Title
Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies

Journal
American Journal of Human Genetics, 97(2)

ISSN
0002-9297

Authors
Kichaev, Gleb
Pasaniuc, Bogdan

Publication Date
2015-08-01

DOI
10.1016/j.ajhg.2015.06.007

Peer reviewed

ARTICLE

Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies

Gleb Kichaev[1] and Bogdan Pasaniuc[1,2,3,*]

Localization of causal variants underlying known risk loci is one of the main research challenges following genome-wide association studies. Risk loci are typically dissected through fine-mapping experiments in trans-ethnic cohorts for leveraging the variability in the local genetic structure across populations. More recent works have shown that genomic functional annotations (i.e., localization of tissue-specific regulatory marks) can be integrated for increasing fine-mapping performance within single-population studies. Here, we introduce methods that integrate the strength of association between genotype and phenotype, the variability in the genetic backgrounds across populations, and the genomic map of tissue-specific functional elements to increase trans-ethnic fine-mapping accuracy. Through extensive simulations and empirical data, we have demonstrated that our approach increases fine-mapping resolution over existing methods. We analyzed empirical data from a large-scale trans-ethnic rheumatoid arthritis (RA) study and showed that the functional genetic architecture of RA is consistent across European and Asian ancestries. In these data, we used our proposed methods to reduce the average size of the 90% credible set from 29 variants per locus for standard non-integrative approaches to 22 variants.

Introduction

Genome-wide associations studies (GWASs) have reproducibly identified thousands of risk loci associated with complex traits and diseases.[1–7] Unfortunately, the index variants reported in these studies are typically not biologically causal but rather correlated with the underlying causal variant through linkage disequilibrium (LD).[8] Fine-mapping experiments identify causal variants responsible for the GWAS signal first by gathering dense genetic information, either through targeted sequencing or dense imputation, and second by statistically prioritizing variants that can subsequently be validated in functional studies.[3,6,9,10]

Divergent population histories due to various demographic forces such as bottlenecks and expansions have produced unique genomic landscapes across ethnicities.[11,12] Such differences in LD patterns and variant frequencies across populations can increase the power of statistical fine mapping if they are properly modeled.[3,13–18] Intuitively, if a locus contains a causal variant, the neighborhood of LD partners linked to this variant will be distinct in different populations. Thus, aggregating the strength of association across multiple populations might accentuate the signal from the true causal variant(s) while dampening the noise from correlated variants.

A common approach to combining information across multiple studies is through inverse-variance fixed-effects meta-analysis,[19] which assumes that effect sizes of causal variants are similar across studies or populations. This assumption can be relaxed by a random-effects strategy, although it has been observed that this usually results in a decrease in statistical power.[20] A recent, and more robust, Bayesian meta-analysis framework[15] was proposed to reason over trans-ethnic studies with potential allelic heterogeneity. Both the fixed-effects meta-analysis statistics and the Bayes factors supplied by the latter approach can be readily converted into posterior probabilities of association (PPAs) for the construction of fine-mapping credible sets.[3,21] However, these credible sets are commonly built under the assumption that a locus harbors at most a single causal variant,[3,22–24] which might be invalidated at many risk loci,[17,25,26] leading to miscalibrated credible sets.[27,28] Although conceptually it might be possible to create credible sets on the basis of independent signals identified through conditional analysis, this strategy suffers from necessitating an ad hoc re-definition of the fine-mapping region. Furthermore, multiple causal variants in LD can create at neighboring sites synthetic associations that are potentially stronger than the association at the true causal variants. The iterative conditioning strategy would necessarily select these synthetic SNPs first, thereby dissipating the signal from the true causal variants.[27]

In addition to the strength of association between genotype and phenotype, an orthogonal source of information lies within a variant's functional genomic context. Projects such as ENCODE[29] and ROADMAP[30] have provided a rich atlas of functional information, and numerous groups have reproducibly demonstrated that disease-associated variants are systematically enriched within chromatin marks that delineate active regulatory regions in phenotypically relevant cell types.[31–35] Whereas functional genomic data are often used as a post hoc validation of association findings,[4,10,36] a number of principled approaches have been proposed to jointly integrate functional and association data.[28,35,37,38] In addition to increasing the accuracy of fine mapping, these integrative approaches also provide insights into the genetic architecture of the trait by identifying relevant tissue-specific functional marks without

[1]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90095, USA; [2]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA; [3]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
*Correspondence: pasaniuc@ucla.edu
http://dx.doi.org/10.1016/j.ajhg.2015.06.007.

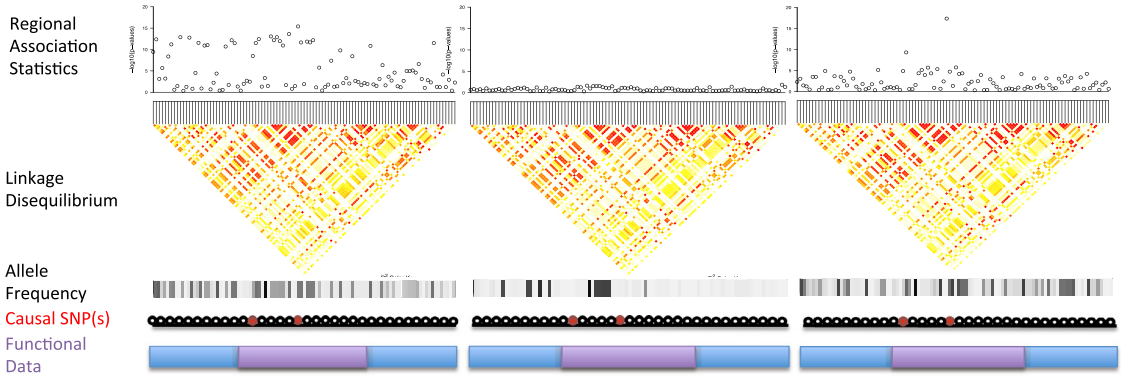

Figure 1. Example of a Fine-Mapping Locus in Three Different Populations
In population 1 (left), the causal variants are present, but strong regional LD makes it difficult to distinguish them from tagging SNPs. In population 2 (middle), the causal variants both have a very low frequency and/or are monomorphic, resulting in no observable association between the SNPs and the trait. In population 3 (right), the causal variants are common and have few tagging SNPs. Our framework jointly models population-specific LD structure and integrates functional genomic data to prioritize causal variants.

making any prior assumptions. However, to the best of our knowledge, functional integrative approaches have not been extended to trans-ethnic fine mapping, and a rigorous assessment of trans-ethnic fine mapping in the presence of multiple causal variants is currently lacking. Although in principle the single-population frameworks that allow for multiple causal variants[27,28] can operate directly on trans-ethnic meta-analysis statistics, they require ad hoc averaging of trans-ethnic LD and do not properly account for heterogeneity by ancestry at causal variants.

In this work, we propose a statistical framework that integrates three sources of information to triangulate causal variants in fine-mapping studies: (1) the strength of association between genotype and phenotype, (2) differential genomic background across ethnic groups, and (3) tissue-specific functional genomic annotations (Figure 1). Different allele frequencies (or sample sizes) across populations induce differential standardized effect sizes at all the variants in a region, even in the presence of no allelic effect-size heterogeneity by ancestry. We model this induced heterogeneity across populations through a multivariate normal (MVN) framework wherein the sets of population-specific association statistics are realizations from population-specific MVN distributions. Similar to the case of a single population,[27,28] this allows us to consider multiple causal variants at any risk locus. We integrate functional genomic data by using Empirical Bayes,[28] which provides a means of selecting functional annotations most relevant to the trait of interest. Most importantly, our proposed approach requires only the summary association data for each population, thereby avoiding the many restrictions that can accompany analysis of individual-level genotype data.

Through extensive simulations, we show that our trans-ethnic framework significantly improves fine-mapping resolution over conventional meta-analysis strategies and demonstrate that considering multiple causal variants in multi-ethnic cohorts yields large gains in fine-mapping efficiency. We showcase our framework by reanalyzing empirical summary data from a large trans-ethnic rheumatoid arthritis (RA [OMIM: 180300]) GWAS.[4] We first demonstrate that the functional architecture of RA is consistent across ethnicities and that there is a strong preponderance of immune-related functional classes that are enriched with causal variants. We then fine map the RA GWAS loci by using functional data and show that our method greatly outperforms current state-of-the art methodologies and uncovers a number of plausible functional variants.

Material and Methods

Multi-population Fine-Mapping Framework

Without loss of generality (given that similar results can be derived for case-control traits), let y be a quantitative phenotype such that $y_i = g_i\beta + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_e^2)$, and g_i denotes a multi-SNP genotype containing {0,1,2} counts of the reference allele at M SNPs for an individual i. The β vector represents allelic effects where the j^{th} entry will be non-zero only if SNP j is causal. Given genotype (G_p) and phenotype (Y_p) data over N_p individuals from population p, a standard approach to measuring association strength at SNP j is through the Wald statistic

$$z_p^j = \frac{\widehat{\beta}_p^j}{\mathrm{SE}\left(\widehat{\beta}_p^j\right)} = \frac{\mathrm{Cov}\left(G_p^j, Y_p\right)\sqrt{N_p}}{\mathrm{Var}\left(G_p^j\right)\sigma_e^2},$$

which asymptotically follows a normal distribution:

$$\mathcal{N}\left(\frac{\beta_p^j\sqrt{\mathrm{Var}\left(G_p^j\right)}}{\sigma_e}\sqrt{N_p}, 1\right).$$

We denote the non-centrality parameter (NCP) of the normal distribution as

$$\lambda_p^j = \frac{\beta_p^j\sqrt{\mathrm{Var}\left(G_p^j\right)}}{\sigma_e}\sqrt{N_p}.$$

Under the null hypothesis that SNP j is not causal (or does not tag a causal variant; see below), $\beta_p^j = 0$ and thus $\lambda_p^j = 0$. If the SNP is causal, then $\beta_p^j \neq 0$, yielding a non-zero λ_p^j, and governs

the power of detecting this variant in an association study (i.e., rejecting the null at some confidence level). Importantly, even when the allelic effects at the causal variants are similar across populations (i.e., $\beta_p^j = \beta_{p'}^j$), different allele frequencies and sample sizes induce population-specific NCPs, yielding larger NCPs at more common SNPs and/or larger studies. This leads to the well-known result that causal variants are more readily detectable in populations in which they are present more frequently.

Pervasive LD at fine-scale resolutions induces correlations between tag SNPs and causal SNPs, thus creating an indirect association between tag SNPs and traits.[13] More specifically, the LD-induced NCP at a SNP j (Λ_p^j) can be approximated as a linear combination of NCPs at the causal SNPs with LD-adjusted weights[13,27,28,39] as

$$\Lambda_p^j = \sum_c r_p^{j,c} \lambda_p^c, \qquad \text{(Equation 1)}$$

where the sum is taken across all causal SNPs c, and $r_p^{j,c}$ is the Pearson correlation coefficient between SNPs j and c in population p. We expand Equation 1 to include all SNPs in the locus by incorporating indicator variable C_p^k, which is set to 1 if SNP k is causal in population p and 0 otherwise:

$$\Lambda_p^j = \sum_{k=1}^{M} r_p^{j,k} \lambda_p^k C_p^k. \qquad \text{(Equation 2)}$$

In vector notation,

$$\Lambda_p = \Sigma_p (\lambda_p \circ C_p), \qquad \text{(Equation 3)}$$

where Σ_p is the LD matrix of Pearson correlations among the M SNPs, C_p is a binary vector indicating which SNPs are causal, and $\circ$ denotes the element-wise multiplication between two vectors. We can now write the probability of the data (i.e., the observed standardized effect sizes, Z scores) given the causal variants (C_p) in population p under a MVN assumption:

$$Z_p \mid \lambda_p, C_p \sim \mathcal{N}(\Lambda_p, \Sigma_p). \qquad \text{(Equation 4)}$$

This allows us to define the total likelihood of the data by marginalizing across all sets of causal SNPs ($\mathcal{C}$) as

$$L(Z_1, Z_2, \cdots, Z_P; \lambda_1, \lambda_2, \cdots, \lambda_P) = \prod_p \sum_{C_p \in \mathcal{C}} P(Z_p \mid \lambda_p, C_p) P(C_p), \qquad \text{(Equation 5)}$$

which we simplify under the assumption that the causal vector set is identical across populations:

$$L(Z_1, Z_2, \cdots, Z_P; \lambda_1, \lambda_2, \cdots, \lambda_P) = \sum_{C \in \mathcal{C}} \prod_p P(Z_p \mid \lambda_p, C) P(C). \qquad \text{(Equation 6)}$$

Here, $P(Z_p|\lambda_p, C_p)$ is defined as the probability density function of the MVN (see Equation 4), and $P(C)$ is the probability of a given causal set. Note that Equation 6 assumes that the causal set is identical across populations but allows for different effect sizes at causal SNPs across populations.

Integration of Functional-Annotation Data

We assume that each variant can potentially have several phenotypically relevant genomic functional annotations (e.g., transcription factor binding site), which can be encoded as binary variable A_{jk} for variant j and annotation k or as a continuous value (e.g., a probabilistic membership of variants in different functional classes). We integrate the functional information through the probability of the causal set C as follows,

$$P(C; \gamma) = \prod_j \left(\frac{1}{1 + \exp(-\gamma^T A_j)} \right)^{C_j} \left(\frac{1}{1 + \exp(\gamma^T A_j)} \right)^{1 - C_j}, \qquad \text{(Equation 7)}$$

where γ is a vector containing the prior log-odds ratio of causality for every functional annotation. We extend the likelihood to incorporate functional data as

$$L(Z_1, Z_2, \cdots, Z_P; \lambda_1, \lambda_2, \cdots, \lambda_P, \gamma) = \sum_{C \in \mathcal{C}} \prod_p P(Z_p \mid \lambda_p, C) P(C; \gamma), \qquad \text{(Equation 8)}$$

which we can further simplify by assuming that data at different loci are independent:

$$L(Z_1, Z_2, \cdots, Z_P; \lambda_1, \lambda_2, \cdots, \lambda_P, \gamma) = \prod_l \sum_{C_l \in \mathcal{C}_l} \prod_p P(Z_{l,p} \mid \lambda_{p,l}, C_l) P(C_l; \gamma). \qquad \text{(Equation 9)}$$

Finally, to obtain posterior probabilities that each SNP is causal, we use Bayes theorem to compute the joint posterior for each causal set,

$$P(C_l \mid Z_{l,1}, Z_{l,2}, \cdots, Z_{l,P}; \lambda_{l,1}, \lambda_{l,2}, \cdots, \lambda_{l,P}, \gamma) = \frac{\prod_p P(Z_{l,p} \mid \lambda_{p,l}, C_l) P(C_l; \gamma)}{\sum_{C_l \in \mathcal{C}_l} \prod_p P(Z_{l,p} \mid \lambda_{p,l}, C_l) P(C_l; \gamma)}, \qquad \text{(Equation 10)}$$

and subsequently marginalize across all $C_l = (C_{1l}, C_{2l}, \ldots, C_{N_l})$ such that $C_{jl} = 1$:

$$P(C_{jl} \mid Z_1, Z_2, \cdots, Z_P; \lambda_{l,1}, \lambda_{l,2}, \cdots, \lambda_{l,P}, \gamma) = \sum_{C_l \in \mathcal{C}_l : C_{jl} = 1} P(C_l \mid Z_1, Z_2, \cdots, Z_P; \lambda_{l,1}, \lambda_{l,2}, \cdots, \lambda_{l,P}, \gamma). \qquad \text{(Equation 11)}$$

Model Fitting

Because of the finite nature of either the sample or the reference panel, the LD matrix in practice could be ill conditioned. We apply a Tikhonov Regularization[40] to all LD matrices to ensure their invertibility and as a result preserve the non-degeneracy and numerical stability of the MVN approximation. Furthermore, because we ensure that all Σ are positive definite, there exists a Cholesky decomposition for each LD matrix and its corresponding inverse. Let $L_p = \text{Chol}(\Sigma_p)^{-1}$; it follows that $\tilde{Z}_p = L_p Z_p \sim \mathcal{N}(L_p \Lambda_p, I)$. In practice, we operate in the transformed-Z-score space ($\tilde{Z}_p$) because it improves numerical stability and reduces computational burden by removing a large, repetitive matrix multiplication when computing the MVN density.

We fit the parameters of the model to the data across all loci by using a variant of the expectation maximization over the functional annotations (γ) and approximate the NCPs by using a simple function of the observed Z scores (see Appendix A). We note that because enumerating over all possible causal sets is combinatorially intractable, we typically restrict the number of causal variants per locus to two or three in practice.

Simulation Data

We benchmarked our proposed framework by using simulations starting from real genotype data. Using the NHGRI catalog of GWAS variants on chromosome 1,[1] we centered 25-kb windows on the lead SNP and used HAPGEN2[41] and 1000 Genomes[12] to simulate individuals from the Asian (n = 286), African (n = 246), and European (n = 379) ancestries. SNPs that were polymorphic with a minor allele frequency $\geq$ 0.01 in at least one population

were retained for analysis. For each simulation, we randomly chose 50 loci and simulated causal variants by drawing causal status according to the logistic prior model described above. Unless otherwise noted, we used the annotations (coding, UTR, promoter, DNase hypersensitivity site [DHS], intronic, and intergenic) and functional enrichments (13.8×, 8.4×, 2.8×, 5.1×, and 0.1×) observed in Gusev et al.[35] for simulations below. We simulated phenotypes under a linear model such that for individual i of population p, their phenotype Y was drawn as $Y_{i,p} = \sum_{j=1}^{N_c} \beta_j \cdot g_{j,i,p} + \epsilon_{i,p}$, where N_c is the total number of causal variants, β_j is the effect size of the j^{th} causal SNP, $g_{j,i,p}$ is the number of copies of the risk allele j for individual i of population p. Following recent works, we simulated similar heritability across populations.[42] The population-specific error term, $\epsilon_{i,p}$, was drawn according to a $\mathcal{N}(0, \sigma^2_{e,p})$, where $\sigma^2_{e,p} = (\sigma^2_{g,p} - h^2_g * \sigma^2_{g,p})/h^2_g$, $\sigma^2_{g,p} = \beta' \mathrm{Cov}(X_p)\beta$, and $\mathrm{Cov}(X_p)$ is the population-specific covariance of the genotypes (LD). The effect size, β_j, was set to be inversely proportional to the average SD of the population allele frequencies; this is roughly equivalent to assuming that each causal SNP explains an equal proportion of the phenotypic variance.[43]

Existing Methods

We compared our proposed methods with other well-established probabilistic methods for fine mapping. First, we investigated MANTRA, a Bayesian trans-ethnic meta-analysis technique proposed by Morris.[15] We obtained the software implementation from the author and ran it with the default settings; we provided the fixation index (F_{ST}) between the three populations as determined in Nelis et al.[44] as the prior for the Bayesian partition model. The output of MANTRA is a Bayes factor, which we subsequently converted to

$$\mathrm{PPA}_i = \frac{BF_i}{\sum_k BF_k}$$

as previously recommended.[3,22,45] Similarly, we calculated posterior probabilities that SNPs are causal strictly on the basis of the inverse-variance fixed-effects[19] meta-analysis by using the PAINTOR (Probabilistic Annotation Integrator)[28] and CAVIARBF[46] frameworks. We note that the CAVIARBF and PAINTOR models require LD as input, which we calculated as the average of the population-specific LD weighted by the sample size of each population. We assessed accuracy by rank ordering SNPs across all fine-mapping loci according to the output of each method and then determined the proportion of identified causal variants as more SNPs were selected. We typically report the median number of SNPs one would need to validate in order to resolve 90% of the causal variants as our main metric of accuracy.

RA Multi-ethnic Fine-Mapping Dataset

We downloaded summary statistics from a large trans-ethnic RA GWAS consisting of over 100,000 individuals (~68,000 of European ancestry and ~36,000 of Asian ancestry).[4] We used the reported genome-wide-significant loci, excluding human leukocyte antigen regions, and centered 100-kb windows around the top SNP, yielding a total of 89 fine-mapping loci. For each of these regions, we estimated LD by using the European and Asian individuals from 1000 Genomes.[12] We integrated 482 publicly available functional annotations comprising 406 DHSs spanning numerous cell types and tissues,[31,47] the seven genomic segmentations of the eight primary ENCODE cell lines,[48] Fantom5 enhancer and transcription start site regions,[49] immune cell enhancers,[10] genic elements derived from GenCode,[50] and overall methylation and acetylation marks from ENCODE.[29] The construction of a phenotypically specific fine-mapping model requires two phases. First, we ran the model marginally on each annotation and subsequently rank ordered all the annotations according to likelihood-ratio statistics.[28,37] Second, we selected the top annotations that were minimally correlated with one another (usually no more than five) to enter a final model to estimate posterior probabilities that variants are causal.

Results

Joint Modeling of Association Statistics across Populations Increases Fine-Mapping Performance

We used simulations to investigate the benefit of jointly modeling population-specific association statistics versus standard meta-analysis approaches. We simulated fine-mapping datasets over 10,000 individuals equally divided among European, Asian, and African ancestries with total heritability of $h^2_g = 0.25$ across 50 loci with genetic architecture similar to that in Gusev et al.[35] The loci were simulated such that in expectation, each locus harbored a single causal variant with allelic effects shared across populations (see Material and Methods). This yielded an average of 15 loci with a single causal variant and 13 loci with multiple causal variants per simulation. In general, we find that trans-ethnic fine-mapping strategies that assume a single causal variant are less optimal than those that allow for multiple causal variants (Table 1). For example, MANTRA meta-analysis requires 1.9 and 96.8 SNPs per locus in order to identify 50% and 90% of the causal variants, respectively, whereas methods that allow multiple causal variants but do not incorporate functional data require 1.2 and 7.0 SNPs per locus to identify 50% and 90% of the causal variants, respectively.[46] Existing integrative fine-mapping methods that leverage functional data[28] applied to fixed-effects meta-analysis statistics achieve accuracy of 1.0 and 5.6 SNPs per locus to find 50% and 90% of the causal variants, respectively. In contrast, our proposed framework resolves causal variants with the greatest efficiency (Figure 2) in that it requires only 0.9 and 5.2 SNPs per locus to find 50% and 90% of the causal variants, respectively (paired t test, $p < 0.001$). Overall, this can be attributed to the fact that our approach models population-specific LD patterns while allowing for multiple causal variants in the presence of functional information.

Recent studies have shown that GWAS findings generally replicate across populations,[42,51] thus suggesting sharing of underlying causal variants. However, it is generally unknown whether these variants contribute to disease risk uniformly across populations. We sought to assess the performance of fine mapping in the situation where the causal variants have either weak or strong heterogeneity by ancestry. In addition to fine mapping datasets in which causal effects were similar across populations (no heterogeneity), we simulated allelic effects inversely proportional to the population-specific allele-frequency SD (weak heterogeneity) and normally distributed allelic effects for each ancestry independently (strong heterogeneity). We found

Table 1. Our Trans-ethnic Integrative Framework Is Superior to Conventional Meta-analysis Strategies and Current State-of-the-Art Methodologies

Heterogeneity Level	Identified Proportion of Causal Variants	Single Causal Variant per Locus		Multiple Causal Variants per Locus		
		Fixed-Effects Meta-analysis	MANTRA[15]	Fixed-Effects CAVIARBF[46]	Fixed-Effects PAINTOR[28,a]	Trans-ethnic PAINTOR[a]
None	0.50	1.9	2.0	1.2	1.0	0.9
	0.75	29.8	30.3	2.9	2.1	1.9
	0.90	96.8	96.8	7.0	5.6	5.2
Weak	0.50	1.9	2.0	1.1	0.9	0.9
	0.75	62.3	62.7	2.9	2.0	1.8
	0.90	118.1	118.6	6.8	4.9	4.1
Strong	0.50	29.0	11.1	12.6	9.6	2.3
	0.75	105.0	92.7	68.6	58.4	19.7
	0.90	143.9	139.8	134.4	121.3	56.5

We simulated 1,000 multi-ethnic fine-mapping datasets under various levels of allelic heterogeneity across populations. For the first two levels of heterogeneity ("none" and "weak"), we invoked the standard infinitesimal assumption on allelic effects either globally or at the population level by setting effect sizes $(\beta_{c,p})$ at the causal SNPs inversely proportional to either the mean allele-frequency SD or the population-specific allele-frequency SD. To simulate strong heterogeneity across ancestries, we drew effect sizes from a standard normal distribution for each population independently and added enough Gaussian noise to maintain $h^2_g = 0.25$. Displayed here is the median number of SNPs selected per locus for identifying a specified proportion of the causal variants.
[a]Methods that also integrate functional data.

that our framework significantly outperformed the fixed-effects meta-analysis followed by probability estimation by existing methods. For example, in the case of weak heterogeneity, our approach required 4.1 as opposed to 4.9 SNPs per locus (19.5% improvement); in addition, in the presence of strong heterogeneity, our approach dramatically outperformed existing meta-analysis strategies by reducing the number of SNPs required for identifying 90% of the causal variants from 121.3 to 56.5 (214% improvement) (Table 1; Figure 2). The increase in performance is likely due to the fact that our framework makes no assumptions pertaining to the population-specific allelic effects at causal SNPs, given that we allow the empirically observed Z scores in each population to dictate the effect size. This allows for arbitrary levels of heterogeneity in the effect size by population, whereas fixed-effects meta-analysis assumes similar effect sizes across populations.

Performance of Trans-ethnic Fine Mapping

The benefit of trans-ethnic fine mapping has been thoroughly documented both in simulations and in empirical data.[3,13,15] However, previous works have assumed a single causal variant at a risk locus, and this assumption is often invalidated in practice. Here, we sought to assess trans-ethnic fine mapping in the presence of multiple causal variants at a risk locus while integrating functional-annotation data. Consistent with previous works,[13] we found that for the same sample size, multi-ethnic cohorts attained superior accuracy over single-population studies. However, allowing for multiple causal variants enabled trans-ethnic fine mapping to perform even better than single-population fine mapping. We observed a near 3- to 4-fold increase in the median resolution for methods that model multiple causal variants but only a 1.4- to 1.6-fold gain for methods that assume a single causal (see Table 2). We attribute this to the

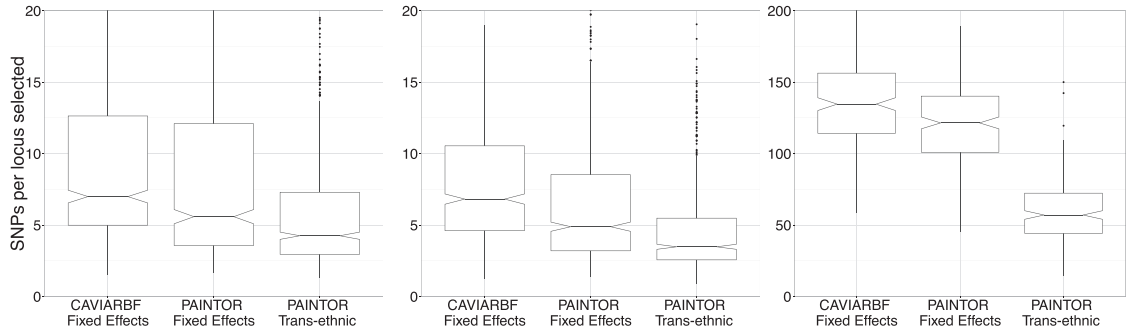

Figure 2. Trans-ethnic PAINTOR Is Most Efficient in Identifying Causal Variants
The distributions of the number of SNPs required for follow-up identification of 90% of the causal variants across 1,000 simulations are displayed as boxplots. The different panels represent increasing levels of effect-size heterogeneity by ancestry: none (left), weak (middle), and strong (right). The widths of the notches in each boxplot roughly correspond to 95% confidence intervals for the median number of SNPs required for resolving 90% of the causal variants. For the sake of clarity, we have cut the y axis to emphasize the significant difference in performance across all three methods.

Table 2. Modeling Multiple Causal Variants in Multi-ethnic Cohorts Yields Larger Relative Gains in Fine-Mapping Efficiency

Ethnic Group	Single Causal Variant		Multiple Causal Variants	
	−	+	−	+
Asians	136.9	134.4	89.3	36.2
Europeans	135.0	130.9	82.9	33.5
Africans	104.0	95.0	34.4	14.7
Trans-ethnic	72.6	58.4	8.5	4.9
Relative	1.4	1.6	4.0	3.0

We simulated fine-mapping datasets with various ethnic compositions and allelic effects shared across populations. Displayed here are four fine-mapping strategies that consider either single or multiple causal variants at each risk locus and either have (+) or do not have (−) access to functional data across different ethnic study designs. The bottom row represents the relative gain in the median 90% causal-variant resolution of trans-ethnic cohorts over the next best-performing group.

much smaller number of sets of causal variants (as a proportion of the total possible sets) that are compatible with the observed association statistics. Diversity in LD patterns across populations additionally penalizes sets of variants that do not contain the true causal variants because they are unlikely to explain the observed data. Consequently, multi-ethnic cohorts not only will have proportionally more LD patterns than single-population cohorts (therefore placing larger penalties on incorrect causal sets) but can also borrow power from populations where the causal variants are present more frequently.

Genetic-Trait Architecture Affects Fine-Mapping Performance

Functional information was demonstrated to improve fine-mapping resolution in a single population,[10,28,37,38] and we investigated the potential gains in a trans-ethnic setting. We simulated two disease architectures by using five functional annotations where causal variants either localize predominantly within a single broad functional class, as observed by Gusev et al.[35] (A1), or have a smaller, more diffuse localization within functionally specific cell types[28] (A2). For each class of disease architectures, we fit six trans-ethnic integrative models such that each successive model incorporated an additional functional annotation into a joint framework. Not surprisingly, when the true genetic architecture of a trait at fine-mapping regions has a strong enrichment of causal variants within a common functional class (i.e., DHSs[35]), these functional annotations will be most informative for the purposes of fine mapping (see Figure 3). On the other hand, more diffuse localization of causal variants requires multiple annotations for maximizing the utility of functional data. For example, for genetic architecture A1, the addition of the DHS annotation yielded a 70% increase in fine-mapping resolution, whereas genetic architecture A2 required all five annotations to improve resolution by 18% (see Figure 3).

Integrative Fine Mapping in a Multi-ethnic RA Dataset

We investigated whether similar results from simulations can be attained in empirical data from a trans-ethnic RA GWAS over more than 100,000 individuals[4] (see Material and Methods). Because the functional genetic architecture of RA across different populations is unknown, we first quantified whether the enrichment of causal variants in various functional annotations is consistent across ancestries. Reassuringly, we saw a strong correspondence in functional enrichment at the fine-mapping loci across all 482 functional categories we investigated ($r = 0.597$; Figure 4). This provides evidence supporting the assumption that a

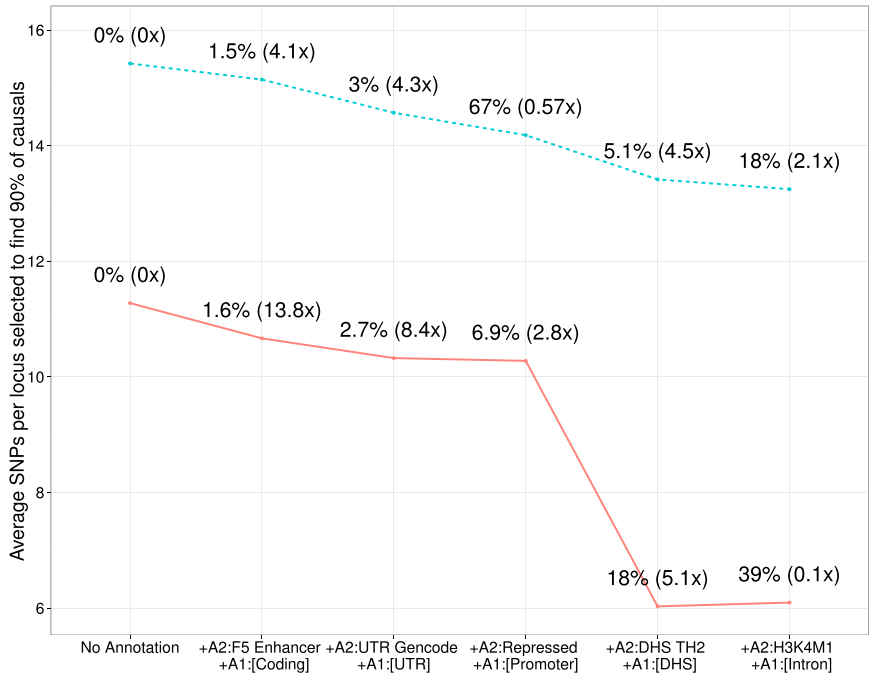

Figure 3. The Underlying Functional Architecture of a Trait Affects Fine-Mapping Performance

We simulated two classes of disease architectures: A1 (solid line) and A2 (dashed line). Architecture A1 was based on the functional enrichment observed in Gusev et al.[35] and had a strong enrichment within a single DHS class. Architecture A2 was simulated with a more diffuse enrichment in various cell types and classes and was based on what we empirically observed in the RA dataset. Displayed on top of each point is the percentage of SNPs falling within that annotation and its corresponding enrichment.

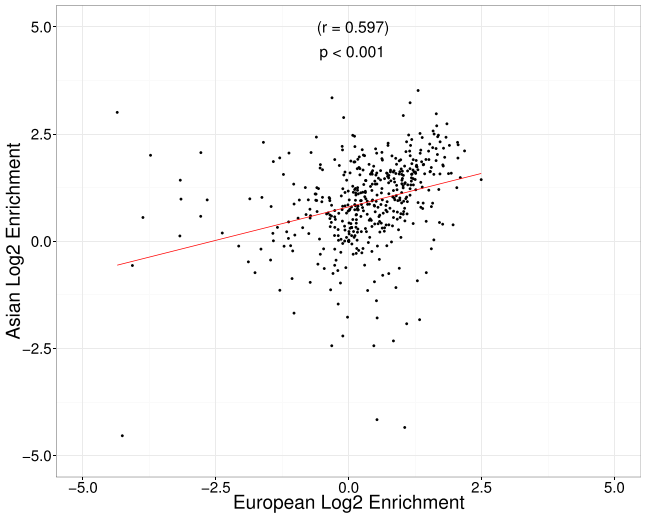

Figure 4. Functional Enrichment Is Consistent across Europeans and Asians
We compared the enrichment across 482 functional annotations at 89 RA-associated loci in Europeans (n ≈ 68,000) and Asians (n ≈ 36,000) separately. Each point represents the estimated enrichment of an annotation in both European and Asian populations.

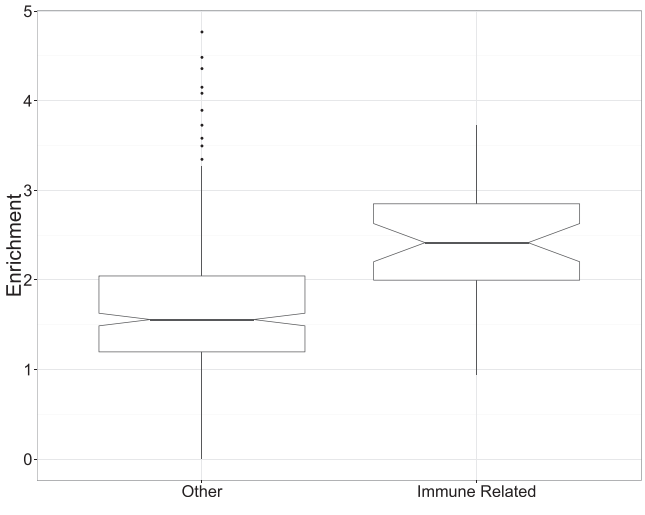

Figure 5. Trans-ethnic Functional Enrichment at RA GWAS Loci Indicates Immune-Related Regulatory Architecture
Here, we compare the enrichment of casual variants within 42 DHSs of immune-related cell types (B cells, T cells, natural killer cells, keratinocytes, monocytes, and thymic cells) and the enrichment of causal variants in 354 DHS annotations of other cell types. The widths of the notches in each boxplot roughly correspond to 95% confidence intervals for the median enrichment.

single functional prior can be applied across populations uniformly in trans-ethnic fine mapping.

Next, we estimated trans-ethnic enrichment for each of the 482 annotations independently to allow the model to discern the most functionally relevant cell types and classes. The enrichment likelihood ratios supplied by this procedure provide a natural way to prioritize functional annotations to move forward with fine mapping.[28] We consistently found a strong and significant enrichment of causal variants within activity regulatory regions of immune-related cell types (see Figure 5), which is largely in line with RA etiology (rank permutation $p < 0.001$). The final trans-ethnic integrative model included annotations of DHS regions specific to three cell types (skin keratinocytes, T helper 2 cells, and B lymphocytes), immune enhancer regions described in Farh et al.,[10] and GENCODE-defined exon regions. We found that simply applying existing multi-causal frameworks[27,28] on the trans-ethnic meta-analysis statistics yielded wider 90% credible sets (it required approximately 28.5 SNPs per locus as opposed to 24.0 SNPs per locus for our proposed framework), thus demonstrating the benefit of modeling population-level LD. Furthermore, the integration of functional data additionally reduced the size of the credible set to 21.7 SNPs per locus (see Table 3), showing that leveraging functional annotations refines trans-ethnic fine-mapping signal.

Next, we explored the plausible causality of the SNPs that attained a high posterior probability under our framework (Table 4). For example, rs968567, which lies within the promoter region of *FADS2* (OMIM: 606149) and was functionally validated to disrupt transcription factor binding and subsequent gene expression,[52] achieved a trans-ethnic posterior probability of 0.29. However, this variant fell within all five functional annotations that our framework deemed important for this trait and, upon appropriate re-weighting, achieved a posterior probability of 0.84. Alternatively, trans-ethnic association can be extremely beneficial on its own. For example, rs12693993, a variant within the coding region of *CD28* (OMIM: 186760), a gene implicated for its importance in T cell development and proliferation and cytokine production, achieved a posterior probability for causality of 0.34 and 0.02 in Europeans and Asians, respectively. However, upon integration of trans-ethnic association with functional data, it achieved a posterior probability for causality of 0.85. The identification of these two SNPs, among others, serves as an important illustration of the benefit of our proposed methodologies.

Discussion

In this work, we introduced a fine-mapping framework that bridges several sources of evidence to prioritize functional SNPs and demonstrated its efficacy in real and simulated datasets. As fine-mapping data become increasingly multi-ethnic[3,4] and functional data become larger and more refined,[30] we believe that our proposed methodology will have increasing relevance. By operating exclusively on summary data, our approach reduces the need to share individual data, which often prohibits large-scale analyses. In addition, a key advantage of our proposed methodology is that it provides an unbiased perspective on which functional genomic data are most relevant to the trait within an Empirical Bayes framework. Rather than relying on careful and manual selection of functional elements when conducting fine mapping,[10,36] we allow the data to dictate

Table 3. Integrative Approaches that Model Population-Level LD Yield the Smallest Credible Sets in Empirical Data

Association Statistics	Average No. of SNPs	
	Without Annotations	With Annotations
Asians	35.2	31.9
Europeans	32.0	28.7
Fixed-effects meta-analysis	28.5	25.0
Trans-ethnic	24.0	21.7

Displayed here is the average number of SNPs per locus in the 90% credible sets for single and multi-population fine mapping of RA-associated loci. To compute credible sets, we first ordered the SNPs across all 89 loci and then took the total number of ordered SNPs that consumed 90% of the total posterior probability mass. Consistent with simulation findings, integrating multiple populations with functional data improved fine-mapping resolution.

the functional relevance of a particular annotation. As the catalog of functional data expands to encompass more diverse cell types and genomic signatures, a principled strategy to parsing these annotations is paramount.

We note that although our model does not assume a priori that there exists allelic heterogeneity by ancestry,[15] by construction, it is capable of handling trans-ethnic heterogeneity whether it is due to a true difference in the per-allelic effects or simply a result of genetic drift that yielded distinct allele frequencies at the causal SNPs. We have found that as the level of heterogeneity across populations increases, our framework increasingly outperforms competing strategies. Although extreme heterogeneity might be unlikely, gene-environment interactions in complex traits can manifest themselves as distinct allelic effects across populations.[53]

We conclude with several limitations of our proposed framework. The efficacy of our proposed method is intimately connected to the underlying functional architecture of the trait being examined. In the scenario where the correct functional annotation is unavailable or the distribution of casual variants is more or less uniform across the functional-annotation categories, our method will most likely underperform fine-mapping strategies that either do not estimate parameters for functional enrichment[27,46] or pre-specify the correct enrichment parameters from other external analyses.[35] However, there is mounting evidence that suggests that casual variants for most complex traits co-localize with epigenetic marks[10,31,35,37] that are now available for the vast majority of human cell types.[54] Finally, additional improvements in performance could be made through a Bayesian treatment of non-centrality parameters within our framework,[46] which we leave as a potential direction for future work.

Appendix A: Optimization Procedure

We optimize the parameters of our model by using expectation maximization. First, we take expectations of the complete data log-likelihood with respect to the posterior

Table 4. Integrating Trans-ethnic Association Strength with Functional Data Promotes a Number of SNPs to Attain a High Posterior Probability for Causality

rsID	Chromosomal Position	European Association (*Z* Score)	Asian Association (*Z* Score)	Posterior Probability without Annotations	Posterior Probability with Annotations	Functional Annotations
rs2476601	chr1: 114,377,568	−26.04	NA	1.00	1.00	coding exons, skin keratinocyte DHSs
rs7731626	chr5: 55,444,683	−9.84	NA	1.00	1.00	GM12865 DHSs, Th2 DHSs, immune enhancers
NA	chr1: 2,523,878	−5.22	−4.18	1.00	1.00	immune enhancers
rs1893592	chr21: 43,855,067	−5.73	−4.01	1.00	1.00	coding exons, immune enhancers
NA	chr19: 10,771,941	−6.13	NA	1.00	1.00	immune enhancers
rs72767222	chr5: 55,440,788	5.11	NA	0.99	0.99	skin keratinocyte DHSs, immune enhancers
rs12715125	chr3: 27,763,427	5.58	NA	0.95	0.99	coding exons, GM12865 DHSs, Th2 DHSs, skin keratinocyte DHSs, immune enhancers
rs71508903	chr10: 63,779,871	7.26	5.88	0.76	0.93	GM12865 DHSs, skin keratinocyte DHSs, immune enhancers
rs12693993	chr2: 204,595,597	−2.74	−1.76	0.68	0.88	Th2 DHSs, skin keratinocyte DHS, immune enhancers
rs968567	chr11: 61,595,564	−4.95	NA	0.29	0.85	coding exons, GM12865 DHSs, Th2 DHSs, skin keratinocyte DHSs, immune enhancers
rs909685	chr22: 39,747,671	6.29	4.62	0.65	0.84	Th2 DHSs, skin keratinocyte DHSs, immune enhancers
rs657075	chr5: 131,430,118	2.54	4.46	0.73	0.82	skin keratinocyte DHSs, immune enhancers

We applied our framework across all 89 GWAS RA loci with relevant functional data. Displayed in this table are SNPs achieving a trans-ethnic posterior probability of greater than 0.8. Abbreviations are as follows: NA, not applicable; Th2, T helper 2 cell.

distribution of causal sets and simplify to obtain a function, Q, that is readily optimized via standard techniques. Let $Z_{l,*}$ represent all P vectors of association statistics $(Z_{l,1}, Z_{l,2}, \cdots, Z_{l,P})$ at locus l, and let $\lambda_{l,*}$ be the corresponding vectors of non-centrality parameters,

$$\begin{aligned} Q(\gamma, \lambda \mid \gamma^{(t)}, \lambda) &= \sum_l \sum_{C_l} P(C_l \mid Z_{l,*}\lambda_{l,*}, \gamma^{(t)}) \ln P(Z_{l,*}; \lambda_{l,*}, \gamma^{(t)}) \\ &= \sum_l \sum_{C_l} P(C_l \mid Z_{l,*}\lambda_{l,*}, \gamma^{(t)}) \left(\ln P(C_l; \gamma^{(t)}) + \sum_p \ln P(Z_{l,p} \mid C_l, \lambda_{l,p}) \right) \\ &= \sum_l \sum_{C_l} P(C_l \mid Z_{l,*}\lambda_{l,*}, \gamma^{(t)}) \ln P(C_l; \gamma^{(t)}) + \sum_l \sum_{C_l} P(C_l \mid Z_{l,*}, \gamma^{(t)}, \lambda_{l,*}) \sum_p \ln P(Z_{l,p} \mid C_l, \lambda_{l,p}) \\ &= Q(\gamma \mid \gamma^{(t)}) + Q(\lambda_p \mid \lambda_p), \end{aligned}$$

thereby decoupling the prior from the likelihood. We simplify $Q(\gamma|\gamma^{(t)})$ to obtain

$$\begin{aligned} Q(\gamma \mid \gamma^{(t)}, \lambda) &= \sum_l \sum_j \sum_{c_{jl} \in 0,1} P(c_{jl} \mid Z_{l,*}; \gamma^{(t)}, \lambda_{l,*}) \ln P(c_{jl}; \gamma^{(t)}) \\ &= -\sum_l \sum_j P(c_{jl} = 1 \mid Z_{l,*}; \gamma^{(t)}, \lambda_{l,*}) \ln\left(1 + \exp(-\gamma^T A_{jl})\right) \\ &\quad -\sum_l \sum_j P(c_{jl} = 0 \mid Z_{l,*}; \gamma^{(t)}, \lambda_{l,*}) \ln\left(1 + \exp(\gamma^T A_{jl})\right), \end{aligned}$$

which is a concave function whose gradient is simply

$$\begin{aligned} \frac{\partial Q(\gamma \mid \gamma^{(t)}, \lambda)}{\partial \gamma} &= \sum_j \sum_l P(c_{jl} = 1 \mid Z_{l,*}; \gamma^{(t)}, \lambda_{l,*}) \frac{1}{1 + \exp(-\gamma^T A_{jl})} A_{jl} \\ &\quad - \sum_j \sum_l P(c_{jl} = 0 \mid Z_{l,*}; \gamma^{(t)}, \lambda_{l,*}) \frac{1}{1 + \exp(\gamma^T A_{jl})} A_{jl}. \end{aligned}$$

To avoid potential numerical instability resulting from inverting a Hessian matrix, as would be required for standard Newton-Raphson, we optimize this function Q by using a limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm implemented in the NLopt library. Finally, as previously mentioned, the non-centrality parameter for SNP j at locus l from population p, $\lambda^j_{p,l}$, is set simply as

$$f\left(Z^j_{p,l}\right) = \begin{cases} \arg\min\left(-3.7, Z^j_{p,l}\right) & \text{if} Z^j_{p,l} < 0 \\ \arg\max\left(3.7, Z^j_{p,l}\right) & \text{if} Z^j_{p,l} > 0 \\ 0 & \text{if } Z^j_{p,l} = 0 \text{ (SNP } j \text{ is monomorphic in population } p\text{)}, \end{cases}$$

a strategy that was previously demonstrated to work well in practice.[28] This iterative algorithm is repeated until the change in the log-likelihood is less than 0.01.

Acknowledgments

We would like to thank Alkes Price, Hillary Finucane, Robert Brown, Huwenbo Shi, and Nicholas Mancuso for helpful discussion and feedback on this work. This research was supported by NIH grants R01-HG006399, R01-GM53275, and R21-CA182821.

Received: April 6, 2015
Accepted: June 9, 2015
Published: July 16, 2015

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes, http://www.1000genomes.org/data
DHS maps 1, http://www.uwencode.org/proj/Science_Maurano_Humbert_et_al/
DHS maps 2, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgDnaseUniform/
ENCODE Genome Segmentation, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/
Ensembl, http://www.ensembl.org/index.html
FANTOM 5, http://fantom.gsc.riken.jp/5/data/
Finemapping Data Portal, http://www.broadinstitute.org/pubs/finemapping/?q=data-portal
OMIM, http://www.omim.org
PAINTOR, http://bogdan.bioinformatics.ucla.edu/software/paintor/

RA GWAS summary statistics, http://plaza.umin.ac.jp/~yokada/datasource/software.htm

References

1. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.
2. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al.; Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.
3. Mahajan, A., Go, M.J., Zhang, W., Below, J.E., Gaulton, K.J., Ferreira, T., Horikoshi, M., Johnson, A.D., Ng, M.C., Prokopenko, I., et al.; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium; Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; Mexican American Type 2 Diabetes (MAT2D) Consortium; Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples (T2D-GENES) Consortium (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat. Genet. *46*, 234–244.
4. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al.; RACI consortium; GARNET consortium (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature *506*, 376–381.
5. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.
6. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al.; LifeLines Cohort Study; ADIPOGen Consortium; AGEN-BMI Working Group; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GLGC; ICBP; MAGIC Investigators; MuTHER Consortium; MIGen Consortium; PAGE Consortium; ReproGen Consortium; GENIE Consortium; International Endogene Consortium (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197–206.
7. Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Mägi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Justice, A.E., et al.; ADIPOGen Consortium; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GEFOS Consortium; GENIE Consortium; GLGC; ICBP; International Endogene Consortium; LifeLines Cohort Study; MAGIC Investigators; MuTHER Consortium; PAGE Consortium; ReproGen Consortium (2015). New genetic loci link adipose and insulin biology to body fat distribution. Nature *518*, 187–196.
8. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.
9. Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., et al.; National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC); United Kingdom Inflammatory Bowel Disease Genetics Consortium; International Inflammatory Bowel Disease Genetics Consortium (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. *43*, 1066–1073.
10. Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.
11. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al.; International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
12. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
13. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E., and Halperin, E. (2010). Leveraging genetic variability across populations for the identification of causal variants. Am. J. Hum. Genet. *86*, 23–33.
14. Ong, R.T.-H., Wang, X., Liu, X., and Teo, Y.-Y. (2012). Efficiency of trans-ethnic genome-wide meta-analysis and fine-mapping. Eur. J. Hum. Genet. *20*, 1300–1307.
15. Morris, A.P. (2011). Transethnic meta-analysis of genomewide association studies. Genet. Epidemiol. *35*, 809–822.
16. Teo, Y.-Y., Ong, R.T., Sim, X., Tai, E.S., and Chia, K.-S. (2010). Identifying candidate causal variants via trans-population fine-mapping. Genet. Epidemiol. *34*, 653–664.
17. Udler, M.S., Meyer, K.B., Pooley, K.A., Karlins, E., Struewing, J.P., Zhang, J., Doody, D.R., MacArthur, S., Tyrer, J., Pharoah, P.D., et al.; SEARCH Collaborators (2009). FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. Hum. Mol. Genet. *18*, 1692–1703.
18. Stacey, S.N., Sulem, P., Zanon, C., Gudjonsson, S.A., Thorleifsson, G., Helgason, A., Jonasdottir, A., Besenbacher, S., Kostic, J.P., Fackenthal, J.D., et al. (2010). Ancestry-shift refinement mapping of the C6orf97-ESR1 breast cancer susceptibility locus. PLoS Genet. *6*, e1001029.
19. Evangelou, E., and Ioannidis, J.P. (2013). Meta-analysis methods for genome-wide association studies and beyond. Nat. Rev. Genet. *14*, 379–389.
20. Wang, X., Chua, H.-X., Chen, P., Ong, R.T.-H., Sim, X., Zhang, W., Takeuchi, F., Liu, X., Khor, C.-C., Tay, W.-T., et al. (2013). Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. Hum. Mol. Genet. *22*, 2303–2311.
21. Liu, C.-T., Buchkovich, M.L., Winkler, T.W., Heid, I.M., Borecki, I.B., Fox, C.S., Mohlke, K.L., North, K.E., and Adrienne Cupples, L.; African Ancestry Anthropometry Genetics Consortium; GIANT Consortium (2014). Multi-ethnic fine-mapping of 14 central adiposity loci. Hum. Mol. Genet. *23*, 4738–4744.
22. Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M., Auton, A., Myers, S., Morris, A., et al.; Wellcome Trust Case Control Consortium (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat. Genet. *44*, 1294–1301.

23. Beecham, A.H., Patsopoulos, N.A., Xifara, D.K., Davis, M.F., Kemppinen, A., Cotsapas, C., Shah, T.S., Spencer, C., Booth, D., Goris, A., et al.; International Multiple Sclerosis Genetics Consortium (IMSGC); Wellcome Trust Case Control Consortium 2 (WTCCC2); International IBD Genetics Consortium (IIBDGC) (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat. Genet. *45*, 1353–1360.
24. Onengut-Gumuscu, S., Chen, W.-M., Burren, O., Cooper, N.J., Quinlan, A.R., Mychaleckyj, J.C., Farber, E., Bonnie, J.K., Szpak, M., Schofield, E., et al.; Type 1 Diabetes Genetics Consortium (2015). Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat. Genet. *47*, 381–386.
25. Meyer, K.B., O'Reilly, M., Michailidou, K., Carlebur, S., Edwards, S.L., French, J.D., Prathalingham, R., Dennis, J., Bolla, M.K., Wang, Q., et al.; GENICA Network; kConFab Investigators; Australian Ovarian Cancer Study Group (2013). Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. Am. J. Hum. Genet. *93*, 1046–1060.
26. Trynka, G., Hunt, K.A., Bockett, N.A., Romanos, J., Mistry, V., Szperl, A., Bakker, S.F., Bardella, M.T., Bhaw-Rosun, L., Castillejo, G., et al.; Spanish Consortium on the Genetics of Coeliac Disease (CEGEC); PreventCD Study Group; Wellcome Trust Case Control Consortium (WTCCC) (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat. Genet. *43*, 1193–1201.
27. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497–508.
28. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. *10*, e1004722.
29. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
30. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
31. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.
32. Trynka, G., and Raychaudhuri, S. (2013). Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases. Curr. Opin. Genet. Dev. *23*, 635–641.
33. Karczewski, K.J., Dudley, J.T., Kukurba, K.R., Chen, R., Butte, A.J., Montgomery, S.B., and Snyder, M. (2013). Systematic functional regulatory assessment of disease-associated variants. Proc. Natl. Acad. Sci. USA *110*, 9607–9612.
34. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. *45*, 124–130.
35. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. *95*, 535–552.
36. Hazelett, D.J., Rhie, S.K., Gaddis, M., Yan, C., Lakeland, D.L., Coetzee, S.G., Henderson, B.E., Noushmehr, H., Cozen, W., Kote-Jarai, Z., et al.; Ellipse/GAME-ON consortium; Practical consortium (2014). Comprehensive functional annotation of 77 prostate cancer risk loci. PLoS Genet. *10*, e1004102.
37. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.
38. Chung, D., Yang, C., Li, C., Gelernter, J., and Zhao, H. (2014). GPA: A statistical approach to prioritizing GWAS results by integrating pleiotropy information and annotation data. arXiv, arXiv:1401.4764, http://arxiv.org/abs/1401.4764v1.
39. Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D.P., Patterson, N., and Price, A.L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics *30*, 2906–2914.
40. Tikhonov, A.N., and Arsenin, V.Y. (1977). Solutions of Ill-Posed Problems (John Wiley & Sons).
41. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304–2305.
42. Coram, M.A., Duan, Q., Hoffmann, T.J., Thornton, T., Knowles, J.W., Johnson, N.A., Ochs-Balcom, H.M., Donlon, T.A., Martin, L.W., Eaton, C.B., et al. (2013). Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. Am. J. Hum. Genet. *92*, 904–916.
43. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. *43*, 519–525.
44. Nelis, M., Esko, T., Mägi, R., Zimprich, F., Zimprich, A., Toncheva, D., Karachanak, S., Piskácková, T., Balascák, I., Peltonen, L., et al. (2009). Genetic structure of Europeans: a view from the North-East. PLoS ONE *4*, e5472.
45. Franceschini, N., van Rooij, F.J., Prins, B.P., Feitosa, M.F., Karakas, M., Eckfeldt, J.H., Folsom, A.R., Kopp, J., Vaez, A., Andrews, J.S., et al.; LifeLines Cohort Study (2012). Discovery and fine mapping of serum protein loci through transethnic meta-analysis. Am. J. Hum. Genet. *91*, 744–753.
46. Chen, W., Larrabee, B.R., Ovsyannikova, I.G., Kennedy, R.B., Haralambieva, I.H., Poland, G.A., and Schaid, D.J. (2015). Fine mapping causal variants with an approximate bayesian method using marginal test statistics. Genetics. Published online May 6, 2015. http://dx.doi.org/10.1534/genetics.115.176107.
47. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.
48. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

49. Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., Itoh, M., et al.; FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.
50. Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2015). Ensembl 2015. Nucleic Acids Res. *43*, D662–D669.
51. Marigorta, U.M., and Navarro, A. (2013). High trans-ethnic replicability of GWAS results implies common causal variants. PLoS Genet. *9*, e1003566.
52. Lattka, E., Eggers, S., Moeller, G., Heim, K., Weber, M., Mehta, D., Prokisch, H., Illig, T., and Adamski, J. (2010). A common FADS2 promoter polymorphism increases promoter activity and facilitates binding of transcription factor ELK1. J. Lipid Res. *51*, 182–191.
53. Malaria Genomic Epidemiology Network; Malaria Genomic Epidemiology Network (2014). Reappraisal of known malaria resistance loci in a large multicenter study. Nat. Genet. *46*, 1197–1204.
54. Ernst, J., and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol. *33*, 364–376.