Please use this page to enter your 'general comments' once the discussion phase ends (on Friday at 8pm)
Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits

Joseph K. Pickrell1,2,*

Annotations of gene structures and regulatory elements can inform genome-wide association studies (GWASs). However, choosing the relevant annotations for interpreting an association study of a given trait remains challenging. I describe a statistical model that uses association statistics computed across the genome to identify classes of genomic elements that are enriched with or depleted of loci influencing a trait. The model naturally incorporates multiple types of annotations. I applied the model to GWASs of 18 human traits, including red blood cell traits, platelet traits, glucose levels, lipid levels, height, body mass index, and Crohn disease. For each trait, I used the model to evaluate the relevance of 450 different genomic annotations, including protein-coding genes, enhancers, and DNase-I hypersensitive sites in over 100 tissues and cell lines. The fraction of phenotype-associated SNPs influencing protein sequence ranged from around 2% (for platelet volume) up to around 20% (for low-density lipoprotein cholesterol), repressed chromatin was significantly depleted for SNPs associated with several traits, and cell-type-specific DNase-I hypersensitive sites were enriched with SNPs associated with several traits (for example, the spleen in platelet volume). Finally, reweighting each GWAS by using information from functional genomics increased the number of loci with high-confidence associations by around 5%.

Introduction

A fundamental goal of human genetics is to create a catalog of the genetic polymorphisms that cause phenotypic variation in our species and to characterize the precise molecular mechanisms by which these polymorphisms exert their effects. An important tool in the modern human genetics toolkit is the genome-wide association study (GWAS), in which hundreds of thousands or millions of SNPs are genotyped in large cohorts of individuals and each polymorphism is tested for a statistical association with some trait of interest. In recent years, GWASs have identified thousands of genomic regions that show reproducible statistical associations with a wide array of phenotypes and diseases.1

In general, the loci identified in GWASs of multifactorial traits have small effect sizes and are located outside of protein-coding exons.2 This latter fact has generated considerable interest in annotating other types of genomic elements apart from exons. For example, the ENCODE project has generated detailed maps of histone modifications and transcription factor binding in six human cell lines, partly to interpret GWAS signals that might act via a mechanism of gene regulation.3 Methods for combining potentially rich sources of functional genomic data with GWASs could in principle lead to important biological insights. The development of such a method is the aim of this paper.

There are two lines of research that have motivated my work on this problem. The first is what are often called “enrichment” analyses. In this type of analysis, the researcher examines the most strongly associated SNPs in a GWAS and tests whether they fall disproportionately in specific types of genomic regions. These studies have found, for example, that SNPs identified in GWASs are enriched in protein-coding exons, in promoters, in UTRs,4 and among those that influence gene expression.5 Further, in some cases, SNPs associated with a trait are enriched in gene regulatory regions in specific cell types7–18 or near genes expressed in specific cell types.19,20 However, the methods in these studies are generally not able to consider more than a single annotation at a time (with a few exceptions21,22). Further, they are not set up to answer a question that I find important: consider two independent SNPs with equivalent p values of $1 \times 10^{-7}$ in a GWAS for some trait (note that this p value does not reach the standard threshold of $5 \times 10^{-8}$ for “significance”); the first is a nonsynonymous SNP, and the second falls far from any known gene. What is the probability that the first SNP is truly associated with the trait, and how does this compare to the probability for the second?

A potential answer to this question comes from the second line of research that motivates this work. In association studies where the phenotype being studied is gene expression (studies of expression quantitative trait loci [eQTLs]), statistical models have been developed to identify shared characteristics of SNPs that influence gene expression.23–25 In a hierarchical modeling framework, the probability that a given SNP influences gene expression can then depend on these characteristics. The key fact that makes these models useful in the context of eQTL mapping is that the genome contains a large number of unambiguous eQTLs on which a model can be trained. In the GWAS context, the number of loci unambiguously associated with a given trait has historically been very small; learning the shared properties of two or three loci is not a job well suited to statistical modeling. However, large meta-analyses of GWASs now regularly identify tens 1New York Genome Center, New York, NY 10013, USA; 2Department of Biological Sciences, Columbia University, New York, NY 10027, USA
*Correspondence: jpickrell@nygenome.org
http://dx.doi.org/10.1016/j.ajhg.2014.03.004. ©2014 by The American Society of Human Genetics. All rights reserved.
to hundreds of independent loci influencing a trait (e.g., Lango-Allen et al.\textsuperscript{26} and Teslovich et al.\textsuperscript{27}). The merits of hierarchical modeling in this context\textsuperscript{28-30} are thus worth revisiting. Indeed, Carbonetto and Stephens\textsuperscript{31} have reported success in identifying loci involved in autoimmune diseases by using a hierarchical model that incorporates information about groups of genes known to interact in a pathway.

In this paper, I present a hierarchical model for jointly analyzing GWASs and genomic annotations. I applied this model to GWASs of 18 diseases and traits; for each trait, I used the model to learn the relevant types of genomic information from a set of 450 genome annotations.

**Material and Methods**

In this section, I detail the specifics of the hierarchical model; details of the data used are in Appendix A. The proposed model is most closely related to that developed by Veyrieras et al.\textsuperscript{23} in the context of eQTL mapping. Conceptually, I split the genome into independent blocks, such that the blocks are larger than the extent of linkage disequilibrium (LD) in the population. Each block is allowed to contain either a single polymorphism that causally influences the trait or none. I model the prior probability that any given block contains an association and the conditional prior probability that any given SNP in the block is the causal one. The key is that these probabilities are allowed to vary according to functional annotations—for example, gene-rich regions might be more likely to contain associations, and if there is an association, the causal polymorphism might be more likely to fall in a transcription factor binding site. I then estimate these priors by using an empirical Bayes approach. Software implementing this model is available from GitHub (see Web Resources).

**Computing the Bayes factor**

The basic building block of the model is a linear regression model. Consider a single SNP genotyped in $N$ phenotyped individuals. Assume that each individual has an associated measurement of a quantitative trait (I describe a slight modification for case-control studies later), and let $\mathbf{y}$ be the vector of phenotypes. Let $\mathbf{z}$ be the vector of genotypes (coded 0, 1, or 2 according to counts of an arbitrarily defined allele). I use a standard additive linear model:

$$E[y_i] = \alpha + \beta z_i.$$

(Equation 1)

I would like to compare two models: one where $\beta = 0$ and one where $\beta \neq 0$. A natural way to compare these two models is the Bayes factor,

$$BF = \frac{\int P(\mathbf{y} | \mathbf{z}, H_1) \, d\mathbf{z}}{\int P(\mathbf{y} | \mathbf{z}, H_0) \, d\mathbf{z}}.$$  

(Equation 2)

where $H_1$ and $H_0$ represent the parameters of the alternative and null models, respectively, and are integrated out. To compute Equation 2, I use the approximate Bayes factor from Wakefield.\textsuperscript{32} This Bayes factor has the practically important property that it can be calculated from a summary of the linear regression without access to the underlying genotype vector $\mathbf{z}$. For completeness, I reiterate here the underlying model. If $\hat{\beta}$ is the maximum-likelihood estimator of $\beta$ and $\sqrt{V}$ is the SE of $\hat{\beta}$, Wakefield\textsuperscript{32} suggests a model in which

$$\hat{\beta} \sim N(\hat{\beta}, V).$$

(Equation 3)

Wakefield\textsuperscript{32} places a normal prior on $\beta$, such that $\beta \sim N(0, W)$. Under this model, Equation 2 becomes

$$BF = \frac{\sqrt{1 - r}}{\exp \left[ \frac{Z^2}{2r} \right]}.$$  

(Equation 4)

where $r = W/(V + W)$ and $Z = \hat{\beta}/\sqrt{V}$ (a standard Z-score). Thus, from a Z-score, an estimate of $V$, and the prior variance $W$, one can obtain a Bayes factor measuring how the statistical support for a model in which a SNP is associated with a trait compares to that for a model in which a SNP is not associated with a trait. Note, however, that because of LD in the genome, any true causal association will lead to multiple true statistical associations. In all applications, I set $W = 0.1$ as the prior, such that the majority of the weight of the prior is on small effect sizes (results are robust to some variation in this prior; Figure S16, available online).

**Hierarchical Model**

Now consider a set of $M$ SNPs, each of which has been genotyped in $N$ individuals in a GWAS. The goal is to build a model to identify the shared characteristics of SNPs that causally influence a trait. Because of LD, there will be many genomic associations that are not causal; however, these will all be restricted to a block around the truly causal site. I thus split the genome into contiguous blocks of size $K$ SNPs (in all applications, I set $K = 5,000$, although doubling this block size had little effect on the results; Figure S16), such that there are $M/K$ blocks. The block size is chosen to be much larger than the extent of LD in the population. Let $\Pi_k$ be the prior probability that block $k$ contains a causal SNP associated with the trait. The probability of the data (the set of observed phenotypes) is then

$$P(\mathbf{y}) = \prod_{k=1}^{M/K} (1 - \Pi_k)P_k^0 + \Pi_k P_k^1,$$

(Equation 5)

where $P_k^0$ is the probability of the data in block $k$ under the model where there are no SNPs associated with the trait in the block and $P_k^1$ is the probability of the data in block $k$ under the model where there is one SNP associated with the trait in the block. Further,

$$P_k^1 = \sum_{i \in S_k} \pi_i P_{ik}^1,$$

(Equation 6)

where $S_k$ is the set of SNPs in block $k$, $\pi_i$ is the prior probability (conditional on there being an association in block $k$) that SNP $i$ is the causal SNP in the region, and $P_{ik}^1$ is the probability of the data under the model where this SNP is associated with the trait. Note that this is not a multiple regression model that jointly models the effects of multiple SNPs on a trait (as in, for example, Carbonetto and Stephens\textsuperscript{31}).

It is now possible to allow the prior probabilities—both $\Pi_k$ (the prior probability on the block of SNPs containing an association) and $\pi_i$ (the prior probability that SNP $i$ is the causal SNP under
the assumption that there is a single association in block k)—to
depend on external information. One would also like to avoid sub-
vective variation in PI_k and π_k but instead learn from the data itself
which genomic annotations are most important. Specifically, I
model the regional prior probability as

$$\ln \left( \frac{\Pi_k}{1 - \Pi_k} \right) = \kappa + \sum_{l=1}^{L_1} \gamma_l I_{kl}.$$  \hspace{1cm} (Equation 7)

where L_1 is the number of region-level annotations in the
model, γ_l is the effect associated with annotation l, and I_{kl} takes
the value 1 if region k is annotated with annotation l and 0 oth-
wise. For example, in practice I will estimate a γ parameter for
regions of high or low gene density. I then model the SNP prior
probability as

$$\pi_{ik} = \frac{e^{x_i}}{\sum_{j=1}^L e^{x_j}},$$  \hspace{1cm} (Equation 8)

where

$$x_i = \sum_{l=1}^{L_2} \lambda_l I_{il},$$  \hspace{1cm} (Equation 9)

where L_2 is the number of SNP-level annotations in the model, λ_l is
the effect of SNP annotation l, and I_{il} takes the value 1 if SNP i falls
in annotation l and 0 otherwise. For example, in practice I will
estimate a λ parameter for nonsynonymous SNPs.

**Fitting the Model**

When the terms above are combined, the likelihood of the data
can be written down as

$$L(\bar{y} \mid \theta) = \prod_{k=1}^{M} P(1 - \Pi_k) + \Pi_k \sum_{l=1}^{L_1} \pi_l P_{l}^{\mathcal{F}}$$

$$L(\bar{y} \mid \theta) = \prod_{k=1}^{M} P(1 - \Pi_k) + \Pi_k \sum_{l=1}^{L_1} \pi_l P_{l}^{\mathcal{F}}.$$  \hspace{1cm} (Equation 10)

$$L(\bar{y} \mid \theta) = \prod_{k=1}^{M} P(1 - \Pi_k) + \Pi_k \sum_{l=1}^{L_1} \pi_l P_{l}^{\mathcal{F}}.$$  \hspace{1cm} (Equation 11)

where θ contains all the parameters of the model, most notably the set
of annotation parameters. I maximize this function by using the
Nelder-Mead algorithm implemented in the GNU Scientific Library.

**Shrinkage Estimators of the Annotation Parameters**

Although maximizing Equation 10 gives the maximum-likelihood
estimates of all parameters, one concern is that there might be some
level of overfitting. When comparing models, I instead shrink these
parameters toward 0. Specifically, I define a penalized
log-likelihood function as

$$I(\bar{y} \mid \theta) = \ln(L(\bar{y} \mid \theta)) - p \left( \sum_{l=1}^{L_1} \gamma_l^2 + \sum_{l=1}^{L_2} \lambda_l^2 \right).$$  \hspace{1cm} (Equation 12)

The penalty p on the sum of the squared annotation parameters is
the one used in ridge regression.\(^{35}\) In ridge regression, parameter
estimates under this penalty are equivalent to the posterior means of
the parameter if the prior distribution of the parameter is Gaussian;\(^{35}\) changing the tuning parameter p is equivalent to
changing the prior. I suspect that the interpretation in this model
is similar. Given that this penalized likelihood cannot be used for
formal statistical tests, I tune the p parameter by cross-
validation. An alternative approach here would be to explicitly
put a prior on the enrichment parameters, but in the absence of
a conjugate prior, this would most likely add substantially to the
computational burden for little practical benefit.

**Cross-validation**

To compare models and tune the penalty p in the penalized
likelihood above, I used a 10-fold cross-validation approach. I split
the chromosomal segments into 10 folds. Let \(\theta_f^p\) be
the parameters of the model estimated without the data from fold f
and under penalty p, and let \(I_f(\theta_f^p)\) be the penalized log-likelihood
of the data in fold f under the model optimized without fold f.
Then,

$$I_f(\theta_f^p) = \frac{1}{10} \sum_{f=1}^{10} I_f(\theta_f^p).$$  \hspace{1cm} (Equation 13)

Note that the size of the folds used in this cross-validation means
that each fold excludes more than an entire chromosome. This
means that no individual chromosome can have undue influence
on the parameters included in the model.

**Model Choice**

Consider a single phenotype and a set of L functional annotations
of SNPs (in applications, L is in the hundreds). Including all L SNP
annotations in the model is neither biologically interesting nor
computationally feasible. I thus set out to choose a relatively
sparse model that fits the data. I start with forward selection: for
each of the L annotations, I fit a model including a region-level parameter for regions in the top third of the distribution of gene
density, a region-level parameter for regions in the bottom third of
the distribution of gene density, a SNP-level parameter for
SNPs from 0 to 5 kb from a transcription start site (TSS), a SNP-level
parameter for SNPs from 5 to 10 kb from a TSS, and a SNP-level
parameter for the annotation in question. I then identify the set
of annotations that significantly improve the model fit (as judged
by the likelihood from Equation 10). I then

1. Add the annotation that most significantly improves the
   likelihood to the model.
2. Test a model (including the annotation and those that
   have already been added) for each annotation identified as
   having a significant marginal effect.
3. Go back to step 1 if any annotation remains significant.

At this point, there are generally a small number of annotations
in the model, but the model might be overfit. I then switch to
using the 10-fold cross-validation likelihood in Equation 13. I first
tune the penalty parameter p by finding the value of p that maxi-
mizes the cross-validation likelihood. I then

1. Drop each annotation from the model in turn and evaluate
   the cross-validation likelihood. When dropping annota-
tions, I additionally try dropping the region-level annota-
tions on gene density and the SNP-level annotations on
distance to the nearest TSS.
2. Drop the annotation from the model and return to step 1 if
   a simpler model has a higher cross-validation likelihood
   than the full model.
3. Report the model that has the highest cross-validation likeli-
hood.
Approximating $V_i$
In order to compute the Bayes factor in Equation 4, one needs an estimate of $V_i$, the variance of the estimated effect size of SNP $i$. In principle, this is trivial output from standard regression software; however, it is rarely reported. Instead, $f_i$ is the minor allele frequency of SNP $i$ and is computed from an external sample of the same ancestry as the population in which the association study was done (I used data from the 1000 Genomes Project). Let $N_i$ be the number of individuals in the association study at SNP $i$ (this can vary across SNPs because of missing data). Then,

$$V_i = \frac{1}{N_i f_i (1 - f_i)}.$$  (Equation 14)

Note that this variance is independent of the actual scale of the measurements; this is appropriate because the $Z$-scores are independent of the scale of the measurements as well.

Case-Control Studies
For all of the above, I have considered studies of quantitative traits. For a case-control study, assume that one has summary statistics from logistic regression instead of linear regression. All aspects of the model are identical, except for the approximation of $V_i$. Define $N_{\text{case}}$ and $N_{\text{control}}$ as the numbers of case and control individuals, respectively. Now,

$$V_i = \frac{N_{\text{case}} + N_{\text{control}}}{2N_{\text{case}} N_{\text{control}} f_i (1 - f_i)}.$$  (Equation 15)

The variance here is on a log-odds scale.

Posterior Probabilities of Association
Once the model has been fit, it produces empirical estimates of the prior probability that region $k$ contains an association, $\hat{\pi}_k$, and of the prior probability that SNP $i$ is the causal one, $\hat{\pi}_i$ (on the condition that there is an association). Define a Bayes factor summarizing the evidence of association in the region (see, for example, Maller et al.35) as

$$BF_k^i = \sum_{\pi \in S_k} \pi_i BF_i,$$  (Equation 16)

where $S_k$ is the set of SNPs in region $k$ and $BF_i$ is the Bayes factor for SNP $i$ (Equation 4). The posterior probability that region $k$ contains an association (PPA for posterior probability of association) is then

$$PPA_k^i = \frac{\hat{\pi}_i BF_i^i / (1 - \hat{\pi}_k)}{1 + \Pi_k BF_k^i / (1 - \hat{\pi}_k)}.$$  (Equation 17)

One can also define the posterior probability that any given SNP $i$ in region $k$ is the causal one under the model

$$PPA_i^k = \frac{\pi_i BF_i}{\sum_{s \in S_k} \pi_s BF_s}.$$  (Equation 18)

This is similar to the calculation in Maller et al.35 except that the prior probability $\pi_a$ is allowed to vary across SNPs.

Finally, one can define the posterior probability that any given SNP is causal. This is the posterior probability that the region contains a causal SNP multiplied by the posterior probability (conditional on there being an association in the region) that the SNP is causal. If SNP $i$ falls in region $k$,

$$PPA_i = PPA_k^i PPA_\pi.$$  (Equation 19)

Conditional Analysis
Because many of the annotations I consider are correlated, those ultimately included in the combined model for each trait (Figure 4) might be representatives of a large group of correlated annotations. For biological interpretation of the model, it is thus important to know which of the other annotations are interchangeable with those included in the model.

To test this, I took an approach of conditional analysis. Consider two SNP-level annotations, with annotation parameters $\lambda_1$ and $\lambda_2$. In a joint model, one would jointly estimate both $\lambda_1$ and $\lambda_2$. However, it is interesting to know whether the second annotation adds information above and beyond that provided by the first annotation. I thus first estimate $\lambda_1$ and then fix this parameter to its maximum-likelihood value, $\hat{\lambda}_1$. I then estimate $(\hat{\lambda}_2 / \hat{\lambda}_1)$—that is, I obtain the maximum-likelihood estimate and 95% confidence interval (CI) of $\lambda_2$ conditional on a fixed value of $\lambda_1$. If this CI does not overlap 0, then this is evidence that the second annotation adds more information to the model than does the first annotation.

In practice, I first fit the combined model as described in the section "Model Choice" above. I then returned to the set of annotations that had significant marginal associations. For each annotation in the combined model, I took each of the other annotations in turn and tested whether the included annotation was significantly more informative than the nonincluded annotation. Figure 4 and Figure S12 display the total number of annotations represented by each one included in the combined model.

Imputation of Summary Statistics
I used ImpG v.1.0.16 under the default settings to impute summary statistics from all GWAS. I used as a reference panel all haplotypes from European individuals in phase 1 of the 1000 Genomes Project and only used SNPs with a minor allele frequency greater than 2%. The reference haplotype files were derived from the 1000 Genomes integrated phase 1 v.3.20101123 calls (Web Resources). I used all 379 individuals labeled "European." After imputation, I removed all imputed SNPs with a predicted accuracy (in terms of correlation with the true summary statistics) less than 0.8. Overall, for each GWAS, about 75%–80% of SNPs with a minor allele frequency over 10% were successfully imputed (Figure S1).

To verify that imputation did not induce inflation of the test statistics, I computed the genome-control inflation factor, $\lambda_{GC}$, before and after imputation (Table S1). In all studies, inflation decreased after imputation, sometimes leading to a marked deflation in the test statistics. This is consistent with previous observations from using this software38 and is caused by the imputation’s shrinkage prior, which is imposed for strict avoidance of false-positive associations.

Simulations
To test the performance of the model, I performed simulations that used data from a GWAS of height.28 Using the imputed summary statistics, I split the genome into blocks of 5,000 SNPs and then extracted the blocks with a genome-wide significant SNP
reported in Lango Allen et al.26 In each block, I had a reported Z-score for each SNP. To simulate annotations, I called the SNP with the smallest p value in the region the “causal” SNP. I then simulated annotations by placing all noncasual SNPs in an annotation with rate r1, and all casual SNPs in the annotation with rate r2. I also varied the numbers of blocks included in the model. In each simulation, I randomly assigned SNPs to annotations according to determined rates and then ran the model under the assumption that Πj = 1, that is, all blocks contain a causal SNP. I then calculated power as the fraction of simulations in which the CIs of the annotation effect did not overlap 0.

I chose parameter settings of r1 and r2 such that the enrichment factors were similar to those in observed data (log2 enrichment of 1.4 and 2.6). I chose r1 to be either 0.2 or 0.1. For each set of parameters, I simulated 100 annotations and ran the model separately on each. Shown in Figure S2 is the power of the model. As expected, power increased as r1 or the effect size increased and as the number of loci increased.

Robustness to Choice of Prior and Window Size
The model contains two parameters that are set by the user—the prior variance W on the effect size and the window size defining “independent” blocks of the genome. I empirically tested the robustness of the model to variation in these parameters by using the Crohn disease data set. I ran the model on each annotation by using W = 0.1 and W = 0.5 and additionally including (as in the main analyses) region-level parameters for regions in the top third and bottom third of the distribution of gene density and SNP-level parameters for SNPs located from 0 to 5 kb from a TSS and SNPs 5–10 kb from a TSS. Plotted in Figure S16A are these annotation parameter estimates for all annotations where the 95% CIs did not overlap 0 in at least one run. The estimates from the two runs with different priors are highly correlated. I additionally tested window sizes of 5,000 SNPs and 10,000 SNPs (both with W = 0.1). The annotation effect estimates from these two window sizes are plotted in Figure S16B and again are highly correlated.

Quantifying the Relative Roles of Coding versus Noncoding Changes in Each Phenotype
To generate Figure 3, I fit a model to each GWAS; this model included region-level annotations for regions in the top third and bottom third of the distribution of gene density and SNP-level annotations for nonsynonymous SNPs and SNPs within 5 kb of a TSS. Shown in Figure 3A are the estimates of the enrichment parameter for nonsynonymous SNPs. At each SNP, the result of this model is the posterior probability that the SNP is causal (see Equation 19). If this posterior probability at SNP i is PPAi, then the fraction of causal SNPs that are nonsynonymous, fNS,i, is

\[
 f_{\text{NS},i} = \frac{\sum_i \text{PPA}_i I_{\text{NS},i}}{\sum_i \text{PPA}_i}.
\]

(Equation 20)

where \( I_{\text{NS},i} \) is an indicator variable that takes value 1 if SNP i is nonsynonymous and 0 otherwise. To get error bars on this fraction, I performed a block jackknife. I split the genome into 20 blocks with equal numbers of SNPs. If \( f_{\text{NS},j}^i \) is the estimate of the fraction of casual SNPs that are nonsynonymous (excluding block j), then

\[
 \text{SE} = \sqrt{\frac{19}{20} \sum_{j=1}^{20} (f_{\text{NS},j}^i - f_{\text{NS},i})^2},
\]

(Equation 21)

where \( f_{\text{NS}} = (1/20) \sum_{i=1}^{20} f_{\text{NS},i} \). Figure S3 shows the corresponding results for synonymous SNPs.

Interaction Effects in Annotation Models
As noted in the Results, there were two cases in which the sign of the annotation effect flipped between the single annotation models and the combined models. These were Crohn disease (Table S6) and red blood cell count (Table S18). For red blood cell count, note that SNPs influencing this trait are enriched in the annotation of DNase-I hypersensitive sites in the fetal renal pelvis when this annotation is considered alone (log2 enrichment of 2.48, 95% CI [0.04, 4.17]). This annotation is correlated with the fetal stomach annotation, which has a log2 enrichment of 4.83 (95% CI [3.30, 6.45]) when treated alone. The SNPs in both of these annotations have a log2 enrichment of 2.41 (95% CI [−1.83, 4.23]), which leads to the interaction effect. Essentially, the signal in the fetal stomach is driven by those SNPs that fall in DNase-I hypersensitive sites in the fetal stomach, but not the fetal renal pelvis. This suggests that there are a subset of DNase-I hypersensitive sites that are of particular interest for this phenotype. The interpretation of the Crohn disease example is similar.

Calibrating a “Significance” Threshold
For each genomic region, the method estimates the posterior probability that the region contains a SNP associated with a trait. If the model were a perfect description of reality, this probability could be interpreted literally. However, because the model is not perfect, I sought a more empirical calibration. I used the fact that I initially ran the method on the GWAS data reported by Teslovich et al.27 on four lipid traits. Since then, a GWAS with more individuals (although at a considerably smaller number of SNPs) has been reported for these four traits.17 This latter study contained many of the individuals from the former (which had approximately 90,000 individuals), as well as about 80,000 more individuals. However, the additional individuals were genotyped in the Metabochip,28 which has fewer than 200,000 markers, rather than the more dense standard GWAS arrays. This means that some regions of the genome do not benefit from the larger sample size.

For each region of the genome for each of the four traits, I built a table containing the minimum p value from Teslovich et al.,27 the PPA in the region (computed with the data from Teslovich et al.25), the minimum p value from the 2013 paper,17 and the sample size used for calculating this minimum p value (from the 2013 paper17). I discarded regions where sample size at the SNP with the minimum p value in the replication data set was smaller than 120,000 (given that in these regions, there are essentially no new data). I then coded each region as a “true positive” if the minimum p value from the 2013 paper17 was less than 5 × 10−8 and a “true negative” otherwise. In Figure S15, I plot the number of true positives and false positives that exceeded various p value and PPA thresholds. Note that because the data in the 2013 paper17 are not independent of those in Teslovich et al.27 this comparison is not appropriate for evaluating the relative performance of p values versus the PPA. The goal was simply to find a PPA threshold with similar performance in terms of reducing the number of false positives as the standard p value threshold of 5 × 10−8.

By visual inspection, I set a PPA threshold at 0.9 (Figure S15). At this threshold, the model identified 45 true positives and 0 false positives for high-density lipoprotein (HDL), 43 and 1 for low-density lipoprotein (LDL), 47 and 0 for total cholesterol (TC), and 27 and 0 for triglycerides (TGs). These are similar to the
numbers for a p value threshold of $5 \times 10^{-8}$ (Table S21). Combining the loci identified by both methods led to 48 loci for HDL (versus 43 with a p value threshold), 44 for LDL (versus 40), 51 for TC (versus 51), and 30 for TGs (versus 29). This is on average an increase of 6% in the number of loci identified. Note that this percentage is most likely a lower bound, given that the p values in the replication study are naturally highly correlated to those in the initial study because they use many of the same individuals. A proper comparison would use a completely separate, large set of individuals to determine true positives and true negatives, but such samples are not yet available.

Identification of Subthreshold, High-Confidence Associations

For each fitted model (using the parameters from Tables S3–S20 estimated with the penalized likelihood), I calculated the PPA in each genomic region. I then identified all regions that had a PPA greater than 0.9 but a minimum p value less than $5 \times 10^{-8}$. For each remaining region, I identified the “lead” SNP as the SNP with the largest posterior probability of being the causal SNP in the region. If this SNP was within 500 kb of a SNP with p < $5 \times 10^{-8}$ (this can happen because I use nonoverlapping windows, and sometimes the best SNP is at the edge of the region), I removed it. I also manually removed two regions (surrounding rs8076131 in Crohn disease and surrounding rs11535944 in HDL) where the “new” association was in LD with a previously reported SNP over 500 kb away. Table S22 shows the remaining SNPs; these regions are high-confidence associations that did not reach traditional genome-wide significance.

Results

I assembled a set of 18 GWASs with publicly available summary statistics and a large number of loci (at least around 20) associated with the trait of interest. These included studies of red blood cell traits, platelet traits, Crohn disease, body mass index (BMI), lipid levels, height, bone mineral density, and fasting glucose levels. I used ImpG36 to impute the summary statistics across the genome (Figure 1C). There were several genomic regions with strong evidence of association with HDL (PPA over 0.9) only when the model incorporating functional information was used. Figure 2 shows one such region near the gene NR0B2 (MIM 604630). The model identified the SNP rs6659176 as the most likely candidate to be the causal polymorphism in

For each trait, I set out to identify which of the 450 annotations (if any) were enriched with genetic variants influencing the trait. To do this, I developed a hierarchical model that learns the shared properties of loci influencing a trait. The full details of the model are presented in the Material and Methods but can be summarized briefly. Conceptually, I break the genome into large, nonoverlapping blocks (with an average size of 2.5 Mb). Let the prior probability that any block k contains an association be $\Pi_k$. If there is an association in block $k$, then let the prior probability that any SNP i is the causal SNP be $\pi_{ik}$. The model allows both $\Pi_k$ and $\pi_{ik}$ to depend on annotations of the region and SNP, respectively, and estimates these quantities on the basis of the patterns of enrichment across the whole genome. I tested this approach by using simulations based on real data from a GWAS of height (Material and Methods).

The methodology is best illustrated with an example. I started with an analysis of a GWAS of HDL levels. I first took each genomic annotation individually and estimated its level of enrichment with (or depletion of) loci that influence HDL (in the model, I additionally included a regional effect of gene density and a SNP-level effect of distance to the nearest TSS; see the Material and Methods for details). Figure 1A shows the top 40 annotations, ordered by how well each improved the fit of the model. Loci that influence HDL were most strongly enriched in enhancers identified in the HepG2 cell line and most strongly depleted from genomic regions repressed in that same cell line. HepG2 cells are derived from a liver cancer; the relevance of this cell line to a lipid phenotype makes intuitive sense. However, there were many other additional (correlated) genome annotations enriched with loci that influence HDL (Figure 1A). I thus built a model including multiple annotations to mitigate overfitting in this situation, I used a cross-validation approach (Material and Methods). The best-fitting model is shown in Figure 1B. It includes both enhancers and repressed chromatin identified in HepG2 cells, as well as coding exons and chromatin repressed in K562 cells. Because many of the annotations are correlated, those included in the combined model are the “best” representatives of sets of related annotations. I thus used conditional analysis to define the set of annotations represented by each member of the combined model (Material and Methods; Figure 1A).

A convenient side effect of fitting an explicit statistical model relating properties of SNPs to the probability of association is that the functional information can be used for reweighting the GWAS (Material and Methods). I used the combined model for HDL to reweight the association statistics across the genome (Figure 1C). There were several genomic regions with strong evidence of association with HDL (PPA over 0.9) only when the model incorporating functional information was used. Figure 2 shows one such region near the gene NR0B2 (MIM 604630). The model identified the SNP rs6659176 as the most likely candidate to be the causal polymorphism in...
this region. This SNP has a p value of $1.5 \times 10^{-6}$. However, this SNP falls in a coding exon (in fact, it is nonsynonymous), leading the model to conclude that this p value is in fact strong evidence of association. Indeed, larger studies of HDL have confirmed the evidence of association in this region (p value of $9.7 \times 10^{-16}$ at rs12748152, which has $r^2 = 0.85$ with rs665917617). This region, although not this particular SNP, was also identified in a scan for SNPs influencing multiple lipid phenotypes.\footnote{I applied this method to all 18 traits. I was first interested in estimating for each trait the fraction of associations that can be explained by nonsynonymous polymorphisms versus polymorphisms that do not influence protein sequences. For each trait, I fit a model including promoters (SNPs within 5 kb of a TSS) and nonsynonymous polymorphisms. For all traits, nonsynonymous polymorphisms were enriched among those that influence the trait, although this enrichment was not statistically significant for all traits (Figure 3A). In contrast, synonymous polymorphisms were generally not enriched among polymorphisms that influence traits, with a few notable exceptions (such as height and Crohn disease, see Figure S3). I then used these enrichments to estimate for each trait the fraction of associations that are driven by nonsynonymous polymorphisms (Material and Methods). This fraction varied from around 2% to around 20% and had an average of 10% (Figure 3B). I conclude that the relative importance of changes in protein sequence versus gene expression most likely varies across traits.}

I then used all 450 genome annotations to build models of enrichment for each trait. As for HDL, I first estimated enrichment levels individually for each annotation (Figures S4–S12). Clustering phenotypes according to these enrichment levels recapitulated many known relationships between traits (Figure S13). I then generated a combined model for each trait. The parameters of the combined models are shown in Figure 4 and Figure S14, and details of the exact annotations are in Tables S2–S20. In general, the models generated with this method were sparse and biologically interpretable. A few general patterns emerged from this analysis. Apart from the repeated occurrence of annotations related to protein-coding genes, marks of repressed chromatin were often significantly depleted for SNPs influencing traits. For example, SNPs influencing Crohn disease were depleted from repressed chromatin identified in a lymphoblastoid cell line (Figure 4D; log$_2$ enrichment of $-1.83$, 95% CI $[-3.06, -0.78]$), SNPs influencing height were significantly depleted from repressed chromatin in HeLa cells (Figure 4I; log$_2$ enrichment of $-1.5$, 95% CI $[-2.39,$}
and SNPs influencing red blood cell volume were significantly depleted from repressed chromatin in an erythroblast-derived cell line (Figure 4F; log$_2$ enrichment of $-3.91$, 95% CI $[-6.25, -2.38]$).

I additionally observed cell-type-specific enrichments in enhancer elements and DNase-I hypersensitive sites for SNPs that influence traits. Most of the observed enrichments can be readily interpreted in light of the known biology of the trait. For example, SNPs that influence platelet volume and platelet count were enriched in open chromatin identified in CD34$^+$ cells, known to be on the cell lineage that leads to platelets$^{47}$ (Figures 4A and 4B; log$_2$ enrichment of 1.81, 95% CI $[0.59, 2.86]$ for platelet count; log$_2$ enrichment of 3.02, 95% CI $[1.69, 4.26]$ for platelet volume), and SNPs that influence corpuscular hemoglobin concentration were enriched in open chromatin identified in K562 cells, a cell line derived from a cancer of erythroblasts (Figure S14E; log$_2$ enrichment of 2.67, 95% CI $[0.61, 4.44]$). For some traits, however, the connection between the trait and the tissues identified is not immediately obvious. For example, SNPs associated with platelet density were enriched in open chromatin identified in fetal fibroblasts from the abdomen (log$_2$ enrichment of 3.17, 95% CI $[1.66, 4.36]$). However, DNase-I hypersensitive sites identified in fetal fibroblasts from the back showed an even stronger enrichment (log$_2$ enrichment of 4.21, 95% CI $[2.99, 5.29]$), and sites in common between the two annotations were intermediate (log$_2$ enrichment of 3.74, 95% CI $[2.29, 4.89]$). This led to an interaction where in the joint model, the contribution of the DNase-I hypersensitive sites identified in fetal abdominal fibroblasts was negative. Although this is a statistical explanation for this observation, the biological explanation is not immediately clear. It seems likely that DNase-I hypersensitive sites are a heterogeneous set of different classes of elements and that different experiments are more sensitive, for either technical or biological reasons, to subsets of these elements.

Finally, I explored the potential of this model to identify additional high-confidence associations (as in Figure 2). In order to do this, one needs a threshold for “significance” in this model, ideally with similar properties as the standard p value threshold of $5 \times 10^{-8}$. To calibrate the method, I used the fact that I initially applied the method to a lipid-trait study that identified about 100 loci in a sample size of around 90,000 individuals.$^{27}$ Since then, larger studies have raised the number of loci associated with lipid
traits to 157.17 If a locus with a p value of $5 \times 10^{-8}$ in the larger study is considered a true positive and a locus that does not reach this threshold in the larger study is considered a true negative, it is possible to calibrate a threshold for a PPA by using the replication data (Supplemental Data). I found that a threshold of a regional PPA of 0.9 performed similarly to a stringent p value threshold (Figure S15). Combining the loci from both the standard p value approach and the approach presented here resulted in an approximately 5% increase in the number of identified loci while still maintaining a false-positive rate close to 0 (Figure S15 and Table S21). This is only a modest gain in power; that said, by applying this method to all 18 traits, I identified 49 loci that did not reach a standard statistical-significance threshold of $5 \times 10^{-8}$ but had a PPA over 0.9 (Table S22). On the basis of the above results for lipids, the level of evidence that these loci are true positives is approximately the same as that for loci that have $p = 5 \times 10^{-8}$ in a standard GWAS. Indeed, the majority of these loci have since been identified in larger cohorts than those used in this paper (Table S22).

Discussion

In this paper, I have developed a statistical model for identifying genomic annotations that are most relevant to the biology of a given phenotype. I have shown that this model is able to scan through hundreds of genomic annotations to identify a sparse set of biologically interpretable annotations without prior knowledge of the biology of the phenotype.

Linking GWASs to Biology

Perhaps the most striking observation is that chromatin annotated as repressed in a given cell type was often depleted of SNPs that influence traits. Given that approximately 60%–70% of the genome falls in this annotation in any given cell type (Supplemental Data), this information could dramatically limit the number of SNPs considered during fine mapping of loci identified in GWASs. Additionally, I identified several nonobvious connections between tissue and phenotypes. For example, SNPs that influence platelet volume were enriched in DNase-I hypersensitive sites in the spleen (Figure 4A). Although the spleen functions in the removal of platelets from the bloodstream, the connection between its function and platelet volume is unclear. An important next step will be to connect the identified noncoding variants in regulatory regions to changes in gene expression; this is presumably the mechanism by which they exert an effect on phenotypes. Methods for inferring the casual chain connecting variation in DNase-I sensitivity, variation in gene expression, and variation in phenotypes46,48–51 will be essential.

Modeling Assumptions

I have made several modeling assumptions that merit discussion. First, by splitting the genome into blocks on the basis of numbers of SNPs, I have made the implicit assumption that the probability that a genomic region

![Figure 3. Estimated Role of Protein-Coding Changes in Each Trait](image-url)

(A) Estimated enrichment of nonsynonymous SNPs. For each trait, I fit a model including an effect of nonsynonymous SNPs and an effect of SNPs within 5 kb of a TSS. Shown are the estimated enrichment parameters and 95% CIs for the nonsynonymous SNPs.

(B) Estimated proportion of GWAS hits driven by nonsynonymous SNPs. For each trait, using the model fit in (A), I estimated the proportion of GWAS signals driven by nonsynonymous SNPs. This estimate and its SE are shown.
contains a SNP associated with a given phenotype depends on the SNP density rather than the physical size—that is, a short genomic region with a large number of SNPs is a priori as likely to have an association as a long genomic region with few SNPs. I have also made a more restrictive assumption that there can be only a single causal SNP in a given genomic region. This assumption is a natural starting point, but as GWAS sample sizes increase even more, it will begin to be untenable. Advances in methods for joint analysis of multiple SNPs (e.g., Yang et al.52) might provide a way forward in this situation. Finally, note that the model is limited by the types of genomic annotations that are

---

**Figure 4. Combined Models for Nine Traits**

For each trait, I built a combined model of annotations by using the algorithm presented in the Material and Methods. Shown are the maximum-likelihood estimates and 95% CIs for all annotations included in each model. Note that although these are the maximum-likelihood estimates, model choice was done with a penalized likelihood (Material and Methods). For the other nine traits, see Figure S14. In parentheses next to each annotation (except for those relating to distance to TSSs) is the total number of annotations that are statistically equivalent to the included annotation in a conditional analysis (Material and Methods). The annotation of DNase-I hypersensitive sites in fetal fibroblasts from the abdomen (marked by an asterisk) had a positive effect when treated alone; see the main text for discussion.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Gene-rich regions</th>
<th>Gene-poor regions</th>
<th>DNAse (CD34+ cells)</th>
<th>Enhancer (HeLa)</th>
<th>Transcribed (LCL)</th>
<th>DNAse (Fetal spleen)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platelet volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platelet count</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bone density (lumbar spine)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crohn disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasting glucose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean red blood cell volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For each trait, I built a combined model of annotations by using the algorithm presented in the Material and Methods. Shown are the maximum-likelihood estimates and 95% CIs for all annotations included in each model. Note that although these are the maximum-likelihood estimates, model choice was done with a penalized likelihood (Material and Methods). For the other nine traits, see Figure S14. In parentheses next to each annotation (except for those relating to distance to TSSs) is the total number of annotations that are statistically equivalent to the included annotation in a conditional analysis (Material and Methods). The annotation of DNase-I hypersensitive sites in fetal fibroblasts from the abdomen (marked by an asterisk) had a positive effect when treated alone; see the main text for discussion.
available, and the best annotations identified in the model might be “proxies” for the truly relevant annotations. For example, SNPs associated with height were enriched in DNase-I hypersensitive sites identified in the fetal lung (Figure 4D); taken literally, this would suggest that some SNPs influence height through lung development. An alternative possibility, however, is that patterns of open chromatin in the lung (which is of course a heterogeneous tissue) are useful proxies for patterns of open chromatin in a cell type that has not been profiled; this hypothetical cell type could in principle be present in any tissue.

Prospects for Fine Mapping GWAS Loci with Functional Genomic Data
I have primarily focused on using the model to identify annotations relevant to a trait of interest, although I have also explored using this information to identify high-confidence associations that do not reach genome-wide significance. A third natural application, which I have not explored, is the possibility to fine map GWAS loci by using functional genomic information. Indeed, the posterior probability that each SNP in a given genomic region is the causal one is explicitly included in the model. However, in current applications, around 20% of common SNPs are neither genotyped nor successfully imputed; this is a major limitation to fine mapping and cannot be overcome with statistical means. As GWASs move to even denser genotyping or sequencing, I expect that revisiting this issue will be fruitful.

Appendix A
In this appendix, I include a detailed description of the data used in the paper.

Genetic Investigation of Anthropometric Traits Consortium Data
I downloaded summary statistics from large GWASs of height46 and BMI41 (Web Resources). The height summary statistics consisted of 2,469,635 SNPs either directly genotyped or imputed in an average of 129,945 individuals. I removed all SNPs with a sample size of fewer than 120,000 individuals. The BMI summary statistics consisted of 2,471,516 summary statistics either directly genotyped or imputed in an average of 120,569 individuals. I removed all SNPs with a sample size of fewer than 110,000 individuals. I then imputed summary statistics at SNPs identified in the 1000 Genomes Project as described in the Material and Methods.

Genetic Factors for Osteoporosis Consortium Data
I downloaded summary statistics from large GWASs of bone mineral density42 (Web Resources). There were two traits in these data: bone density measured in the femoral neck and bone density measured in the lumbar spine. The GWAS on femoral neck bone density consisted of 2,478,337 SNPs, and the GWAS on lumbar spine bone density consisted of 2,468,080 SNPs. Because the sample size at each SNP was not reported, I used the overall study sample sizes of 32,961 (for femoral neck bone density) and 31,800 (lumbar spine bone density) as approximations of the sample size at each SNP and imputed summary statistics as described in the Material and Methods.

International Inflammatory Bowel Disease Genetics Consortium Data
I downloaded summary statistics from a large GWAS of Crohn disease40 (Web Resources). The downloaded data consisted of 953,242 SNPs. Because the sample size at each SNP was not reported, I used the overall study sample sizes of 6,299 case and 15,148 control individuals as approximations of the sample size at each SNP and imputed summary statistics as described in the Material and Methods. Note that summary statistics from a GWAS of ulcerative colitis were also available from this site; however, these data contained a number of false-positive associations that were filtered by Jostins et al.40 according to criteria that were not available to me. I thus only used the Crohn disease association study.

Meta-analyses of Glucose and Insulin-Related Traits Consortium Data
I downloaded summary statistics from a large GWAS of fasting glucose levels43 (Web Resources). The downloaded data consisted of 2,628,880 SNPs. Because the sample size at each SNP was not reported, I used the overall study sample size of 58,074 as an approximation of the sample size at each SNP and imputed summary statistics as described in the Material and Methods.

Global Lipids Genetics Consortium Data
I downloaded summary statistics from a large GWAS of lipid traits27 (Web Resources). These data consisted of summary statistics for association studies of four traits: LDL cholesterol, HDL cholesterol, TGs, and TC. The HDL data consisted of 2,692,429 SNPs genotyped or imputed in an average of 88,754 individuals, the LDL data consisted of 2,692,564 SNPs genotyped or imputed in an average of 84,685 individuals, the TC data consisted of 2,692,413 SNPs genotyped or imputed in an average of 89,005 individuals, and the TG data consisted of 2,692,560 SNPs genotyped or imputed in an average of 85,691 individuals. For all traits, I removed SNPs with a sample size of fewer than 80,000 individuals and imputed summary statistics as described in the Material and Methods.

To calibrate significance thresholds, I additionally used summary statistics from the 2013 Global Lipids Genetics Consortium paper17 (Web Resources).

Red Blood Cell Trait Data
I obtained summary statistics from a large GWAS of red blood cell traits15 from the European Genome-phenome Archive (accession number EGAS00000000132). I
downloaded summary statistics from association studies of six traits: hemoglobin levels, mean cell hemoglobin (MCH) levels, mean corpuscular hemoglobin concentration (MCHC), mean cell volume (MCV), packed cell volume (PCV), and red blood cell count (RBC). The hemoglobin-level data consisted of 2,593,078 SNPs genotyped or imputed in 50,709 individuals, the MCH data consisted of 2,586,785 SNPs genotyped or imputed in an average of 43,127 individuals, the MCHC data consisted of 2,588,875 SNPs genotyped or imputed in an average of 46,469 individuals, the MCV data consisted of 2,591,132 SNPs genotyped or imputed in an average of 47,965 individuals, the PCV data consisted of 2,591,079 SNPs genotyped or imputed in an average of 44,485 individuals, and the RBC data consisted of 2,589,454 SNPs genotyped or imputed in an average of 44,851 individuals. I removed all SNPs with a sample size of fewer than 50,000 individuals (for hemoglobin levels) or 40,000 individuals (for the other traits) and imputed summary statistics as described in the Material and Methods.

Platelet Traits
Summary statistics from a large GWAS of platelet traits were generously provided by Nicole Soranzo. The data consisted of summary statistics from association studies of two traits: platelet counts and mean platelet volume. The platelet-count data consisted of 2,705,636 SNPs genotyped or imputed in an average of 44,217 individuals, and the platelet-volume data consisted of 2,690,858 SNPs genotyped or imputed in an average of 16,745 individuals. I removed all SNPs with a sample size of fewer than 40,000 individuals (for platelet counts) or 15,000 individuals (for platelet volume) and imputed summary statistics as described in the Material and Methods.

DNase-I-Hypersensitivity Data
I downloaded DNase-I-hypersensitivity data from two sources. The first was a set of regions defined as DNase-I hypersensitive by Maurano et al. in 349 samples. I downloaded .bed files for 349 samples (Web Resources) on February 13, 2013. These samples include 116 samples from cell lines or sorted blood cells and 333 samples from primary fetal tissues. These latter samples were sampled from several tissues at various time points; I treated each track as independent rather than pooling data from tissues, given that different experiments might have slightly different properties. The tissues in this latter group were fetal heart, fetal brain, fetal lung, fetal kidney, fetal intestine (large and small), fetal muscle, fetal placenta, and fetal skin.

The second was a set of regions defined as DNase-I hypersensitive by the Crawford lab in the context of the ENCODE project. I downloaded .bed files for 53 samples (Web Resources) on March 29, 2013. I restricted myself to the files labeled as being generated at Duke University. Each experiment defined a set of regions of open chromatin in a particular cell type or cell line.

The “Duke” DNase-I hypersensitive sites were all of exactly 150 bases in length, and each annotation covered approximately 1% of the genome (range = 0.4%–1.9% of the genome). The “Maurano” DNase-I hypersensitive sites were on average 514 bases long, and each covered on average 2.7% of the genome (range = 0.9%–5.1% of the genome).

Chromatin State Data
I downloaded the “genome segmentations” of the six ENCODE cell lines (Web Resources) on December 18, 2012. I used the “combined” segmentation from two algorithms. This segmentation split the genome into nonoverlapping regions described as CTCF binding sites, enhancers, promoter-flanking regions, repressed chromatin, transcribed regions, TSSs, and weak enhancers. This segmentation was done independently in each of six cell lines for a total of 42 annotations.

Overall, the “repressed chromatin” mark covered the largest fraction of the genome on average 66% (ranging from 60% for HUVEC cells to 70% for H1 ES cells). The “transcribed” mark covered on average 13% of the genome, the “CTCF” mark 1% of the genome, the “enhancer” mark 0.9% of the genome, the “TSS” mark 0.7% of the genome, the “weak enhancer” mark 0.4% of the genome, and the “promoter-flanking” mark 0.2% of the genome. The remainder of the genome was not mappable by short reads, and it was thus excluded from these annotations.

Gene Models
I downloaded the Ensembl gene annotations from the UCSC Genome Browser on May 21, 2013. Annotations of nonsynonymous and synonymous status for all SNPs in phase 1 of the 1000 Genomes Project were obtained (Web Resources). Coding exons covered about 3% of the genome, whereas 3’ and 5’ UTRs covered 2% and 0.6% of the genome, respectively.

Supplemental Data
Supplemental Data include 16 figures and 22 tables and can be found with this article online at http://www.cell.com/ajhg.

Acknowledgments
I thank David Reich, Nick Patterson, Alkes Price, and Po-Ru Loh for helpful discussions and suggestions and Jonathan Pritchard, Graham Coop, Graham McVicker, Anil Raj, and two anonymous reviewers for comments on a previous version of this manuscript. I also thank Nicole Soranzo for providing access to the platelet studies and Luke Jostins for assistance in obtaining the Crohn disease data. This work was supported by NIH postdoctoral fellowship GM103098 to J.K.P.

Received: November 26, 2013
Accepted: March 11, 2014
Published: April 3, 2014
Web Resources

The URLs for data presented herein are as follows:
1000 Genomes Project annotated VCFs, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/analysis_results/functional_annotaton/annotated_vcf/
DNase-I hypersensitive sites (Duke), http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/iddPeaks/
UCSC Genome Browser, http://genome.ucsc.edu

References


