scenarios with between 1 and 4 total causal haplotypes and random sample subsets of [800, 1600, 2400, 3200]. Happler was executed with at most L=3 trees and 3 iterations, and its haplotypes were constructed to have a minimum MAF of 0.05 and maximum p-value of 5e-8.

We measured the performance of happler by computing the LD between each observed haplotype output by happler and its best match among the causal (simulated) haplotypes. Observed haplotypes were uniquely paired to causal haplotypes according to a linear assignment via scipy.optimize.linear_sum_assignment, which implements minimum weight matching in bipartite graphs. This assignment maximizes the total LD between the pairings while ensuring that a causal haplotype is never paired with two or more observed haplotypes. When happler reported more observed haplotypes than the number of true causal haplotypes, we reported the remaining haplotype that could not be paired separately. After haplotype construction, observed haplotypes as well as individual genotypes and phenotypes were used as input to SuSiE v0.14.2 to perform fine-mapping under default settings with L=10.

2.2.4 Parameter tuning and benchmarking

We extended our simulation framework to benchmark and tune thresholds ($\Delta BIC_{ext}, \Delta BIC_{ind}$) for the two model comparison steps used by happler at each node in the tree-building process. We randomly subsetted the 1000 Genomes data[39] to four sets of 800, 1600, 2400, and 3200 samples. Causal haplotypes within each Geuvadis eQTL locus were randomly

constructed from pairs of alleles using the process described above at the lowest sample size of 800. The resulting haplotypes were filtered to ensure an approximately even distribution of LD values between the two alleles defining the haplotype. For this, we assigned haplotypes to 10 equally spaced bins between 0.05 and 0.95 based on the LD of their alleles and chose 10 haplotypes at random from each bin for a total of 100 haplotypes. Thus, each of the 100 haplotypes comes from a distinct locus. We used haptools to simulate phenotypes according to the model described above. Each haplotype was used to simulate phenotypes with effect sizes (β_j) of [0.05, 0.1, 0.15, 0.2, 0.25, 0.3]. This process was repeated five times for each of the four sample subsets.

First, we sought to optimize and evaluate the BIC threshold used when deciding whether to extend a node of the tree. For each simulated phenotype, we computed the BIC value of the model with the haplotype $BIC(y{\sim}h_{hap})$ and the BIC value of the model with only the first allele in the haplotype $BIC(y{\sim}h_{parent})$. The alleles of each haplotype are ordered such that the allele with the best (lowest) BIC value based on individual association tests for each allele appears first. We also extended the haplotype by the next best node in the tree (so the haplotype now has three alleles) and computed the BIC a third time: $BIC(y{\sim}h_{extension})$. For each simulated phenotype, we labeled the ΔBIC of $BIC(y{\sim}h_{parent}) - BIC(y{\sim}h_{hap})$ as a "positive" and the ΔBIC of $BIC(y{\sim}h_{extension}) - BIC(y{\sim}h_{hap})$ as a "negative." We evaluated the ability of the ΔBIC_{ext} score to classify between positives and negatives by computing the performance metrics AUROC and average precision.

We used a similar process to evaluate the performance of the test for an independent association. For this, the positives corresponded with simulations in which the haplotype was simulated as causal $y{\sim}h_{hap}$, while the negatives represented simulations in which the parent of

the haplotype and the haplotype's last allele were additively causal as in $y \sim h_{parent} + Z_{child}$. For both cases, we used the ΔBIC_{ind} of those two models, $BIC(y \sim h_{parent} + Z_{child}) - BIC(y \sim h_{hap})$, as the scoring metric. We evaluated the ability of this score to classify between the positive and negative cases by computing the same performance metrics.

To choose appropriate thresholds for both ΔBIC_{ext} and ΔBIC_{ind}, for each simulation setting we determined ΔBIC_{ext} or ΔBIC_{ind} thresholds that result in a false discovery rate (FDR) of at most 5%. We found optimal thresholds in each setting are relatively robust to changes in sample size, with ΔBIC_{ext} thresholds mostly below 18 and ΔBIC_{ind} all below 14. We therefore defined a default setting of $\Delta BIC_{ext} > 20$ and $\Delta BIC_{ind} > 15$. After executing happler, we also recompute a final value of ΔBIC_{ext} ($\Delta BIC_{ext,final}$) against the original phenotypes using a separate script, which differs from the ΔBIC_{ext} values computed during each iteration, which are based on phenotypes that have been adjusted by candidate signals arising from other trees in the forest. Overall, we found that $\Delta BIC_{ext,final}$ of the output haplotype is a useful quality metric, with high $\Delta BIC_{ext,final}$ values indicating the highest confidence haplotypes.

2.2.5 Geuvadis data and preprocessing

Previously computed expression quantitative trait loci (eQTL) summary statistics based on 359 individuals of European ancestry were obtained from the Geuvadis gene expression project at https://ftp.ebi.ac.uk/biostudies/fire/E-GEUV-/001/E-GEUV-1/Files/E-GEUV-1/analysis_results/EUR373.gene.cis.FDR5.best.rs137.txt.gz[40]. We lifted over 2,440 eQTLs identified at FDR 5% corresponding to 12,606 unique genes from hg19 to hg38 using the UCSC liftover[41] tool v1.3.1 with chain file https://hgdownload.soe.ucsc.edu/gbdb/hg19/liftOver/hg19ToHg38.over.chain.gz. Individual-level gene expression values were obtained from https://www.ebi.ac.uk/biostudies/files/E-GEUV-

1/E-GEUV-1/analysis_results/GD660.GeneQuantRPKM.txt.gz[40]. Individual-level phased genotypes from the 1000 Genomes Project[42] corresponding to these individuals were obtained from https://www.internationalgenome.org/data-portal/data-collection/30x-grch38. Prior to fine-mapping, we regressed out the following covariates from gene expression values: 36 PEER factors[43], sex, and the first 5 genotype PCs, where the PCs were computed as part of a previous study[44]. We fine-mapped 1 Mbp regions centered around each eQTL. For downstream processing, we used PLINK2 v2.00a5.10 to extract only biallelic SNPs with a minimum minor allele frequency (MAF) of 0.1. Happler was executed with at most L=3 trees and 3 iterations, and its haplotypes were constructed to have a minimum MAF of 0.1 and maximum p-value of 5e-6. SuSiE v0.11.42 was used for the fine-mapping step with L=10.

We computed the R^2 of the model fit via statsmodels.regression.linear_model.OLSResults.rsquared[36] for each candidate haplotype based on a linear model $\vec{y} = \beta_h \ \overrightarrow{Z_h} + \vec{\epsilon_j}$ with standardized haplotype counts $\overrightarrow{Z_h}$ and phenotypes $\vec{y}$. For comparison we built a second model consisting of additive and independent contributions of the alleles defining each haplotype: $\vec{y} = \sum_{i=1..h} \beta_i \overrightarrow{Z_i} + \vec{\epsilon_j}$ where $\overrightarrow{Z_i}$ represents the genotypes for variants corresponding to the *i*th allele used to construct the haplotype and *h* is the total number of alleles in the haplotype (typically two, but can in some cases be more than two). The R^2 of the model fit was computed in the same way as described above.

2.2.6 GTEx data and preprocessing

We utilized available gene expression and genotype data from the Genotype-Tissue Expression (GTEx) Project[45] v8 available through dbGaP accession number phs000424.v8.p2 to validate associations of haplotypes with gene expression identified in Geuvadis. We focused on 131 samples with expression data available for EBV Transformed Lymphocytes. SNP Genotypes

used in this study were from dbgap (file GTEx_Analysis_2017-06-05_v8_WholeGenomeSeq_866Indiv.vcf.gz), and filtered using PLINK2 with the following parameters: –hwe 0.001 and –geno 0. We then performed principal component analysis (PCA) of available SNP genotypes using PLINK2[38] to infer the set of samples with predicted European. Based on visual inspection of the first two genotype PCs, we classified a sample as European if PC1>-0.014 and PC2 > 0.01 resulting in 109 European samples.

Inverse normal transformed expression values for each gene were obtained from the GTEx website (GTEx_Analysis_v8_eQTL_expression_matrices.tar). We computed PEER factors using PEER[43] v1.0. For PEER factor computations, we further removed genes with TPM<1 in more than 90% of individuals. Prior to downstream analyses we regressed out the top 5 genotype principal components (PCs), sex, and 11 PEER factors. For each candidate haplotype-gene pair, we computed the R^2 of the model fit using an additive model including both of the individual SNPs contributing to each haplotype as well as the haplotype itself.

2.2.7 All of Us data and preprocessing

We applied happler to fine-map GWAS signals for platelet count identified in the All of Us Research Program[31] v8 dataset. We chose this trait since it has been measured in a large number of samples, is highly polygenic[46], and we have already performed fine-mapping for this trait previously to identify several high confidence candidate causal variants[47].

Prior to downstream analysis, we removed samples whose value for sequencing metrics ins_del_ratio, snp_het_homvar_ratio, indel_het_homvar_ratio, ti_tv_ratio, or singleton metrics falls outside 8 standard deviations of the means across all samples. We also subsetted to only the samples that were included within AoU's list of unrelated individuals and which had passed sex concordance checks performed by AoU. We restricted to samples which had self-identified race

as White and had genetically inferred ancestry as 'eur' according to AoU's ancestry prediction file. We focused on this cohort since it had the largest available sample size.

We then obtained data corresponding to concept ID 3024929 (“Platelets [#/volume] in Blood by Automated count”). Data points with units “thousand per microliter”, “per cubic millimeter”, or “/mm3” were kept. We additionally kept data points with “no value” or "No matching concept" listed as the unit since these followed a similar distribution to the retained measurements.

After restricting to data points from individuals passing sample preprocessing above, we filtered values more than 5 standard deviations away from the mean. We then obtained the median phenotype value per year that measurements were available, then took the median of medians across all years. We recorded the year corresponding to the final value used as the age covariate for downstream association testing. After preprocessing, a total of 105,446 samples remained for analysis.

We applied fine-mapping to regions we previously identified in Supplementary Table 2 of Mirmira *et al.*[47] for platelet count in AoU. Before running the happler workflow, we regressed out all covariates (first 10 PCs, age, and sex at birth) from the normalized trait values. We ran happler with the phased short-read SNP/indel VCF files provided by AoU. Using PLINK2[38], we removed multiallelic variants and variants with HWE less than 1e-100. Happler was run with the following parameters: --hap-maf 0.0001 --max-signals 3 --max-iterations 3 --discard-multiallelic --remove-SNPs --indep-thresh 15 -t 20 --chunk-size 500 --out-thresh 5e-08. We ran SuSiE with L=10 (maximum number of independent signals) and all other parameters set to their default values. As input, we used normalized phenotype values and individual variant genotypes.

2.3 Results

2.3.1 Benchmarking happler using simulated data

We first evaluated happler under a range of simulation settings in which we simulated phenotypes assuming a single causal haplotype and varied the simulated effect size (range 0.05-0.3), sample size (range 800-3200), and LD between alleles (range 0-1) of each causal haplotype. In each case, samples were randomly split 50/50 into training (haplotype construction) and validation (fine-mapping) cohorts. We first measured the performance of the haplotype construction step of happler by computing the absolute value of the Pearson correlation between the best output (observed) haplotype and causal (simulated) haplotypes. As expected, the correlation with the causal haplotype tends to increase at higher effect sizes and larger sample sizes (**Figure 4A**).

We found that the correlation of the output haplotype with the causal haplotype tends to increase with the LD between alleles on the causal haplotype (**Figure 4B**). In cases for which the alleles of the haplotype are in high LD, happler tends to output haplotypes containing only a single allele, which is already a strong proxy for the causal haplotype. Haplotypes for which the alleles are in modest LD showed lower correlation, in most cases because the termination tests (**Methods**) required to extend the haplotype to the second allele did not achieve the required statistical significance. As expected, this trend improves with increasing sample size (**Figure 4B**) which results in improved power for both tests.

We next evaluated happler's performance in cases of multiple, independent causal haplotypes and tested the effect of the input parameter L specifying the maximum number of signals to identify. We simulated between 1-4 causal haplotypes per locus while allowing happler to build L=3 trees. In cases where the number of causal haplotypes was less than or equal

to L, haplotype construction identified haplotypes that were strongly correlated with each of the input haplotypes (**Figure 4C**), indicating the iterative procedure is correctly identifying independent signals. As expected, when the number of trees built was higher than the number of simulated signals, not all output haplotypes were matched to a causal haplotype. Further, when simulating 4 causal haplotypes but setting L=3, only three matching haplotypes were identified. Observed haplotypes that do not have a match among the causal haplotypes are assigned a lower PIP (mean of 0.07330.00465) than the matching haplotypes (mean of 0.26630.00379).

We then evaluated the performance of the *fine-mapping step* of happler, which uses variant-level genotypes in addition to candidate haplotypes output during the haplotype construction step as input to an existing fine-mapping tool (in this case, SuSiE21). This step allows us to obtain PIPs for the haplotypes and assess whether the haplotypes are better candidates for the causal variant than any other individual variant (**Figure 3B**). We assessed performance of the fine-mapping step on the original (training) cohort vs. on an independent validation cohort. For this, we performed additional simulations in which the true causal variant is an individual SNP, rather than a haplotype. We then performed haplotype construction on the training cohort to obtain the best candidate haplotypes. Finally we evaluated whether the PIP of a simulated causal SNP would be affected by inclusion of a non-causal haplotype. The PIP remains unaffected in the validation cohort, whereas fine-mapping in the training cohort incorrectly assigns lower PIPs to the truly causal SNP, often incorrectly assigning the non-causal haplotype a high PIP. In both cohorts, the rate at which haplotypes were falsely identified as causal (at a PIP threshold of 0.5) is lower at higher delta BIC thresholds and is most attenuated when the fine-mapping step is performed in a separate validation cohort. Taken together, these results indicate that using an orthogonal

validation cohort for fine-mapping is critical to avoid assigning non-causal haplotypes high PIPs due to over-fitting.

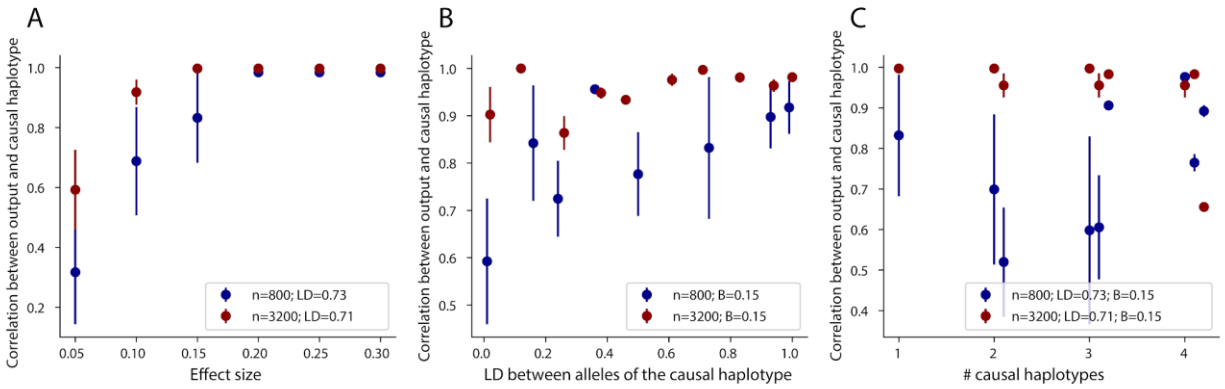


Figure 4. Benchmarking happler's haplotype construction step using simulations. We simulated traits based on causal 2-SNP haplotypes with a range of effect sizes, sample sizes, and LD values between the two SNPs of each haplotype. We used happler to construct candidate haplotypes for each simulated trait and evaluated performance by computing the correlation between output haplotypes and simulated causal haplotypes. For **A-C**, results are shown for two sample sizes: 800 (dark blue) and 3200 (dark red). In all cases dots give the mean and error bars denote +/- 1 standard error based on 10 replicates. In each of the 10 replicates the simulated causal haplotype is the same but the random seed is changed. **A. Correlation between the output and causal haplotype vs. effect size.** The x-axis gives the simulated effect size and the y-axis gives the correlation with the causal haplotype. The legend indicates the LD between the alleles of the haplotype for the two simulated sample sizes. **B. Correlation between the output and causal haplotype vs. the LD between the haplotype alleles.** The x-axis gives the LD between the two alleles and the y-axis gives the correlation of the output with the causal haplotype. For both sample sizes, results are shown for simulations with an effect size of 0.15. **C. Correlation between output and causal haplotypes in the case of multiple causal haplotypes.** We simulated between 1-4 causal haplotypes but performed haplotype construction allowing for a maximum of 3 signals. Output haplotypes were matched with simulated haplotypes using a procedure described in **Methods** and the plot shows the correlation with the matched causal haplotype for each output haplotype.

After establishing the need to perform fine-mapping in an independent cohort, we next assessed the performance of fine-mapping in this cohort using simulations described above in which a haplotype is the true causal variant. We found that as expected the PIP of candidate haplotypes increases with effect size. Further, fine-mapping could better identify the haplotype as causal in cases where the LD between the haplotype's alleles is modest (**Figure 5A**), since in the case of high LD the individual alleles of the haplotype are in high LD with the haplotype itself,

making it difficult for fine-mapping to distinguish between them. For comparison, we performed traditional fine-mapping using SuSiE with only individual variants as input. As expected, in cases of modest LD between the haplotype alleles, happler correctly identifies the candidate haplotype as causal with high posterior probability (**Figure 5B**), whereas SuSiE based on variants alone is not able to identify a strong candidate causal variant (**Figure 5C**).

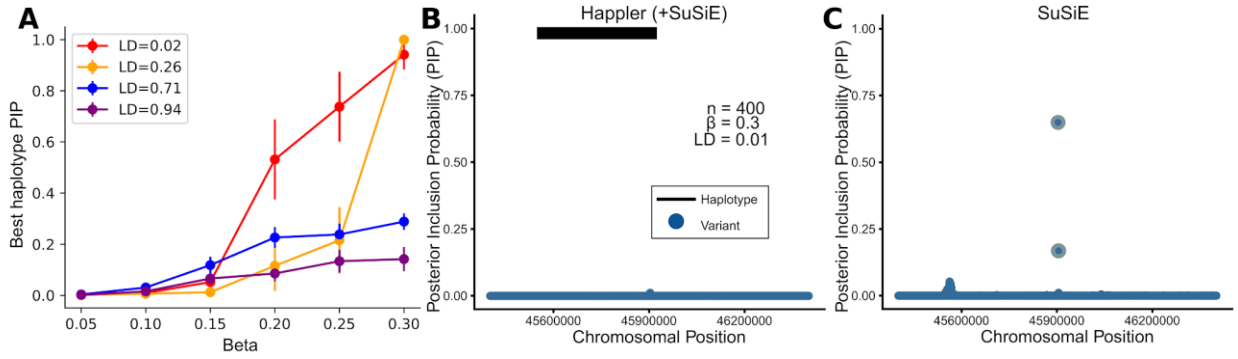


Figure 5. Benchmarking haplotype-based fine-mapping with simulations. Observed haplotypes identified during the haplotype construction step, plus individual variants and simulated phenotypes described above were input to SuSiE for fine-mapping. Plots show the case of a single simulated causal haplotype with n=777 samples. **A. Effect size vs. posterior inclusion probability (PIP).** The x-axis represents the simulated effect size and the y-axis represents the best PIP obtained for an observed haplotype. Dots give the mean across 10 simulations and error bars represent +/- 1 standard error. Colors denote the LD between the two alleles of the simulated haplotype. **B. Happler PIPs for all variants in an example region.** PIPs are shown for all individual variants (blue) and the observed haplotype (black bar spanning the location of the two alleles defining the haplotype) in an example fine-mapped region with simulation parameters LD between alleles=0.06, n=777, Effect size=0.15. **C. SuSiE PIPs for all variants in the same example region.** The plot is the same as in **B**, but based on running SuSiE with only variants (no haplotypes) as input. In this case SuSiE fails to identify a high confidence causal variant.

2.3.2 Haplotype-based fine-mapping eQTLs in the Geuvadis cohort

We applied happler to fine-map eQTL signals identified in lymphoblastoid cell lines (LCLs) for 359 European samples for which RNA-sequencing is available from the Geuvadis[30] cohort and genotypes are available from the 1000 Genomes Project[42]. We focused on 2,406 genes for which significant eQTLs were previously identified and performed fine-mapping in a 1Mb region centered on each lead SNP. Haplotype construction identified candidate haplotypes at a

total of 35 eQTL loci under default settings. Of the 35 candidate haplotypes, 34 haplotypes consisted of two alleles, and a single haplotype contained 3 alleles..

Next, we evaluated candidate haplotypes in an independent cohort using available LCL expression data and matched genotypes available for 109 samples of predicted European ancestry from the GTEx Project[45]. Of the 35 genes with candidate haplotypes, 33 were present and could be assessed in GTEx. Effect sizes of candidate haplotypes are highly correlated across datasets (Pearson R^2=0.90; two-sided P=7.26e-13; n=33). We observed that the majority of haplotypes with $\Delta BIC_{ext,final} > 25$ exhibited a higher model R^2 than their individual alleles in GTEx (14/19, 73.7%), indicating that the haplotype better explains the signal compared to its component alleles even in an independent cohort. Similarly, haplotype frequencies and the LD between alleles in each haplotype were highly consistent across datasets.

To further validate our haplotypes, we used the GTEx cohort to perform fine-mapping for each haplotype identified in Geuvadis. We provided SuSiE with both the haplotypes and the individual variants as input (**Figure 6A**). Of the 14 total haplotypes identified above exhibiting a higher model R^2 than their component alleles, five haplotypes obtained PIPs above 0.9. All other tested haplotypes did not achieve PIPs above this threshold. These haplotypes were associated with the expression of the following genes: *PAAF1*, *TBX6*, *ERAP1*, *PIP5K1C*, and *CHI3L2*. As an example, we depict plots of variant and haplotype associations with *PAAF1* (**Figure 6B-C**), a gene linked to the regulation of proteasome function. Happler identified a candidate fine-mapped haplotype in the *PAAF1* locus spanning chr11:73935581-73910990 in GRCh38, defined by the CA allele of 11:73910990:CA:C and the TTTC allele of 11:73935581:T:TTTC. To investigate the contribution of this haplotype to the overall eQTL signal in the locus, we plotted eQTL summary statistics for the locus and conditioned on either the haplotype itself, or one or both of its

component alleles. We saw that the original signal (**Figure 6B**, left panel) is not reduced after conditioning on the individual alleles of each haplotype (**Figure 6B**, second and third panels). We observe an intermediate reduction when conditioning on each allele of the haplotype together (**Figure 6B**, fourth panel). On the other hand, the signal is completely removed when conditioning on the haplotype itself (**Figure 6B**, fifth panel). The haplotype in this locus also exhibited a stronger association than its individual alleles in GTEx (haplotype: p=2.26e-07, 11:73910990:CA:C: p=0.66, 11:73935581:T:TTTC: p=9.75e-04) (**Figure 4C**). Overall this is consistent with the locus being best explained by the *combination* of alleles on the haplotype rather than by independent contributions of the two alleles defining the haplotype.

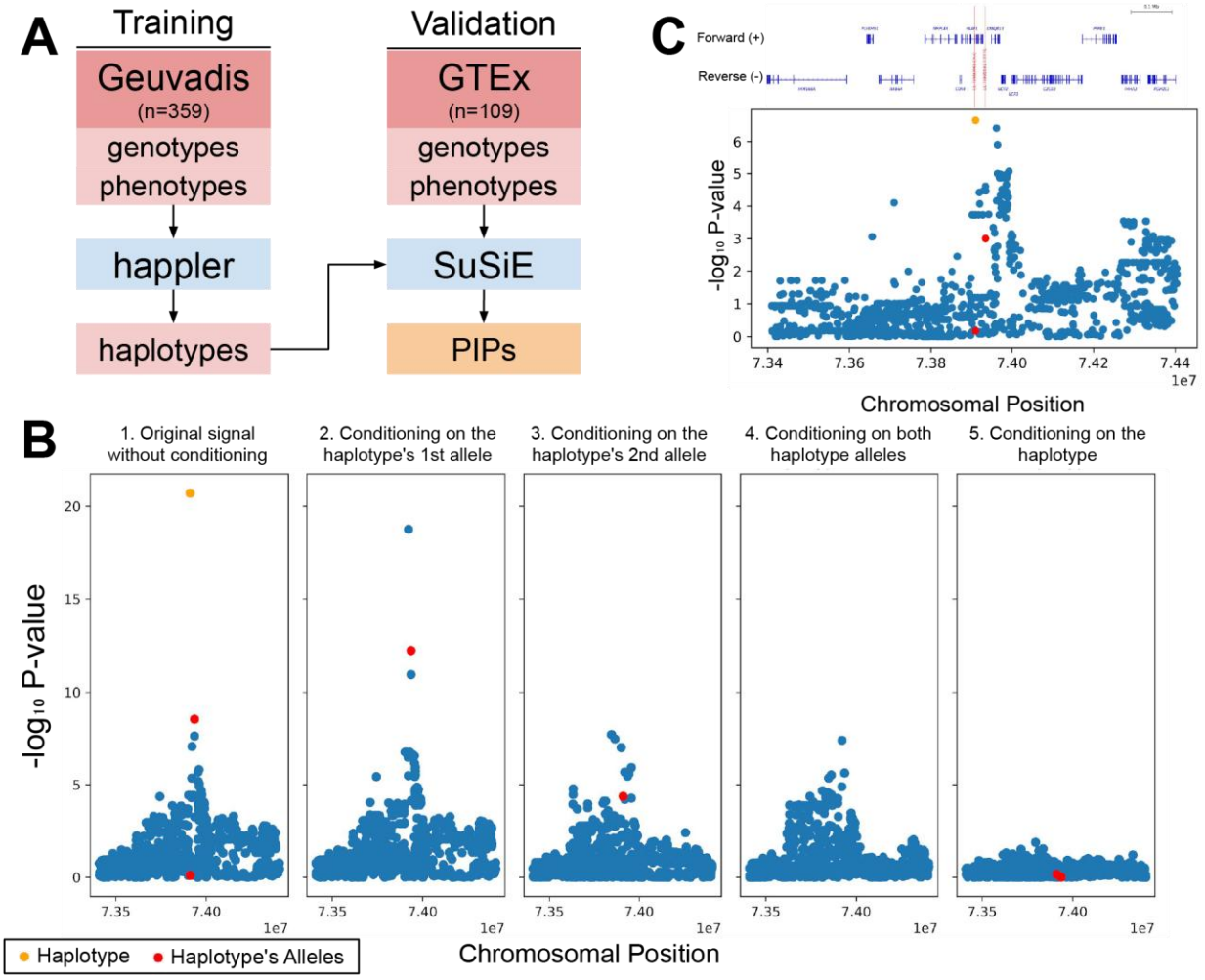


Figure 6. An application of happler to fine-map eQTLs in the Geuvadis cohort. We applied happler to 2440 eQTL loci in the Geuvadis dataset and identified 35 haplotypes associated with gene expression. **A. Schematic of Geuvadis analysis.** We used the Geuvadis cohort (n=359) for training and discovery of the haplotypes. We computed the PIPs of each haplotype by fine-mapping them alongside individual variants with SuSiE in a GTEx validation cohort (n=109). **B. Summary statistics for an example locus in Geuvadis when conditioning on a haplotype vs its alleles.** The haplotype identified consists of alleles from hg38:chr11:73935581 (T/TTTC) and hg38:chr11:73910990 (CA/C) and was fine-mapped as an eQTL for the gene *PAAF1*. The original eQTL signal in Geuvadis (**B,** first panel) is not reduced compared with conditioning on each of the haplotype's individual alleles (**B**, second and third panels) or on both alleles together (**B**, fourth panel). However, the signal is completely removed when conditioning on the haplotype in the locus (**B**, fifth panel). In each plot, blue dots show association statistics for individual variants. Red dots show association statistics of the individual alleles of the haplotype. The orange dot represents the haplotype itself. **C. Summary statistics for the example locus in GTEx.** The same haplotype in GTEx has a stronger association (p=2.26e-07, PIP=0.985) than any individual allele in the 1 Mbp locus.

2.3.3 Haplotype tagging of rare causal variants in AoU

We sought to evaluate the ability of our haplotypes to tag rare causal variants in the AoU research cohort v8 (n=105446). A previous study by Mirmira et al.[47] had found causal signals associated with platelet count in non-Africans, with many candidate causal variants in those signals having MAFs below 0.05. For each of 201 platelet-count GWAS loci identified by Mirmira et al, we ran SuSiE at an MAC of 20 and labeled the SNP with the highest PIP > 0.9 as a candidate causal variant (**Figure 7A**). Happler was then executed repeatedly at each locus with a delta BIC of 20 at increasing MAF thresholds of 0.01 to 0.1 (**Figure 7A**). Loci without a haplotype or a candidate causal variant with MAF < 0.1 were discarded. Across all MAF thresholds, our haplotypes generally exhibited a higher LD with the candidate causal variant than the best tagging biallelic variant, especially once the causal variant was filtered out (**Figure 7B-C**).

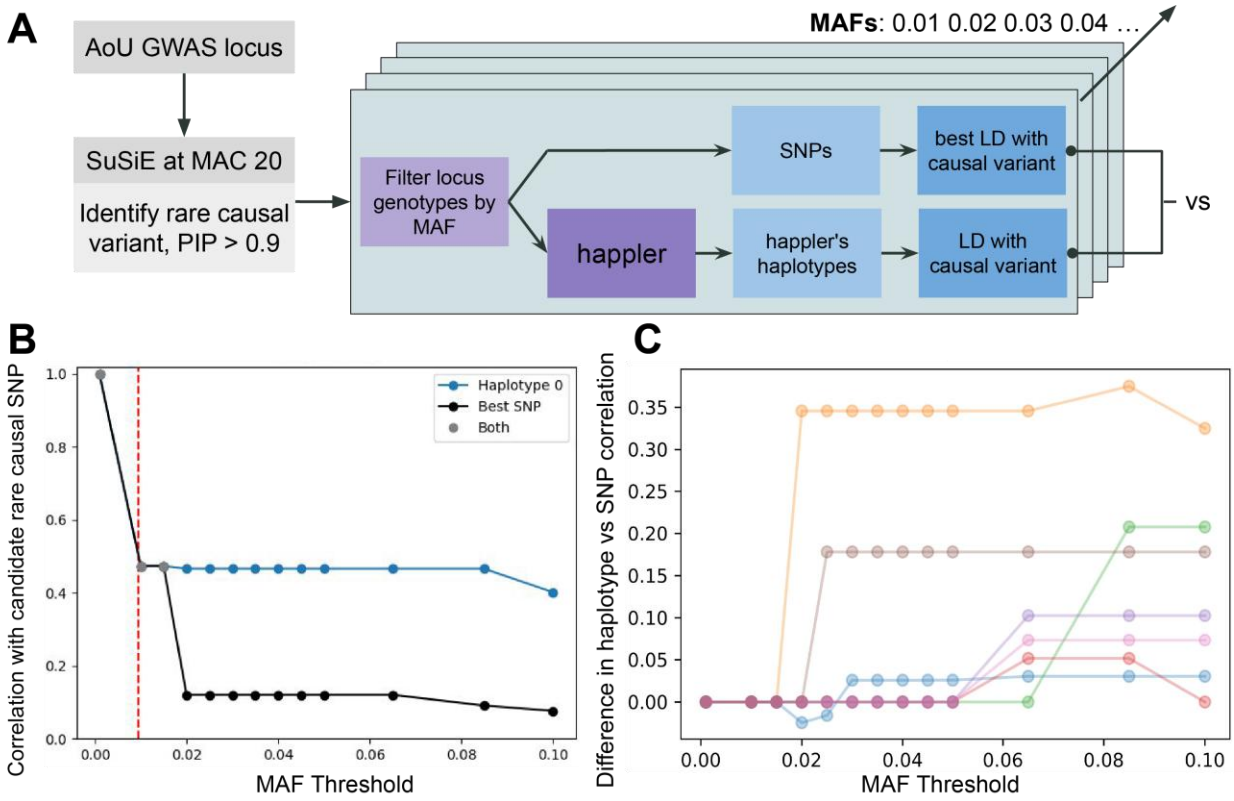


Figure 7. Tagging rare causal variants in All of Us. We applied happler to 201 platelet count GWAS loci in the All of Us v8 dataset and identified 7 rare candidate causal, biallelic variants associated in non-Africans. **A. Analysis flowchart.** A candidate rare causal variant was identified for each locus via SuSiE at an MAC threshold of 20 and PIP > 0.9. At increasing MAF thresholds, we ran happler and then computed the correlation between happler's haplotypes and the original rare candidate causal variant (17:44528237. We compared this to the correlation with the best biallelic variant above the MAF threshold. **B. Example locus 17:43860386-44869051.** The haplotype correlates more strongly with the rare causal variant than the best biallelic variant once the rare causal variant is filtered by MAF (dashed/red line). **C. Difference in correlation between the haplotype and the best SNP.** At high MAF thresholds, the haplotype has a stronger correlation with the rare causal variant than the best SNP.

2.4 Discussion

Here we presented happler, a novel haplotype-based fine-mapping method. Compared to existing fine-mapping strategies, happler can identify both individual variants as well as haplotypes, or combinations of variants, that may harbor the causal drivers underlying association signals. We benchmark happler using simulations of haplotype effects and find that in settings with sufficient sample size it can accurately identify the simulated causal haplotypes.

An exception is cases where the variants of the haplotype are in strong LD, in which case the haplotype effect cannot be distinguished from the effect of the individual variants. We further demonstrate that performing traditional variant-level fine-mapping can result in inconclusive or incorrect results when the truly causal haplotype is not included in the input set of variants. Finally, we demonstrate the utility of happler on real data by fine-mapping eQTL signals from the Geuvadis cohort, which identified multiple signals that are better explained by haplotypes rather than individual variants.

The happler method faces several limitations. First, we take a greedy approach to identify candidate haplotypes, in which we start by identifying the variant with the strongest trait association and incrementally extend the haplotype by additional variants providing the strongest haplotype signal. This is a heuristic strategy to avoid exploring the entire $|2^n|$ search space of *n*-variant haplotypes. Our simulations show this strategy works in most cases, but this approach is likely not optimal. Second, our haplotype construction step applies lenient thresholds on the termination tests described to identify significant haplotypes, and does not account for the large number of hypothesis tests performed throughout tree building. However, we have found that we can tolerate a high false positive rate during haplotype construction, since non-causal haplotypes identified by this stage typically are assigned low PIPs during the fine-mapping stage.

Finally, a major remaining challenge is interpretation of the haplotype effects identified. Importantly, the individual variants defining the haplotypes identified by happler are in most cases probably not causal themselves. Rather, candidate haplotypes are likely in strong LD with something else that is causal. For example, an identified haplotype could tag an unobserved rare variant or structural variant. Alternatively, the haplotype could harbor combinations of variants that exhibit *cis* interaction effects, such as combinations of protein coding variants that

individually have no effect but together alter protein function. Additional work is required to understand the biological impact of the haplotypes identified from real data.

Our current study focused on analysis of eQTL data in a relatively small cohort (n=359), a sample size our simulations suggest is underpowered to identify most haplotype signals. In future studies, we aim to apply happler at biobank scale to identify haplotypes that may underlie GWAS signals identified for a wide range of complex traits in humans. Overall, we envision this will help identify signals that may be driven by an underlying causal variant that was not directly included in the GWAS and downstream analyses, and will ultimately lead to new insights into the biology of complex traits.

2.5 Code availability

Happler is available at: https://github.com/gymrek-lab/happler

2.6 Acknowledgements

This work was partially supported by NIH/NHGRI grant RM1HG011558 (M.G.). A.R.M. was also supported by the National Science Foundation Graduate Research Fellowship [DGE-2038238] and grants from the National Institutes of Health [T32GM008806 and T32GM139790].

Chapter 2, in part is currently being prepared for submission for publication of the material. "Haplotype-based fine-mapping of variant associations for complex traits." Arya R Massarat, Michael Lamkin, Tara Mirmira, Utkarsh Jain, Jonathan Margoliash, Yang Li, Melissa Gymrek. The dissertation author was the primary researcher and author of this material.

REFERENCES

1. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).

2. Margoliash, J., Fuchs, S., Li, Y., Zhang, X., Massarat, A., Goren, A. & Gymrek, M. Polymorphic short tandem repeats make widespread contributions to blood and serum traits. *Cell Genomics* **3**, 100458 (2023).

3. Saint Pierre, A. & Génin, E. How important are rare variants in common disease? *Brief. Funct. Genomics* **13**, 353–361 (2014).

4. Mukamel, R. E., Handsaker, R. E., Sherman, M. A., Barton, A. R., Zheng, Y., McCarroll, S. A. & Loh, P.-R. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science* **373**, 1499–1505 (2021).

5. Saini, S., Mitra, I., Mousavi, N., Fotsing, S. F. & Gymrek, M. A reference haplotype panel for genome-wide imputation of short tandem repeats. *Nat. Commun.* **9**, 4397 (2018).

6. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

7. Benner, C., Spencer, C. C. A., Havulinna, A. S., Salomaa, V., Ripatti, S. & Pirinen, M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

8. Williams, A. L., Jacobs, S. B. R., Moreno-Macías, H., Huerta-Chagoya, A., Churchhouse, C., Márquez-Luna, C., García-Ortíz, H., José Gómez-Vázquez, M., Burtt, N. P., Aguilar-Salinas, C. A., González-Villalpando, C., Florez, J. C., Orozco, L., Haiman, C. A., Tusié-Luna, T., Altshuler, D., Williams, A. L., Márquez-Luna, C., Huerta-Chagoya, A.,

Ripke, S., José Gómez-Vázquez, M., Manning, A. K., Moreno-Macías, H., García-Ortíz, H., Neale, B., Burtt, N. P., Aguilar-Salinas, C. A., Reich, D., Stram, D. O., Fernández-López, J. C., Romero-Hidalgo, S., Altshuler, D., Florez, J. C., Tusié-Luna, T., Patterson, N., Haiman, C. A., Aguilar-Delfín, I., Martínez-Hernández, A., Centeno-Cruz, F., Mendoza-Caamal, E., Revilla-Monsalve, C., Islas-Andrade, S., Córdova, E., Rodríguez-Arellano, E., Soberón, X., Orozco, L., Florez, J. C., González-Villalpando, C., González-Villalpando, M. E., Haiman, C. A., Henderson, B. E., Monroe, K., Wilkens, L., Kolonel, L. N., Le Marchand, L., Riba, L., Ordóñez-Sánchez, M. L., Rodríguez-Guillén, R., Cruz-Bautista, I., Rodríguez-Torres, M., Muñoz-Hernández, L. L., Sáenz, T., Gómez, D., Alvirde, U., Burtt, N. P., Onofrio, R. C., Brodeur, W. M., Gage, D., Murphy, J., Franklin, J., Mahan, S., Ardlie, K., Crenshaw, A. T., Winckler, W., Prüfer, K., Shunkov, M. V., Sawyer, S., Stenzel, U., Kelso, J., Lek, M., Sankararaman, S., Williams, A. L., Patterson, N., MacArthur, D. G., Reich, D., Derevianko, A. P., Pääbo, S., Jacobs, S. B. R., Churchhouse, C., Gopal, S., Grammatikos, J. A., Smith, I. C., Bullock, K. H., Deik, A. A., Souza, A. L., Pierce, K. A., Clish, C. B., Altshuler, D., Fennell, T., Farjoun, Y., Genomics Platform*, B., Gabriel, S., Stram, D. O., Gross, M. D., Pereira, M. A., Seielstad, M., Koh, W.-P., Tai, E.-S., Flannick, J., Fontanillas, P., Morris, A., Teslovich, T. M., Burtt, N. P., Atzmon, G., Blangero, J., Bowden, D. W., Chambers, J., Shin Cho, Y., Duggirala, R., Glaser, B., Hanis, C., Kooner, J., Laakso, M., Lee, J.-Y., Tai, E.-S., Ying Teo, Y., Wilson, J. G., Haiman, C. A., Henderson, B. E., Monroe, K., Wilkens, L., Kolonel, L. N., Le Marchand, L., Puppala, S., Farook, V. S., Thameem, F., Abboud, H. E., DeFronzo, R. A., Jenkinson, C. P., Lehman, D. M., Curran, J. E., Blangero, J., Duggirala, R., Burtt, N. P., Cortes, M. L., Altshuler, D., Florez, J. C., Haiman, C. A., Henderson, B. E., Aguilar-Salinas, C. A., González-Villalpando, C., Orozco, L., Tusié-Luna, T., The SIGMA

Type 2 Diabetes Consortium, Writing team, Analysis team, Clinical research, study design and metabolic phenotyping: D. in M. S., Massachusetts General Hospital, Mexico City Diabetes Study, Multiethnic Cohort, UNAM/INCMNSZ Diabetes Study, Sample quality control and whole-genome genotyping, Neanderthal analysis team, Functional analysis and metabolite profiling, Replication genotyping and analysis: Broad Institute of Harvard and MIT, Singapore Chinese Health Study, T2D-GENES Consortium, Texas Biomedical Research Institute and University of Texas Health Science Center at San Antonio, Scientific and project management, & Steering committee. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97–101 (2014).

9. Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, A. D., Haines, J. L. & Pericak-Vance, M. A. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921–923 (1993).

10. Atkinson, E. G. & others. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat Genet* **53**, 195–204 (2021).

11. Naslavsky, M. S., Suemoto, C. K., Brito, L. A., Scliar, M. O., Ferretti-Rebustini, R. E., Rodriguez, R. D., Leite, R. E. P., Araujo, N. M., Borda, V., Tarazona-Santos, E., Jacob-Filho, W., Pasqualucci, C., Nitrini, R., Yaffe, K., Zatz, M. & Grinberg, L. T. Global and local ancestry modulate APOE association with Alzheimer's neuropathology and cognitive outcomes in an admixed sample. *Mol Psychiatry* **27**, 4800–4808 (2022).

12. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719 (2011).

13. Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. & Lee, J. J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, (2015).

14. Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., Ramachandran, S., Stefancsik, R., Stewart, J., Whetzel, P., Wilson, R., Hindorff, L., Cunningham, F., Lambert, S. A., Inouye, M., Parkinson, H. & Harris, L. W. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977 (2022).

15. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).

16. Mukamel, R. E., Handsaker, R. E., Sherman, M. A., Barton, A. R., Zheng, Y., McCarroll, S. A. & Loh, P.-R. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science* **373**, 1499–1505 (2021).

17. Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., Genovese, G., Rose, S. A., Handsaker, R. E., Daly, M. J., Carroll, M. C., Stevens, B. & McCarroll, S. A. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).

18. Lescai, F., Chiamenti, A. M., Codemo, A., Pirazzini, C., D'Agostino, G., Ruaro, C., Ghidoni, R., Benussi, L., Galimberti, D., Esposito, F., Marchegiani, F., Cardelli, M., Olivieri, F., Nacmias, B., Sorbi, S., Tagliavini, F., Albani, D., Martinelli Boneschi, F., Binetti, G., Santoro, A., Forloni, G., Scarpini, E., Crepaldi, G., Gabelli, C. & Franceschi, C. An APOE

haplotype associated with decreased ε4 expression increases the risk of late onset Alzheimer's disease. *J. Alzheimers Dis. JAD* **24**, 235–245 (2011).

19. Wang, M. & Lin, S. Detecting associations of rare variants with common diseases: collapsing or haplotyping? *Brief. Bioinform.* **16**, 759–768 (2015).

20. Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., Genovese, G., Rose, S. A., Handsaker, R. E., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Daly, M. J., Carroll, M. C., Stevens, B. & McCarroll, S. A. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).

21. Fallin, D., Cohen, A., Essioux, L., Chumakov, I., Blumenfeld, M., Cohen, D. & Schork, N. J. Genetic Analysis of Case/Control Data Using Estimated Haplotype Frequencies: Application to APOE Locus Variation and Alzheimer's Disease. *Genome Res.* **11**, 143–151 (2001).

22. Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. & Poland, G. A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* **70**, 425–434 (2002).

23. Zaykin, D. V., Westfall, P. H., Young, S. S., Karnoub, M. A., Wagner, M. J. & Ehm, M. G. Testing Association of Statistically Inferred Haplotypes with Discrete and Continuous Traits in Samples of Unrelated Individuals. *Hum. Hered.* **53**, 79–91 (2002).

24. Salem, R. M., Wessel, J. & Schork, N. J. A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Hum. Genomics* **2**, 39 (2005).

25. Howey, R. & Cordell, H. J. Imputation Without Doing Imputation: A New Method for the Detection of Non-Genotyped Causal Variants. *Genet. Epidemiol.* **38**, 173–190 (2014).

26. Zheng, J., Rodriguez, S., Laurin, C., Baird, D., Trela-Larsen, L., Erzurumluoglu, M. A., Zheng, Y., White, J., Giambartolomei, C., Zabaneh, D., Morris, R., Kumari, M., Casas, J. P., Hingorani, A. D., UCLEB Consortium, Evans, D. M., Gaunt, T. R. & Day, I. N. M. HAPRAP: a haplotype-based iterative method for statistical fine mapping using GWAS summary statistics. *Bioinformatics* **33**, 79–86 (2017).

27. Al Bkhetan, Z., Chana, G., Soon Ong, C., Goudey, B. & Ramamohanarao, K. eQTLHap: a tool for comprehensive eQTL analysis considering haplotypic and genotypic effects. *Brief. Bioinform.* **22**, bbab093 (2021).

28. Zhang, B. C., Biddanda, A., Gunnarsson, Á. F., Cooper, F. & Palamara, P. F. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nat. Genet.* **55**, 768–776 (2023).

29. Massarat, A. R., Lamkin, M., Reeve, C., Williams, A. L., D'Antonio, M. & Gymrek, M. Haptools: a toolkit for admixture and haplotype analysis. *Bioinformatics* **39**, btad104 (2023).

30. Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D. G., Lek, M., Lizano, E., Buermans, H. P. J., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S. B., Donnelly, P., McCarthy, M. I., Flicek, P., Strom, T. M., Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, Á., Antonarakis, S. E., Häsler, R., Syvänen, A.-C., van Ommen, G.-J., Brazma,

A., Meitinger, T., Rosenstiel, P., Guigó, R., Gut, I. G., Estivill, X. & Dermitzakis, E. T. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

31. All of Us Research Program Investigators, Denny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., Jenkins, G. & Dishman, E. The 'All of Us' Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).

32. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

33. Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., Mohammadi, P., Park, Y., Parsana, P., Segrè, A. V., Strober, B. J., Zappala, Z., Cummings, B. B., Gelfand, E. T., Hadley, K., Huang, K. H., Lek, M., Li, X., Nedzel, J. L., Nguyen, D. Y., Noble, M. S., Sullivan, T. J., Tukiainen, T., MacArthur, D. G., Getz, G., Addington, A., Guan, P., Koester, S., Little, A. R., Lockhart, N. C., Moore, H. M., Rao, A., Struewing, J. P., Volpi, S., Brigham, L. E., Hasz, R., Hunter, M., Johns, C., Johnson, M., Kopen, G., Leinweber, W. F., Lonsdale, J. T., McDonald, A., Mestichelli, B., Myer, K., Roe, B., Salvatore, M., Shad, S., Thomas, J. A., Walters, G., Washington, M., Wheeler, J., Bridge, J., Foster, B. A., Gillard, B. M., Karasik, E., Kumar, R., Miklos, M., Moser, M. T., Jewell, S. D., Montroy, R. G., Rohrer, D. C., Valley, D., Mash, D. C., Davis, D. A., Sobin, L., Barcus, M. E., Branton, P. A., Abell, N. S., Balliu, B., Delaneau, O., Frésard, L., Gamazon, E. R., Garrido-Martín, D., Gewirtz, A. D. H., Gliner, G., Gloudemans, M. J., Han, B., He, A. Z., Hormozdiari, F., Li, X., Liu, B., Kang, E. Y., McDowell, I. C., Ongen, H., Palowitch, J. J., Peterson, C. B., Quon, G., Ripke, S., Saha, A., Shabalin, A. A., Shimko, T. C., Sul, J. H., Teran, N. A., Tsang, E.

K., Zhang, H., Zhou, Y.-H., Bustamante, C. D., Cox, N. J., Guigó, R., Kellis, M., McCarthy, M. I., Conrad, D. F., Eskin, E., Li, G., Nobel, A. B., Sabatti, C., Stranger, B. E., Wen, X., Wright, F. A., Ardlie, K. G., Dermitzakis, E. T., Lappalainen, T., Aguet, F., Ardlie, K. G., Cummings, B. B., Gelfand, E. T., Getz, G., Hadley, K., Handsaker, R. E., Huang, K. H., Kashin, S., Karczewski, K. J., Lek, M., Li, X., MacArthur, D. G., Nedzel, J. L., Nguyen, D. T., Noble, M. S., Segrè, A. V., Trowbridge, C. A., Tukiainen, T., Abell, N. S., Balliu, B., Barshir, R., Basha, O., Battle, A., Bogu, G. K., Brown, A., Brown, C. D., Castel, S. E., Chen, L. S., Chiang, C., Conrad, D. F., Cox, N. J., Damani, F. N., Davis, J. R., Delaneau, O., Dermitzakis, E. T., Engelhardt, B. E., Eskin, E., Ferreira, P. G., Frésard, L., Gamazon, E. R., Garrido-Martín, D., Gewirtz, A. D. H., Gliner, G., Gloudemans, M. J., Guigo, R., Hall, I. M., Han, B., He, Y., Hormozdiari, F., Howald, C., Kyung Im, H., Jo, B., Yong Kang, E., Kim, Y., Kim-Hellmuth, S., Lappalainen, T., Li, G., Li, X., Liu, B., Mangul, S., McCarthy, M. I., McDowell, I. C., Mohammadi, P., Monlong, J., Montgomery, S. B., Muñoz-Aguirre, M., Ndungu, A. W., Nicolae, D. L., Nobel, A. B., Oliva, M., Ongen, H., Palowitch, J. J., Panousis, N., Papasaikas, P., Park, Y., Parsana, P., Payne, A. J., Peterson, C. B., Quan, J., Reverter, F., Sabatti, C., Saha, A., Sammeth, M., Scott, A. J., Shabalin, A. A., Sodaei, R., Stephens, M., Stranger, B. E., Strober, B. J., Sul, J. H., Tsang, E. K., Urbut, S., van de Bunt, M., Wang, G., Wen, X., Wright, F. A., Xi, H. S., Yeger-Lotem, E., Zappala, Z., Zaugg, J. B., Zhou, Y.-H., Akey, J. M., Bates, D., Chan, J., Chen, L. S., Claussnitzer, M., Demanelis, K., Diegel, M., Doherty, J. A., Feinberg, A. P., Fernando, M. S., Halow, J., Hansen, K. D., Haugen, E., Hickey, P. F., Hou, L., Jasmine, F., Jian, R., Jiang, L., Johnson, A., Kaul, R., Kellis, M., Kibriya, M. G., Lee, K., Billy Li, J., Li, Q., Li, X., Lin, J., Lin, S., Linder, S., Linke, C., Liu, Y., Maurano, M. T., Molinie, B., Montgomery, S. B., Nelson, J., Neri, F. J.,

Oliva, M., Park, Y., Pierce, B. L., Rinaldi, N. J., Rizzardi, L. F., Sandstrom, R., Skol, A., Smith, K. S., Snyder, M. P., Stamatoyannopoulos, J., Stranger, B. E., Tang, H., Tsang, E. K., Wang, L., Wang, M., Van Wittenberghe, N., Wu, F., Zhang, R., Nierras, C. R., Branton, P. A., Carithers, L. J., Guan, P., Moore, H. M., Rao, A., Vaught, J. B., Gould, S. E., Lockart, N. C., Martin, C., Struewing, J. P., Volpi, S., Addington, A. M., Koester, S. E., Little, A. R., GTEx Consortium, Lead analysts:, Laboratory, D. A. & C. C. (LDACC):, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:, Laboratory, D. A. & C. C. (LDACC)—Analysis W. G., Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, & Biospecimen Collection Source Site—NDRI. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

34. Huang, H., Fang, M., Jostins, L., Umićević Mirkov, M., Boucher, G., Anderson, C. A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., D'Amato, M., Deffontaine, V., Dmitrieva, J., Docampo, E., Elansary, M., Farh, K. K.-H., Franke, A., Gori, A.-S., Goyette, P., Halfvarson, J., Haritunians, T., Knight, J., Lawrance, I. C., Lees, C. W., Louis, E., Mariman, R., Meuwissen, T., Mni, M., Momozawa, Y., Parkes, M., Spain, S. L., Théâtre, E., Trynka, G., Satsangi, J., van Sommeren, S., Vermeire, S., Xavier, R. J., Weersma, R. K., Duerr, R. H., Mathew, C. G., Rioux, J. D., McGovern, D. P. B., Cho, J. H., Georges, M., Daly, M. J. & Barrett, J. C. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).

35. Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T. & Posthuma, D. Genome-wide association studies. *Nat. Rev. Methods Primer* **1**, 1–21 (2021).

36. Skipper, S. & Josef, P. statsmodels: Econometric and statistical modeling with python. *9th Python Sci. Conf.* (2010).

37. Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Human Genome Structural Variation Consortium, Paul Flicek, null, Germer, S., Brand, H., Hall, I. M., Talkowski, M. E., Narzisi, G. & Zody, M. C. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).

38. Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. & Lee, J. J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).

39. Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Human Genome Structural Variation Consortium, Paul Flicek, null, Germer, S., Brand, H., Hall, I. M., Talkowski, M. E., Narzisi, G. & Zody, M. C. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).

40. Emmanouil Dermitzakis, N. K. & Lappalainen, T. RNA-sequencing of 465 lymphoblastoid cell lines from the 1000 Genomes. (2012).

41. Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., Sultan-Qurraie, A., Thomas, D. J., Trumbower, H., Weber, R. J., Weirauch, M., Zweig, A. S., Haussler, D. & Kent, W. J. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).

42. Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. A., Nickerson, D. A., Schmidt, J. P., Sherry, S. T., Wang, J., Wilson, R. K., Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Lander, E. S., Altshuler, D. M., Gabriel, S. B., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P., Resch, A. M., Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Bentley, D. R., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Borodina, T. A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Mardis, E. R., Wilson, R. K., Fulton, L., Fulton, R., Sherry, S. T., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J.,

Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., McVean, G. A., Durbin, R. M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T. M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Schmidt, J. P., Davies, C. J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Campbell, C. L., Kong, Y., Marcketta, A., Gibbs, R. A., Yu, F., Antunes, L., Bainbridge, M., Muzny, D., Sabo, A., Huang, Z., Wang, J., Coin, L. J. M., Fang, L., Guo, X., Jin, X., Li, G., Li, Q., Li, Y., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G. T., Garrison, E. P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A. N., Wu, J., Zhang, M., Daly, M. J., DePristo, M. A., Handsaker, R. E., Altshuler, D. M., Banks, E., Bhatia, G., del Angel, G., Gabriel, S. B., Genovese, G., Gupta, N., Li, H., Kashin, S., Lander, E. S., McCarroll, S. A., Nemesh, J. C., Poplin, R. E., Yoon, S. C., Lihm, J., Makarov, V., Clark, A. G., Gottipati, S., Keinan, A., Rodriguez-Flores, J. L., Korbel, J. O., Rausch, T., Fritz, M. H., Stütz, A. M., Flicek, P., Beal, K., Clarke, L., Datta, A., Herrero, J., McLaren, W. M., Ritchie, G. R. S., Smith, R. E., Zerbino, D., Zheng-Bradley, X., Sabeti, P. C., Shlyakhter, I., Schaffner, S. F., Vitti, J., Cooper, D. N., Ball, E. V., Stenson, P. D., Bentley, D. R., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Humphray, S., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E. E., Batzer, M. A., Konkel, M. K., Walker, J. A., MacArthur, D. G., Lek, M., Sudbrak, R., Amstislavskiy, V. S., Herwig, R., Mardis, E. R., Ding, L., Koboldt, D. C., Larson, D., Ye, K., Gravel, S., The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Coriell Institute for Medical Research, European

Molecular Biology Laboratory, E. B. I., Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Bilkent University, Boston College, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, McGill University & National Eye Institute, N. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

43. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).

44. Ziaei Jam, H., Li, Y., DeVito, R., Mousavi, N., Ma, N., Lujumba, I., Adam, Y., Maksimov, M., Huang, B., Dolzhenko, E., Qiu, Y., Kakembo, F. E., Joseph, H., Onyido, B., Adeyemi, J., Bakhtiari, M., Park, J., Javadzadeh, S., Jjingo, D., Adebiyi, E., Bafna, V. & Gymrek, M. A deep population reference panel of tandem repeat variation. *Nat. Commun.* **14**, 6711 (2023).

45. Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., Mohammadi, P., Park, Y., Parsana, P., Segrè, A. V., Strober, B. J., Zappala, Z., Cummings, B. B., Gelfand, E. T., Hadley, K., Huang, K. H., Lek, M., Li, X., Nedzel, J. L., Nguyen, D. Y., Noble, M. S., Sullivan, T. J., Tukiainen, T., MacArthur, D. G., Getz, G., Addington, A., Guan, P., Koester, S., Little, A. R., Lockhart, N. C., Moore, H. M., Rao, A., Struewing, J. P., Volpi, S., Brigham, L. E., Hasz, R., Hunter, M., Johns, C., Johnson, M., Kopen, G., Leinweber, W. F., Lonsdale, J. T., McDonald, A., Mestichelli, B., Myer, K., Roe, B., Salvatore, M., Shad, S.,

Thomas, J. A., Walters, G., Washington, M., Wheeler, J., Bridge, J., Foster, B. A., Gillard, B. M., Karasik, E., Kumar, R., Miklos, M., Moser, M. T., Jewell, S. D., Montroy, R. G., Rohrer, D. C., Valley, D., Mash, D. C., Davis, D. A., Sobin, L., Barcus, M. E., Branton, P. A., Abell, N. S., Balliu, B., Delaneau, O., Frésard, L., Gamazon, E. R., Garrido-Martín, D., Gewirtz, A. D. H., Gliner, G., Gloudemans, M. J., Han, B., He, A. Z., Hormozdiari, F., Li, X., Liu, B., Kang, E. Y., McDowell, I. C., Ongen, H., Palowitch, J. J., Peterson, C. B., Quon, G., Ripke, S., Saha, A., Shabalin, A. A., Shimko, T. C., Sul, J. H., Teran, N. A., Tsang, E. K., Zhang, H., Zhou, Y.-H., Bustamante, C. D., Cox, N. J., Guigó, R., Kellis, M., McCarthy, M. I., Conrad, D. F., Eskin, E., Li, G., Nobel, A. B., Sabatti, C., Stranger, B. E., Wen, X., Wright, F. A., Ardlie, K. G., Dermitzakis, E. T., Lappalainen, T., Aguet, F., Ardlie, K. G., Cummings, B. B., Gelfand, E. T., Getz, G., Hadley, K., Handsaker, R. E., Huang, K. H., Kashin, S., Karczewski, K. J., Lek, M., Li, X., MacArthur, D. G., Nedzel, J. L., Nguyen, D. T., Noble, M. S., Segrè, A. V., Trowbridge, C. A., Tukiainen, T., Abell, N. S., Balliu, B., Barshir, R., Basha, O., Battle, A., Bogu, G. K., Brown, A., Brown, C. D., Castel, S. E., Chen, L. S., Chiang, C., Conrad, D. F., Cox, N. J., Damani, F. N., Davis, J. R., Delaneau, O., Dermitzakis, E. T., Engelhardt, B. E., Eskin, E., Ferreira, P. G., Frésard, L., Gamazon, E. R., Garrido-Martín, D., Gewirtz, A. D. H., Gliner, G., Gloudemans, M. J., Guigo, R., Hall, I. M., Han, B., He, Y., Hormozdiari, F., Howald, C., Kyung Im, H., Jo, B., Yong Kang, E., Kim, Y., Kim-Hellmuth, S., Lappalainen, T., Li, G., Li, X., Liu, B., Mangul, S., McCarthy, M. I., McDowell, I. C., Mohammadi, P., Monlong, J., Montgomery, S. B., Muñoz-Aguirre, M., Ndungu, A. W., Nicolae, D. L., Nobel, A. B., Oliva, M., Ongen, H., Palowitch, J. J., Panousis, N., Papasaikas, P., Park, Y., Parsana, P., Payne, A. J., Peterson, C. B., Quan, J., Reverter, F., Sabatti, C., Saha, A., Sammeth, M., Scott, A. J., Shabalin, A. A., Sodaei, R.,

Stephens, M., Stranger, B. E., Strober, B. J., Sul, J. H., Tsang, E. K., Urbut, S., van de Bunt, M., Wang, G., Wen, X., Wright, F. A., Xi, H. S., Yeger-Lotem, E., Zappala, Z., Zaugg, J. B., Zhou, Y.-H., Akey, J. M., Bates, D., Chan, J., Chen, L. S., Claussnitzer, M., Demanelis, K., Diegel, M., Doherty, J. A., Feinberg, A. P., Fernando, M. S., Halow, J., Hansen, K. D., Haugen, E., Hickey, P. F., Hou, L., Jasmine, F., Jian, R., Jiang, L., Johnson, A., Kaul, R., Kellis, M., Kibriya, M. G., Lee, K., Billy Li, J., Li, Q., Li, X., Lin, J., Lin, S., Linder, S., Linke, C., Liu, Y., Maurano, M. T., Molinie, B., Montgomery, S. B., Nelson, J., Neri, F. J., Oliva, M., Park, Y., Pierce, B. L., Rinaldi, N. J., Rizzardi, L. F., Sandstrom, R., Skol, A., Smith, K. S., Snyder, M. P., Stamatoyannopoulos, J., Stranger, B. E., Tang, H., Tsang, E. K., Wang, L., Wang, M., Van Wittenberghe, N., Wu, F., Zhang, R., Nierras, C. R., Branton, P. A., Carithers, L. J., Guan, P., Moore, H. M., Rao, A., Vaught, J. B., Gould, S. E., Lockart, N. C., Martin, C., Struewing, J. P., Volpi, S., Addington, A. M., Koester, S. E., Little, A. R., GTEx Consortium, Lead analysts:, Laboratory, D. A. & C. C. (LDACC):, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:, Laboratory, D. A. & C. C. (LDACC)—Analysis W. G., Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, & Biospecimen Collection Source Site—NDRI. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

46. Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M. A., Lambourne, J. J., Sivapalaratnam, S., Downes, K., Kundu, K., Bomba, L., Berentsen, K., Bradley, J. R., Daugherty, L. C., Delaneau, O., Freson, K., Garner, S. F., Grassi, L., Guerrero, J., Haimel, M., Janssen-

Megens, E. M., Kaan, A., Kamat, M., Kim, B., Mandoli, A., Marchini, J., Martens, J. H. A., Meacham, S., Megy, K., O'Connell, J., Petersen, R., Sharifi, N., Sheard, S. M., Staley, J. R., Tuna, S., van der Ent, M., Walter, K., Wang, S.-Y., Wheeler, E., Wilder, S. P., Iotchkova, V., Moore, C., Sambrook, J., Stunnenberg, H. G., Di Angelantonio, E., Kaptoge, S., Kuijpers, T. W., Carrillo-de-Santa-Pau, E., Juan, D., Rico, D., Valencia, A., Chen, L., Ge, B., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yang, Y., Guigo, R., Beck, S., Paul, D. S., Pastinen, T., Bujold, D., Bourque, G., Frontini, M., Danesh, J., Roberts, D. J., Ouwehand, W. H., Butterworth, A. S. & Soranzo, N. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).

47. Mirmira, T., Ma, N., Margoliash, J., Rivera, W. G. G., Amariuta, T., Frazer, K. A., Goren, A. & Gymrek, M. Characterization of shared and ancestry-specific signals driving complex traits using multi-ancestry fine-mapping. *MedRxiv Prepr. Serv. Health Sci.* 2025.11.13.25339614 (2025) doi:10.1101/2025.11.13.25339614.