Functional Annotation Signatures of Disease Susceptibility Loci Improve SNP Association Analysis

Edwin S. Iversen, Jr.,[*] Gary Lipton,[†] Merlise A. Clyde,[‡] and Alvaro N. A. Monteiro[§]

November 6, 2013

Corresponding Author:
Edwin S Iversen, Jr.
Department of Statistical Science
Duke University
Durham, NC 27708–0251
USA
iversen@stat.duke.edu
(919) 681–8442 (Office)
(919) 684–9524 (FAX)

Running Title: Functional Signatures Improve SNP Association Analysis

Key Words: Association Study; GWAS; SNPs; Functional Annotations; Bayesian Analysis; ENCODE Project.

[*] Department of Statistical Science, Duke University, Durham, NC 27708.

[†] Department of Statistical Science, Duke University, Durham, NC 27708.

[‡] Department of Statistical Science, Duke University, Durham, NC 27708.

[§] Cancer Epidemiology Program, H. Lee Moffitt Cancer Center & Research Institute, 12902 Magnolia Drive, Tampa, FL 33612.

Abstract

We describe the development and application of a Bayesian statistical model for the prior probability of phenotype–genotype association that incorporates data from past association studies and publicly available functional annotation data regarding the susceptibility variants under study. The model takes the form of a binary regression of association status on a set of annotation variables whose coefficients were estimated through an analysis of associated SNPs housed in the GWAS Catalog (GC). The set of functional predictors we examined includes measures that have been demonstrated to correlate with the association status of SNPs in the GC and some whose utility in this regard is speculative: summaries of the UCSC Human Genome Browser ENCODE super–track data, dbSNP function class, sequence conservation summaries, proximity to genomic variants included in the Database of Genomic Variants (DGV) and known regulatory elements included in the Open Regulatory Annotation database (ORegAnno), PolyPhen–2 probabilities and RegulomeDB categories. Because we expected that only a fraction of the annotation variables would contribute to predicting association, we employed a penalized likelihood method to reduce the impact of non–informative predictors and evaluated the model's ability to predict GC SNPs not used to construct the model. We show that the functional data alone are predictive of a SNP's presence in the GC. Further, using data from a genome–wide study of ovarian cancer, we demonstrate that their use as prior data when testing for association is practical at the genome–wide scale and improves power to detect associations.

1 Introduction

The purpose of genetic association studies is to discover genetic loci that contribute to an inherited trait, identify the variants behind these associations and ascertain their functional role in determining the phenotype (Manolio, 2010). Modern association studies bring to bear on this problem high coverage genotype data, comprehensive databases of genetic variation that allow imputation of most common ungenotyped variants to high accuracy and extensive, publicly available, *in silico* resources housing a growing assortment of genomic data that allow functional characterization of vast regions of the human genome. In the typical genome–wide association study (GWAS), the first two forms of data are combined to reconstruct genotypes to a desired density and these genotypes are then systematically tested for association with the phenotype. The functional annotation data are most frequently used in *post hoc* interpretation of evident associations raised by the analysis (Freedman et al., 2011).

To date, functional annotation data have rarely played more than an indirect role in assessing evidence for association. For example, they may be used to suggest candidate genes and SNPs for study or to support links between candidate SNPs and genes. While methods to incorporate functional annotation data *a priori* in genetic association analyses exist, they are infrequently used. The prevailing approach to this is via a two–staged hierarchical model in which coefficients in the stage I generalized linear model for phenotype given genotype and exposure measurements are regressed, in stage II, on the annotation data (Witte et al., 1994; Aragaki et al., 1997; Hung et al., 2004, 2007). This is limited to analysis of a modest number of variants and does not make use of prior data derived from previous association studies to inform the nature of that relationship.

It is becoming increasingly clear that a widening array of annotation data correlates with a variant's having been associated with a human phenotype (Hindorff et al., 2009; Nicolae et al., 2010; The ENCODE Project Consortium, 2012; Schaub et al., 2012). In what follows, we describe a formal approach to inference for association that combines functional annotation data (through a prior distribution) with genotype data (through a sampling model for the phenotype given genetic and other covariate data). We construct the prior distribution through careful analysis of SNPs housed in the GWAS Catalog (Hindorff et al., 2009). We refer to the linear combination of the annotation variables defined by this model and evaluated for a given SNP as its 'functional annotation signature.' We show that functional signatures so derived are predictive of the association status of SNPs not used in their creation and that, when coupled with genetic association data following the method we describe, improve the efficiency of association testing in a GWAS study of ovarian cancer.

2 Results

The ultimate goal of association studies is to identify the set of common polymorphisms that influence a phenotype. This goal is approached through a statistical analysis designed to measure the evidence in favor of association followed by a decision rule used to declare each variant's true status as 'associated,' 'uncertain,' or 'unassociated.' The data that inform these analyses usually comprise phenotype labels, SNP genotype data and a set of non–genetic covariates in addition to functional annotations of the variants under study. The statistical analysis may take many forms, varying according to choice of modeling approach and inferential paradigm (Frequentist or Bayesian). The approach we develop here relies on Bayesian inference but can also be applied when the genetic association summaries are p–values. In this paradigm, prior data on a quantity of interest (such as the binary association status of a genetic variant) are updated to reflect evidence in the current data set.

A Bayesian analysis of genetic association data returns an estimate of the odds of association of each marker given the available data. When the data take two distinct forms — here subject–level phenotype, genotype and covariate data and variant–level functional annotations — the odds of association may be calculated in two stages, either by incorporating functional data prior to or following evaluation of the genetic data. The latter represents the heuristic typically followed in practice, whereby functional data is evaluated in an informal way (from the probabilistic point of view) conditional on evidence for association. Here we describe a model–based framework for combining functional and association data following the second factorization. We focus on the case–control study design for purposes of illustrating integration of the *a priori* (to association data) models for functional annotation data we describe below into analyses of genetic association data. Details of the models and their assumptions are provided in Methods.

When the functional data are incorporated as prior information, the odds of a SNP's association given the functional and subject–level data can be written as the product of the Bayes factor (BF) in favor of association and the prior odds of association given the functional data. The BF is the ratio of the integrated likelihood of the phenotype data given the covariate and genotype data assuming the SNP is associated to the integrated likelihood of the phenotype data given the covariate data only (i.e. assuming the SNP is not associated). It is a commonly used Bayesian statistical measure of association and is calculated by the SNPTEST (Marchini et al., 2007) and BIMBAM (Servin and Stephens, 2007) packages for analysis of GWAS data. Alternately, Sellke et al. (2001) show that an upper bound on the Bayes factor in favor of association is approximately equal to $-1/(ep\log_e(p))$ when $p < (1/e)$ and 1.0 otherwise, where p is the p–value for association. This allows the method to be used in conjunction with standard frequentist association testing software.

In short, the functional annotation data are incorporated into an analysis by formally updating the prior odds of association given the annotation data by a standard measure of genetic association. This process is depicted schematically in Figure 1. In what follows, we describe the model used to calculate prior odds of association and demonstrate its use in a GWAS of ovarian cancer. In it, the log of a SNP's prior odds of association, its 'functional signature,' is a linear combination of the functional data.

2.1 Functional Signatures of Known Associations

We constructed the functional annotation signatures by estimating the multivariate relationship between a set of functional annotation variables and the binary association status of a set of SNPs. Figure 2 provides a schematic of our approach. In brief, we identified a set of associated SNPs and, for each, we chose a matching, unassociated 'control' SNP. We divided the matched pairs into 'training' and 'validation' sets and used the former to construct a series of models to predict association status given the function data and used the latter to compare the performance of these models. We chose the model that demonstrated the best predictive accuracy in the validation data, as measured by concordance index, to define the functional annotation signatures.

We began by constructing a matched case–control study *of SNPs* in which the cases were drawn from the GWAS Catalog (Hindorff et al., 2009) and the controls were identified from the HapMap database, Release 27, Phases II and III merged genotypes. We identified 2,093 case SNPs and, for each, identified one control SNP matched on chromosome, minor allele frequency and the genotyping platform(s) it appeared on. Since SNPs in the GWAS Catalog are arguably more frequently tags than the directly associated variant, we followed Hindorff et al. (2009) and identified 'LD partners' for each case and control SNP. We grouped each case and control SNP together with its LD partners to form blocks.

Using on–line bioinformatics resources, we assembled a set of functional annotation variables representing a variety of contextual descriptions and empirical measurements with which we annotated each of the 48,889 case, control and LD partner SNPs. We included annotation variables shown to be correlated with presence in the GWAS Catalog or that we believed likely to be so. These were: dbSNP function designation; summaries of ENCODE Project (The ENCODE Project Consortium, 2007, 2011) data on transcription levels assayed by RNA–seq (Mortazavi et al., 2008; Langmead et al., 2009), measures of signal enrichment for H3K4Me1, H3K27Ac and H3K4Me3 histone modifications associated with enhancer and promoter activity (Bernstein et al., 2006; Mikkelsen et al., 2007), evidence for overlap with a DNaseI hypersensitivity cluster (Sabo et al., 2006, 2004) and evidence for transcription factor binding (Euskirchen et al., 2004, 2007; Martone et al., 2003; Robertson et al., 2007; Rozowsky et al., 2009); PhyloP evolutionary conservation scores (Siepel et al., 2006); indicators for whether or not the variant falls in a region of known copy number variation, a region containing insertions or deletions or a region with inversions (Iafrate et al., 2004; Zhang et al., 2006); PolyPhen–2 (Adzhubei et al., 2010) probability that a mutation is damaging; and RegulomeDB score (Boyle et al., 2012). The latter represents a synthesis of regulatory data derived from ENCODE and other sources. While not a comprehensive set, they covered the major annotation classes available at the time of analysis and are readily available to individuals executing an association study. The infrastructure and methods described here are easily updated to accommodate new variables as they become generally available. Table 1 lists the 57 variables that we used to construct the functional signatures of association.

The 48,889 SNPs included in the analysis were grouped into 2,093 case and an equal number of control blocks. We randomly selected 1,675 of these matched case–control pairs for development of the model (the 'training set') and left the remaining 418 pairs for model evaluation and comparison (the 'evaluation set'). We modeled the probability that a SNP is associated given its functional data using a logistic regression model. Further, we assumed that each case block contained one or more associated SNPs and that each control block contained none.

While the assembled list of functional predictors includes measures that have been demonstrated to correlate with the association status of SNPs in the GC, it also includes a number of measures whose utility in this regard was unclear. Hence, we expected that only a fraction of the 57 variables would contribute to predicting phenotype association. We used shrinkage priors (Hans, 2009; Richardson et al., 2011) to reflect this belief and chose the normal–exponential–gamma (NEG) distribution for its ability to penalize heavily weakly determined predictors and to penalize weakly those that are well determined (Griffin and Brown, 2007; Hoggart et al., 2008; Griffin and Brown, 2010). Further details of the model and the Markov chain Monte Carlo (MCMC) algorithm used for inference can be found in Methods.

Table 2 provides a summary of the coefficient estimates obtained for the binary regression of association status on the 57 functional annotation variables. Because all variables in the model were standardized, coefficients measure the difference in the log–odds of phenotype association attributed to an increase of one standard deviation in the covariate when the others remain fixed. The majority of predictive variation (51%) in the functional scores as measured in the control block SNPs from the validation set, is due to the Broad promoter/enhancer ChIP–seq principal components (PCs) and nearly all ($> 97\%$) of this variation is due to PCs 1, 2, 4, 5, 6, 8 and 13. Each PC is a linear combination of the 75 summary statistics of the 25 assays. Supplemental Figure 1 depicts the loadings (weights in the linear combinations) for these PCs as they depend on histone modification, cell line and summary statistic. Grossly, PC 1 measures total signal strength across all cell lines and histone modifications, PC 2 contrasts average signal strength of the H3k4me3 assay with variation over all assays, while the remaining PCs each contrast signal in one subset of cell lines with that in another (PC 4: HMEC and NHEK *versus* GM12878 and K562; PC 5:

GM12878, HMEC and NHEK *versus* HSMM, HUVEC and NHLF; PC 6: H1-hESC, HepG2 and HSMM *versus* GM12878 and HUVEC; PC 8: K562 *versus* H1-hESC; and PC 13: HepG2 *versus* H1-hESC).

The sequence conservation PCs collectively make the next largest contribution, explaining 16% of variation in the functional scores; PCs 1 and 3 explain > 97% of this. Each PC is a linear combination of the summary statistics of the 28 and 44 species PhyloP scores, each for all species and restricted to placental mammals. Supplemental Figure 2 graphs the loadings for these PCs as they depend on number of species, depth of alignment and summary statistic. Briefly, PC 1 measures total signal strength across scores with the scores based on the 28–way alignment weighted more heavily than those based on the 44–way alignment, while PC 3 contrasts the 28–way with 44-way scores.

The CalTech RNA–seq PCs collectively explain 10% of the signature, with PCs 1, 2, 4 and 8 contributing 87% of this. Supplemental Figure 3 depicts the loadings for these PCs as they depend on cell line and summary statistic. PC 1 provides a measure of total signal strength across all cell lines, while the remaining PCs each contrast signal in one subset of cell lines with that in another (PC 2: H1-hESC and K562 *versus* GM12878 and NHEK; PC 4: GM12878 and H1-hESC *versus* K562, NHEK and HepG2; PC8: HUVEC *versus* NHEK).

RegulomeDB score explains the next largest fraction (8%) of variation. It is represented by six variables, each indicating a functional category; category 7 serves as the reference ('baseline'). Categories 2, 4, 5 and 6 explain 99% of this variation suggesting that other annotation variables in the model better characterize the probability of phenotype association for variants in categories 1 and 3. Virtually all (96%) of the 2.5% contribution to variation made by the DGV variables is due to the copy number and inversion variables. Finally, the dbSNP functional class variables are the only remaining that contribute more (=1.7%) than 1% of the variation in functional scores. Virtually all (99%) of this contribution is due to the non–synonymous designation within which the PolyPhen–2 probability contributes significant resolution to the model.

We estimated the concordance indexes (equivalent to AUC, area under the ROC curve) for each model using the 418 matched case–control block pairs in the validation set as a tool for comparing the accuracy of their out–of–sample predictions. Table 3 provides the estimates of concordance and associated 95% interval estimates. While the concordance statistics are not discernibly different from one another, the best out–of–sample predictive ability is achieved using the model with the prior distribution having the strongest shrinkage properties, i.e. the 'NEG3' model.

2.2 Application to an Ovarian Cancer Multi–GWAS Study

Here we compare the ranks assigned to a group of variants in a GWAS analysis when those ranks are calculated with and without the functional annotation data. Each in the group of variants is assumed to have known association status (associated/unassociated) with epithelial ovarian cancer, where this determination is based on confirmatory studies subsequent to the GWAS. The group is constructed as follows. There are currently 11 published, genome–wide significant loci for epithelial ovarian cancer. Nine of the 11 have come to light through analysis of genome–wide SNP data. These are rs3814113 (Song et al., 2009), rs8170 (Bolton et al., 2010), rs2072590, rs2665390, rs7814937, rs9303542 (Goode et al., 2010), rs11782652, rs7084454, rs757210 (Pharoah et al., 2013). The remaining two (rs10069690 and rs2077606) were identified by candidate gene/pathway investigations (Bojesen et al., 2013; Permuth-Wey et al., 2013); all 11 have been evaluated in very large confirmatory studies. We consider these to be 'true positive' variants. Our analysis of data from the large–scale follow–up study of GWAS candidates described in Pharoah et al. (2013) allowed us to identify a group of variants with strong evidence *against* association that we treat here as 'true negatives.'

Table 4 summarizes the GWAS results for the true positive and true negative SNPs when the analysis is conducted with (subscript 'A+F') and without (subscript 'A') the functional signatures and where the association summaries (Bayes factors) are calculated directly (columns labeled 'Bayesian Analysis') and approximated from the results of standard likelihood ratio tests using the method of Sellke et al. (2001) (columns labeled 'P–Value Approximation'). We focus here on results of the Bayesian analysis, noting that the approximate method yields very similar results. Note that the candidate SNPs are ranked substantially lower than the GWAS 'hits.' Indeed, the evidence in the association data related to these variants is actually *against* association (both of their Bayes factors are less than 1.0). The GWAS hits are all ranked in the top 50,000 (of approximately 2.5 million) by the same measure and all have Bayes factors of at least 3 to 1 in favor of association.

Only two of the truly associated SNPs (rs11782652 and rs9303542) are ranked higher when the functional data are ignored than when they are used, however their respective changes in rank are small. The median (alt. average) rank of the truly associated SNPs was 5,272 (178,246) without and 3,532 (80,143) with the functional data included. If design constraints allowed only for followup of the top 5,000 variants, a larger fraction (7/11) would be discovered with addition of the functional data than without (5/11); with followup of 10,000 variants, these fractions become 8/11 and 7/11. In contrast, when the function data were included the median (alt. average) rank among a set of 'true negative' SNPs increased from 181,116 (438,664) to 244,393 (517,810), while the number selected for followup fell from 244 to 204 under the 5K scenario and from 443 to 373 under the 10K scenario.

Functional signatures of tag SNPs correlate with function of tagged SNPs. While a few of the functional variables, such as the function class designation 'nonsynonymous,' incorporated in the signature are base pair specific, most map to contiguous regions of 100's or 1000's of base pairs. Hence, the functional signatures associated with nearby SNPs are correlated. Figure 3 is a plot of the correlation between the functional signatures of adjacent SNPs that passed QC in the ovarian cancer GWAS described above as a function of the distance, measured in base pairs (BPs), between the two variants. This correlation is greater than 0.72 (alt 0.68) for more than 80% (alt 97.5%) of adjacent variants, corresponding to those at distances of 1470 (alt 4376) BPs or less. Hence, while there are gains to be realized in doing so, it is not necessary to impute to and annotate at the highest possible density to realize an increase in power to detect association through the use of functional signatures, a fact we demonstrated empirically above. Note that typical BP distances between tagged (not genotyped or imputed) variants and their nearest tag will be on the order of one half of the distances reported here for adjacent tags.

3 Discussion

Using the GWAS Catalog as a sampling frame, we developed a model for the probability that a given polymorphism is associated with an observable human phenotype given a set of functional annotation variables and demonstrated that this model has the ability to predict a set of phenotype associated variants not used in the model building exercise. We demonstrate several methods for incorporating functional annotation signatures defined by this model and evaluated for a SNP's annotation data as prior data and show through example that by doing so we improve the efficiency of GWAS scale analysis to identify true positive associations for follow–up study.

The approach we describe is computationally tractable and scalable to modern genome–wide analysis. Our use of penalized regression techniques to model the functional data and construct the function signatures allows us to consider a relatively large number of individual annotation variables while controlling for over–fitting. We evaluated sensitivity of the model's out–of–sample predictions to choice of shrinkage prior and found that the most aggressive choice we examined, the model whose results are summarized herein, resulted in the best out–of–sample concordance estimates. Our approach can be expanded and adapted to incorporate more detailed annotation data such as was recently released by the ENCODE consortium (The ENCODE Project Consortium, 2012) or generated experimentally in individual labs.

In principle, estimates of the parameters in the model for SNP association status given the functional data can be refined via Bayesian updating as part of an association analysis. This requires an additional layer of analysis that is feasible, but computationally demanding to implement on a genome–wide scale. However, the value of this will be limited in settings where there are few truly associated SNPs and/or the case–control data supporting associations are weak, i.e. the vast majority of applications. Here, Bayesian updating will yield estimates equivalent to those using the approach we describe above up to Monte Carlo simulation error. Indeed, we formally compared the two approaches using the ovarian cancer GWAS data and found little change in the median ranks of the true positive (3,532 *versus* 3,705) and true negative SNPs (244,393 *versus* 248,459). This suggests that the added value of Bayesian updating to the functional signatures will typically be limited.

Performance for our integrative approach likely depends on the depth, specificity and density of coverage of the available annotation data. The current study defines a starting point and benchmark in each of these dimensions. In particular, while the depth of annotation considered here is sufficient to noticeably improve inference for association, it is clear from recent ENCODE Project Consortium publications that it reflects only a small fraction of the complexity present in the regulatory landscape. Further, none of the annotation variables are tailored to the outcome phenotype; indeed, the ENCODE super track data enter the model through linear combinations of the cell–line specific measurements, effectively averaging over cell type. Many regulatory processes are cell–type–specific (The ENCODE Project Consortium, 2012; Schaub et al., 2012) and hence will be more informative for a given phenotype if measured in the appropriate cell type. However, determining the relevant annotation data, assuming it exists, for a given phenotype requires domain expertise and more careful modeling to create functional signatures. While Bayesian updating did not improve inferences in the ovarian cancer GWAS example, a generalization of it that couples the existing signature structure with context–specific annotations such as cell type specific eQTL data and an independent prior distribution on its multivariate adjusted effect is one approach to improving specificity.

Finally, our analyses have been carried out entirely at the HapMap III density. Our approach succeeds at this density because the functional signatures of SNPs nearby, at distances typical of HapMap III, are highly correlated and hence the functional signatures of HapMap III polymorphisms essentially tag function of nearby polymorphisms not in the database. As coverage (genotype/imputation density) of the typical association study becomes more complete, the need to rely on correlations between functional signatures will diminish and their power to assist in identifying and localizing associations is expected to increase. Association analyses at the density of the 1000 Genomes Project database (The 1000 Genomes Project Consortium, 2012) are now possible and will likely become common. The specificity of the functional signatures should improve when reconstructed and applied at this density as we plan to do as we continue to develop this approach.

4 Methods

4.1 Association Analysis Given Annotation Data

Let $\mathbf{G}$ be an n by p matrix of SNP genotypes, D be an n by 1 vector of disease indicators where $D_i = 1$ if individual i has the disease and $D_i = 0$ otherwise, $\mathbf{X}$ be an n by r matrix of covariates used in the association model and $\mathbf{F}$ be a p by m matrix of SNP–level functional annotation data where n is the number of individuals, p is the number of SNPs, r is the number of covariates and m is the number of annotation variables. Finally, let A be a p by 1 vector of 0-1 indicators of the (unknown) association status of the variants where $A_s = 1$ if SNP s is associated with the phenotype of interest.

In what follows, we specify the likelihood for the association indicator given the association ($\mathbf{X}$, D, G) and function ($\mathbf{F}$) data. To this end, we let $\Pr(A \mid D, \mathbf{X}, \mathbf{G}, \mathbf{F}) \propto \prod_{s=1}^{p} \Pr(A_s \mid D, \mathbf{X}, G_s, F_s)$. This relies on two assumptions: (1) that the A_s's are conditionally independent given ($\mathbf{X}$, D, $\mathbf{G}$, $\mathbf{F}$) and (2) that the A_s's are conditionally independent of other variants ($\mathbf{G}_{-s}$, $\mathbf{F}_{-s}$) given ($\mathbf{X}$, D, G_s, F_s). The notation $\mathbf{G}_{-s}$ indicates the matrix obtained by removing column s from $\mathbf{G}$.

Further, we assume that the disease phenotype data are conditionally independent of the functional data for SNP s given the association status of that SNP, the covariate data and the genotype data for that SNP and that the association status indicator for SNP s is conditionally independent of the covariate data and its genotype data given its functional data. The latter assumption may be violated, for example, if the genotype data G_s carries information about function (e.g. minor allele frequency) not included in $\mathbf{F}$. Given this, the odds of association of SNP s given its association and functional data can be written as the product of the (prior) odds of its association given its functional data times the (integrated) likelihood ratio or Bayes Factor (BF) of the phenotype given the SNP genotype and other covariate data, i.e.

$$\begin{aligned}
\text{odds}(A_s = 1 \mid D, \mathbf{X}, G_s, F_s) &= \frac{\Pr(A_s \,=\, 1 \mid D, \mathbf{X}, G_s, F_s)}{\Pr(A_s \,=\, 0 \mid D, \mathbf{X}, G_s, F_s)} \\
&= \frac{\Pr(D \mid A_s \,=\, 1, \mathbf{X}, G_s) \Pr(A_s \,=\, 1 \mid F_s)}{\Pr(D \mid A_s \,=\, 0, \mathbf{X}, G_s) \Pr(A_s \,=\, 0 \mid F_s)} \\
&= \mathbf{BF}_s \times \text{odds}(A_s = 1 \mid F_s),
\end{aligned}$$

We describe estimation of the association summary Bayes factor below.

Given the binary, logistic link model developed below for association status given the functional data and the parameters α and β, $\text{odds}(A_s \,|\, F_s) \;=\; \exp{(\alpha \,+\, F_s\beta)}$ and hence, given α and β

$$\Pr(A_s \;=\; 1 \,|\, D, \mathbf{X}, G_s, F_s, \alpha, \beta) \;=\; \frac{\mathbf{BF}_s \exp{(\alpha \,+\, F_s\beta)}}{1 \,+\, \mathbf{BF}_s \exp{(\alpha \,+\, F_s\beta)}}. \tag{1}$$

Provided that estimates of α and β are available from an external analysis such as described in the next section, one can estimate $\Pr(A_s \;=\; 1 \,|\, D, \mathbf{X}, G_s, F_s)$ by

$$\sum_{i=1}^{I} \Pr(A_s \;=\; 1 \,|\, D, \mathbf{X}, G_s, F_s, \alpha_i, \beta_i)/I$$

where the α_i and β_i are samples from the posterior distribution from an analysis such as described in Section 2.1.

The above procedure depends on estimates of the marginal likelihoods,

$$\Pr(D \,|\, \mathbf{X}, G_s, A_s = a) \;=\; \int \Pr(D \,|\, \mathbf{X}, G_s, A_s = a, \theta_a)\Pr(\theta_a),$$

of the association data for each SNP under H_o ($A_s = 0$) and under H_a ($A_s = 1$). $\Pr(D \,|\, \mathbf{X}, G_s, A_s = a, \theta_a)$ is a logistic regression of the disease status indicator, D, on the covariates, $\mathbf{X}$, and SNP genotype, G_s, and with coefficient vector θ_1 under H_a and is a logistic regression D on $\mathbf{X}$ with coefficient vector θ_0 under H_o. We place independent normal mean 0, standard deviation 10 prior distributions on all components of θ_0 and θ_1, with exception of the coefficient of G_s, which is accorded a normal mean 0, standard deviation 0.25 prior distribution, as the majority of log–odds estimates cited in the GWAS catalog are smaller than 0.5 in absolute value. We estimate the SNP–specific marginal likelihoods under each hypothesis of association using the Laplace approximation (Kass and Raftery, 1995) implemented in software described in Wilson et al. (2010) and available from the authors.

Since it is not always convenient or possible to directly calculate Bayes factors, we consider the performance of our method when applied to Bayes factors estimated from p–values using the approximation described in Sellke et al. (2001). These authors show that the Bayes factor *against* association can be approximated by the function $-ep_s \log_e(p_s)$ when $p_s < 1/e$ and 0.0 otherwise, where p_s is a p–value from a standard test of association, and that this function provides a lower bound for that quantity that is sharp (i.e. accurate) for $p_s < 1/e$. As a consequence, its multiplicative inverse provides a sharp upper bound on the Bayes factor *in favor of* association. We evaluate our method using both this and the ratio of Laplace approximations described above to calculate $\mathbf{BF}_s$ in Equation 1.

4.2 Construction of Functional Signatures

In what follows, we detail the steps we took to assemble the case–control study of SNPs used to build and evaluate the models for a variant's association status given the functional data. These comprised identification of a representative set of phenotype–associated SNPs to serve as 'cases' in the analysis and a matched set of 'controls' and the collection of a set of measurements related to function to serve as annotations for the variants. The process is depicted in Figure 4 and described in detail below.

Sampling Frame. Many genomewide genotyping arrays were designed to over–sample variants with characteristics related to their ability to explain phenotypic variation, such as proximity to coding regions, type of variant (e.g. missense) and minor allele frequency (MAF). Hence, a comparison of case SNPs identified using such assays to control SNPs drawn randomly from the HapMap or dbSNP, for example, may lead to spurious associations between assay design variables and the SNP's association with a human phenotype. In order to avoid confounding due to the selection method employed in the design of the genomewide genotyping platform, we constructed a sampling frame of SNPs by combining SNPs on the Affymetrix GeneChip Human Mapping 500K Array Set and the Illumina HumanHap550 Genotyping BeadChip, as this generation of arrays and their predecessors cover most of the reported findings in the GWAS Catalog attributed to an Affymetrix or Illumina product and these products were the most commonly used. We labeled SNPs in the sampling frame according to whether they appeared only on the Affymetrix list, only on the Illumina list or on both and confined attention to those variants appearing in both the

Genome Browser's dbSNP 130 and HapMap Release 27 tables (see Supplemental Table 1) and having a MAF estimated in HapMap's CEU sample to be 0.05 or larger. The sampling frame comprises 803,991 SNPs with 421,072 unique to Illumina, 305,672 to Affymetrix and the remaining 77,247 common to both.

Case and Control Selection. The GWAS Catalog is subject to constant update and versions are available from several locations. We downloaded the GWAS Catalog from the Genome Browser (time stamp and location in Supplemental Table 1). We confined attention to non–CNV variants in the GWAS Catalog discovered by association studies utilizing an Affymetrix and/or an Illumina genomewide array and present in the sampling frame. We randomly chose a single representative of each set of SNPs appearing multiple times in the GWAS Catalog or sharing one or more 'LD partners' (see below). This left 2093 unique case SNPs, 1306 of which were unique to Illumina, 403 unique to Affymetrix and 384 in common. We randomly matched one control SNP drawn from the sampling frame to each case SNP on chromosome, platform (Illumina only, Affymetrix only, on both) and MAF rounded to the nearest 0.02. We excluded SNPs in the sampling frame in LD ($R^2 > 0$) with one or more case SNPs as reported in the HapMap Release #27 LD files (see Supplemental Table 1) or sharing an LD partner with another control SNP.

LD Partner Identification. SNPs in the GWAS Catalog are arguably more likely to tag the variant that is directly associated with the phenotype than to be that variant (Hindorff et al., 2009). Hence, following Hindorff et al. (2009) we identified and annotated each case and control SNP's 'LD partners.' We defined LD partners as those SNPs with $R^2 \geq 0.8$ with a case or control SNP as reported in the HapMap Release #27 LD files. Hindorff et al. (2009) chose a threshold of 0.9 but noted that their results were nearly the same when using thresholds of 1.0 and 0.8. We identified 20,924 LD partners of the case SNPs and 23,779 LD partners of the control SNPs.

Annotation Data. All data drawn from the UCSC Genome Browser (Rhead et al. (2010)) used the "March 2006 (NCBI36/hg18)" assembly. Supplemental Table 1 provides locations, revision dates and references for each of the annotation files referred to below. In what follows, we describe each class of annotation variable, its source and the parameterization we use for it in the models we fit.

Variants described in dbSNP (Sherry et al., 2001) release 130 are classified according to their predicted function as determined by their locations relative to known genes in the reference assembly. Variants that fall within the coding sequence of a known gene are further described as 'non–synonymous' if they result in a change to the associated amino acid or 'synonymous' if they do not. A variant may have several such designations; for purposes of our analysis, we confine attention to each variant's primary designation. Those observed among the SNPs included in our analysis are 'unknown,' 'coding–synon,' 'intron,' 'near–gene–3,' 'near–gene–5,' 'nonsense,' 'missense,' 'untranslated–3,' and 'untranslated–5.' Given the small number ($n = 5$) of nonsense variants, we created a 'coding–nonsynonymous' designation by combining the 'missense' and 'nonsense' categories; similarly, we combined the 'untranslated–3,' and 'untranslated–5' designations into the category 'untranslated.'

Measures of sequence conservation are frequently employed as evidence regarding the disease association status of rare missense variants (Tavtigian et al., 2008). We examined the PhyloP evolutionary conservation scores of Siepel et al. (2006) applied to 28– and 44–species alignments, and to those alignments restricted to the placental mammals and human, for their ability to predict the disease association status of common variants. Each of the four relevant Genome Browser tables provides the sum of the score, its sum of squares and the number of nucleotides that contribute to these statistics within ranges of contiguous nucleotides. We calculated a standardized score (mean divided by standard deviation) for each range and alignment and assigned these values to SNPs within the range. The PhyloP conservation scores exhibited pairwise correlations of up to 0.984. The top four, six, and nine PCs explain 90%, 96%, and 99% of the variability, respectively, in the 24 variables. The top four PCs were used.

The Database of Genomic Variants (DGV; Iafrate et al. (2004); Zhang et al. (2006)) is a compilation of reported genomic alterations spanning more than 1000 bases (>100 in the case of indels) observed in healthy subjects. We formed three variables indicating, respectively, whether (=1) or not (=0) each SNP falls in a region of copy number variation, a region containing insertions or deletions or a region with inversions.

The ENCODE Project (The ENCODE Project Consortium, 2007, 2011) is an ambitious project to identify and characterize the various functional elements present in the human genome sequence and to facilitate public access to the data it generates; its overarching objective is to improve our knowledge of human disease processes by providing a more comprehensive understanding of human molecular biology. Application of ENCODE functional annotation data to the design, analysis and interpretation of GWAS studies is one way in which ENCODE data can quickly be put to use to shed light on human disease processes (The ENCODE Project Consortium, 2011). To this end, we examine the utility of the recently released ENCODE regulation supertrack data available from, and displayed on, the Genome Browser for *a priori* prediction of functional, disease-associated variants. In particular, we include variables (see below) summarizing: transcription levels assayed in six cell lines by RNA–seq (Mortazavi et al., 2008; Langmead et al., 2009) and represented as a density measure of signal enrichment ('raw signal'); density measures of signal enrichment for H3K4Me1 (Histone H3 Lysine 4 monomethylation) associated with enhancer and promoter activity measured in eight cell lines, similarly coded measures of promoter– and enhancer–associated H3K27Ac (Histone H3 Lysine 27 acetylation) in eight cell lines and of promoter–associated H3K4Me3 (Histone H3 Lysine 4 tri–methylation) in nine cell lines (Bernstein et al., 2006; Mikkelsen et al., 2007); evidence for the variant falling within a DNaseI hypersensitivity cluster (Sabo et al., 2006, 2004); and the evidence for transcription factor binding measured via ChIP–seq (Euskirchen et al., 2004, 2007; Martone et al., 2003; Robertson et al., 2007; Rozowsky et al., 2009).

The Broad ChIP–seq, Caltech RNA–seq, and PhyloP signal tracks are summarized at the level of genomic bins. The ChIP–seq signals are measured within 118,084 contiguous bins of 25,600 bases apiece. The RNA–seq and PhyloP signals are measured in sets of non-overlapping, non-uniform bins. Bins are indexed according to the hierarchical scheme described in (Kent et al., 2002).

The ENCODE database provides basic summary statistics (the minimum, range, count, sum and sum of squares) of the signal enrichment density measures within each bin. For purposes of our analysis, we summarized each cell line's bin level data by $\log_e$(sum/count), $\log_e$(maximum), and $\log_e$(z), where z is the standardized score; in bins where sum=0, we set sum=1 (sums are non–negative and range across six orders of magnitude).

The three Broad types (signal enrichment for H3K4Me1, H3K27Ac and H3k4Me3 histone modifications) comprise data on eight, eight and nine cell lines, respectively. We found significant pairwise correlations among the 75 variables (25 each of log(mean), log(maximum), and log(z)), ranging as high as 0.949, and therefore conducted a principal components analysis to identify the linear combinations, i.e. principal components (PCs), that explain most of the variability in the data. The top 18, 27 and 44 PCs explain 90%, 95% and 99% of variability, respectively, in the 75 measures. Finally, we mapped each SNP to the appropriate Broad bin and annotated each with the top 18 PCs for purposes of the analysis.

The Caltech tables comprise RNA–seq raw signal enrichment data on six cell lines. In addition to the three ENCODE variables described above, an indicator variable for a SNP falling within a bin was included. Pairwise correlations among the 24 variables ranged as high as 0.970. The top 11, 12, and 16 PCs explain 92%, 95%, and 99% of the variability. The top 11 PCs were used in the analysis.

The transcription factor ChIP–seq data are summarized by scores, ranging from 6 to 1000, measuring strength of evidence for binding within specified, sometimes overlapping, chromosomal bins ('clusters'). We summarize these data as they apply to each SNP using two variables: the number of clusters it intersects with ('TFBSfreq') and the average $\log_e$(score) ('logTFBS') assigned to those clusters (coded as 0 if the SNP does not intersect with a cluster). Similarly, the DNaseI hypersensitivity data are summarized by scores, ranging from 16 to 1000, within specified chromosomal bins ('clusters'). We summarize these data as they apply to each SNP using (1) an indicator for the variant falling within a clusters and (2) the $\log_e$(score) assigned to that cluster.

The Open REGulatory ANNOtation database (ORegAnno) Montgomery et al. (2006); Griffith et al. (2008) is a curated collection of regulatory elements. The Genome Browser ORegAnno table provides start and stop coordinates and annotations for elements in the database. For purposes of our analysis, we summarize these data with an indicator variable for whether or not a variant falls within an ORegAnno regulatory region.

PolyPhen–2 (PPh2; Adzhubei et al. (2010)) assigns to nonsynonymous SNPs a probability of being damaging based on the sequence, phylogenetic and structural information characterizing the amino acid substitution.

RegulomeDB (Boyle et al., 2012) annotates SNPs with known and predicted regulatory elements in the intergenic regions of the human genome. Each SNP is assigned one of seven categories based on its likelihood of affecting protein binding.

Model. For purposes of the analysis, we assumed that blocks and the SNPs within the blocks were independent conditional on the functional data. We modeled the probability that a SNP s in block b was an associated SNP, π_{sb}, given the functional data for that SNP, F_{sb}, using the logistic regression model $\mathrm{logit}(\pi_{\mathrm{sb}}) = \alpha + \mathrm{F}_{\mathrm{sb}}\beta$. We assumed that there was at least one associated SNP in each case block and that there were no associated SNPs in control blocks. Hence, each case block contributed the factor $[1 - \prod_{s=1}^{n_b}(1 - \pi_{sb})]$ to the likelihood, while each control block contributed $\prod_{s=1}^{n_b}(1 - \pi_{sb})$. As a result, we expected *at least* 1,675 of the 48,888 SNPs in the training set to be phenotype associated. This corresponds to $\alpha = -3.34$ (columns of F are centered); if 10% (alt 20%) of case blocks contain two phenotype associated SNPs, $\alpha = -3.24$ (alt -3.15). Hence the normal mean -3.24, standard deviation 0.1 prior distribution we placed on α is consistent with our expectation that there were fewer than 2,178 ($= 1.3 \times 1,675$) true phenotype associated variants among the case blocks.

Our specification of the prior distribution on β was guided by the observation that, in the normal model with the normal–exponential–gamma (NEG) distribution as prior on the mean and the variance known, the posterior mode is identically zero when the maximum likelihood estimator (MLE) is in a neighborhood around zero, but rapidly converges to the MLE as the MLE diverges from zero (this setting approximates the more general one in which the NEG distribution is used as the prior distribution for a parameter whose likelihood is approximately normal). The NEG distribution is specified by its shape and scale parameters and the width of the threshold neighborhood is a function of these parameters. For purposes of our analysis, we chose parameter values for which no more than 10% of the coefficients are outside of the threshold region with probability 0.90, *a priori*. We placed independent NEG prior distributions on the components of β; in addition, we also considered the model with independent standard normal distributions on the components of β. Inference for each of these models was carried out using the training set and were evaluated using the evaluation set.

We used random–walk Markov Chain Monte Carlo (MCMC) algorithms (Metropolis et al., 1953; Gilks et al., 1996) to estimate summaries of the posterior distribution under each of the models. We started 10 independent chains per model from starting points drawn from the prior distribution. In each case, step sizes were adusted so that parameter level acceptance ratios fell between 0.3 and 0.5 during an initial, 'burn–in' set of iterations not used for inference. We fixed the step sizes and ran the 10 chains from their leave–off positions for an additional 50,000 iterations per chain. Inspection of trace plots, as well as computation of the Gelman–Rubin (Gelman and Rubin, 1992), Heidelberger–Welch (Heidelberger and Welch, 1983), Raftery–Lewis (Raftery and Lewis, 1996), and Geweke (Geweke, 1992) diagnostics implemented in the CODA package (Plummer et al., 2010) in R (Ihaka and Gentleman, 1996), indicated satisfactory convergence. We thinned the 10 chains by 1,000 and combined them to produce a sample of 500 coefficient vectors.

We used the concordance index (CI) to measure the out–of–sample predictive accuracy of the model. We calculated the CI as the fraction of matched pairs in the 'evaluation set' in which the average probability of association given the functional data over the n_1 SNPs in the case block ('$b1$') was larger than the corresponding average over the n_0 SNPs in the matched control block ('$b0$'); i.e. if

$$\sum_{s=1}^{n_1} \Pr(A_{s,b1} \,|\, F_{s,b1})/n_1 \;>\; \sum_{s=1}^{n_0} \Pr(A_{s,b0} \,|\, F_{s,b0})/n_0,$$

where we estimated $\Pr(A_{s,bn} \,|\, F_{s,bn})$ by

$$\sum_{i=1}^{500} \Pr(A_{s,bn} \;=\; 1 \,|\, F_{s,bn}, \alpha_i, \beta_i)/500,$$

where the α_i and β_i are MCMC samples saved from analysis of the training data.

4.3 Evaluation

We carried out a genome–wide association analysis of serous ovarian cancer using the methods described above. The data for this analysis were drawn from GWAS studies conducted in the US (Permuth-Wey et al., 2011) and the UK (Song et al., 2009). The genotype data from these studies were combined and imputed to HapMap III density, resulting in an data set comprising analyzable genotypes at 2,500,004 SNPs for 7,272 subjects of European ancestry. The association analysis was confined to the 2,004 cases with advanced stage serous ovarian cancer and the 3,272 available controls and was adjusted for study site and the first two principal components of the sample genotypes. We calculated Bayes factors (BFs) as described above and set the prior probability of association to be 0.00001 when estimating posterior probabilities; ranks are invariant to this choice. P–values used in the Bayes factor approximation were from likelihood ratio tests of the model including versus the model excluding a SNP.

Several large scale studies conducted to follow up promising associations from these GWAS have identified the eleven genome–wide significant loci listed in Table 4. We treat these as established or 'true positive' associations for purposes of evaluating the various association measures. In addition, we identified a set of likely unassociated, 'true negative' SNPs from among 22,254 GWAS followup SNPs placed on the iCOGS chip (Pharoah et al., 2013). This analysis included 8,344 cases with advanced stage serous ovarian cancer and 22,913 controls of European ancestry and was adjusted for study site and the first five European ancestry principal components. We identified a subset of 5,155 SNPs with strong evidence *against association* (defined as BF<0.1 on Jeffreys' scale of evidence (Jeffreys, 1961)) to serve as the 'true negatives.'

We compared the rankings of these two sets of SNPs in the original GWAS analysis when association was measured using genotype data only to those obtained with incorporation of the functional signatures. We compared the procedures based on their power to identify the truly associated variants for follow–up assuming budgets allowing for evaluation of the top 5,000 or 10,000 SNPs.

In most association studies, genotypes are determined, through a combination of genotyping and imputation, for only a subset of the universe of variants. In this setting, it is standard to rely on correlations between genotyped variants ('tags') and those that are 'tagged' (not genotyped) to identify and localize associations. Likewise, the utility of functional signatures in a typical study will depend on the degree to which they reflect the likelihood of function of both the tag for which it is calculated and for the set of variants it tags. We evaluated correlations between functional signatures, defined as $(F_s\beta)$, for adjacent pairs of SNPs included in the ovarian cancer

GWAS analysis. We identified the quantile of each adjacent variant pair in the overall distribution of distances measured in base pairs (BPs). For purposes of this analysis, we defined quantiles in increments of 0.025, i.e. with each containing 2.5% of the mass of the distance distribution. We estimated the Pearson correlation between the functional signatures of the adjacent SNP pairs within each quantile and plotted these estimates against BP distance, locating the estimates at the midpoints of the quantile bins.

Data Access

The data used to construct and evaluate the functional signatures we describe are available at `ftp://stat.duke.edu/pub/Users/iversen/FunctionalSignatures/`.

Acknowledgements

Funding: This work was funded by the National Institutes of Health through its Genes and Environment Initiative, grant number R01–HL090559 and through R21–ES020796 (NIEHS) and by the National Science Foundation via DMS–1106891. The ovarian cancer multi–GWAS analyzed herein were provided by the FOCI project (1U19-CA148112) of NCI's GAME–ON consortium.

Disclosure Declaration

Conflict of Interest: None declared.

Figure Legends

Figure 1

Figure 1: Two–staged procedure for integrating variant–level functional annotation data with subject–level genetic association data. At the first stage, functional annotation data are combined to estimate the prior (to observing the genetic association data) probability of association for each variant. At stage two, these estimates are combined with the Bayes Factor (a metric of association) in favor of genetic association via Bayes' formula to estimate the posterior (to observing the functional and genetic association data) probability of association for each variant.

Figure 2

Figure 2: Construction and evaluation of models for (prior) probability of association given the functional annotation data. The purple arrows represent model construction ('training'), while the green arrows represent evaluation of the models. Construction of the training set, validation set and functional annotation database are depicted in Figure 4 and described in Methods. The training data were used to construct a series of models, each distinguished by the coefficients (or 'weights') it assigns to the various annotation variables. We chose the best amongst these by comparing their predictions in the validation set using the concordance index.

Figure 3

Figure 3: Correlation of functional signatures between adjacent HapMap II/III SNPs as a function of base pair distance (black line). Cumulative distribution function(CDF) of base pair distances across the genome (red line).

Figure 4

Figure 4: Construction of data sets and functional annotation database. Case SNPs from the GWAS Catalog are matched with control SNPs from HapMap III to generate training and validation sets. The matched SNP IDs and their locations are used to interrogate several online databases. These results are merged to build the functional annotation database.

Figures

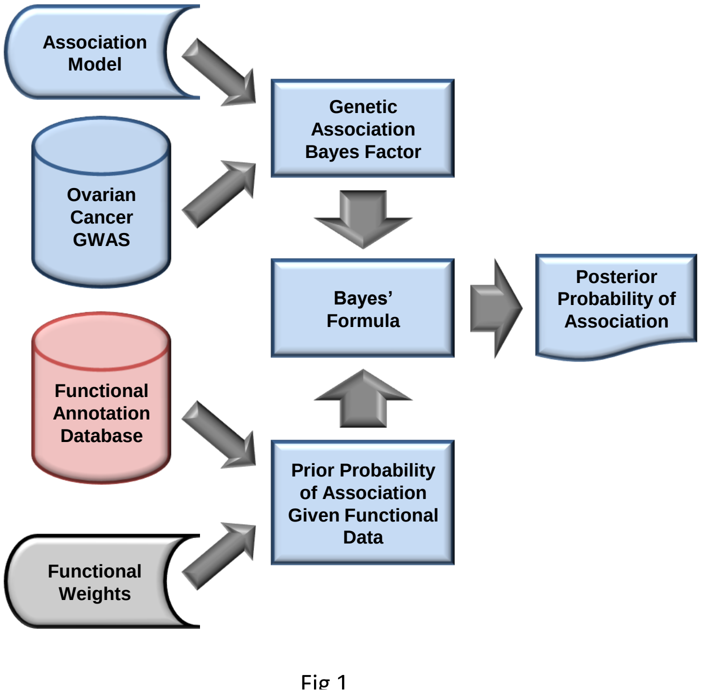

Association Model
Ovarian Cancer GWAS
Functional Annotation Database
Functional Weights
Genetic Association Bayes Factor
Bayes' Formula
Prior Probability of Association Given Functional Data
Posterior Probability of Association
Fig 1

Figure 1

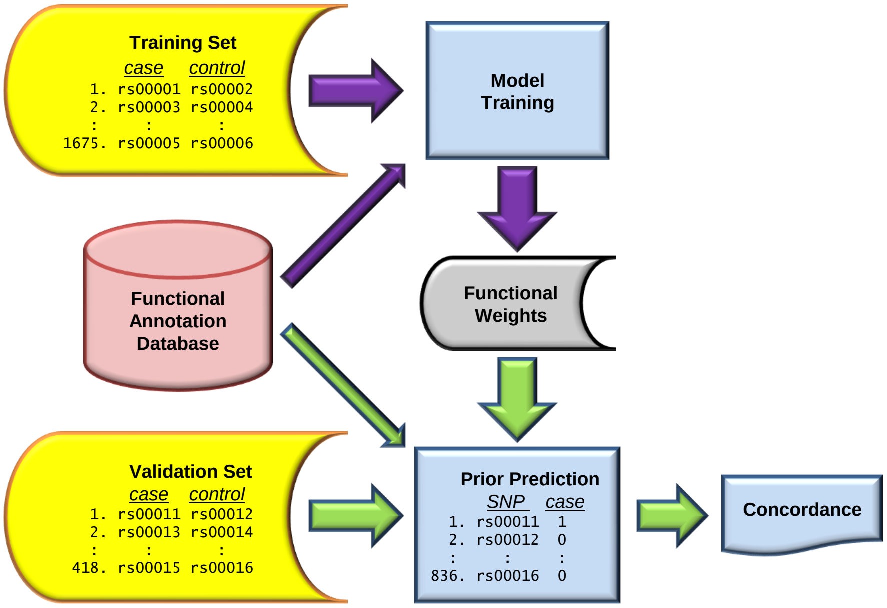


Fig 2

Figure 2

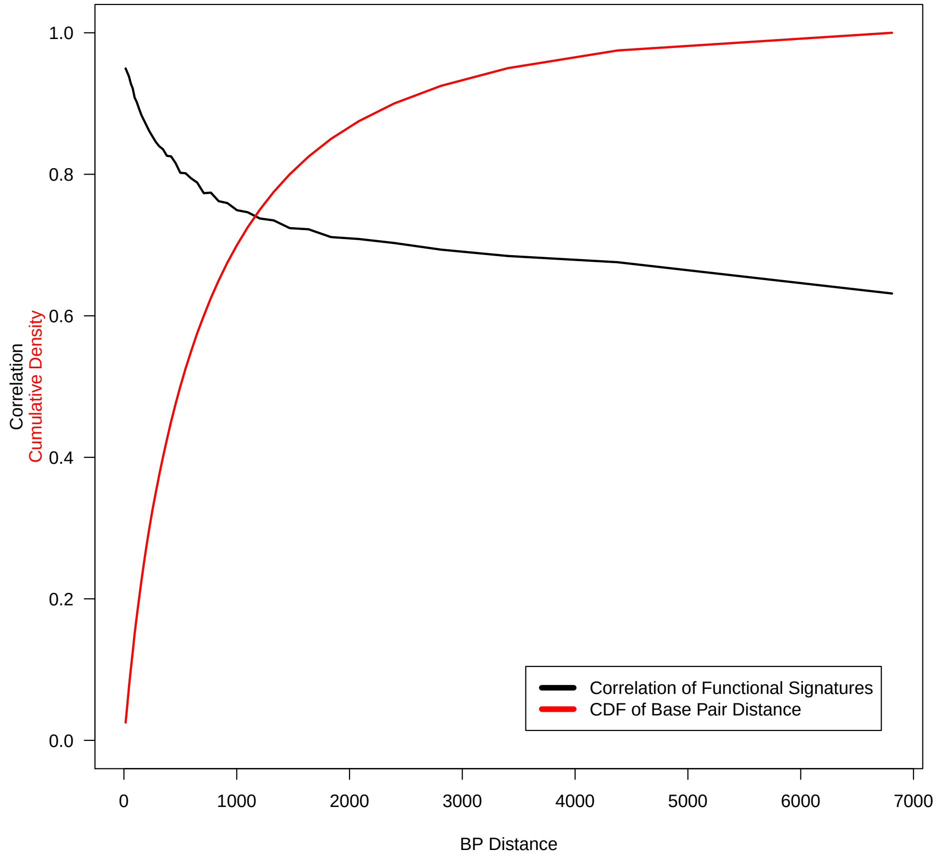

1.0
0.8
0.6
0.4
0.2
0.0
Correlation
Cumulative Density
0
1000
2000
3000
4000
5000
6000
7000
BP Distance
Correlation of Functional Signatures
CDF of Base Pair Distance

Figure 3

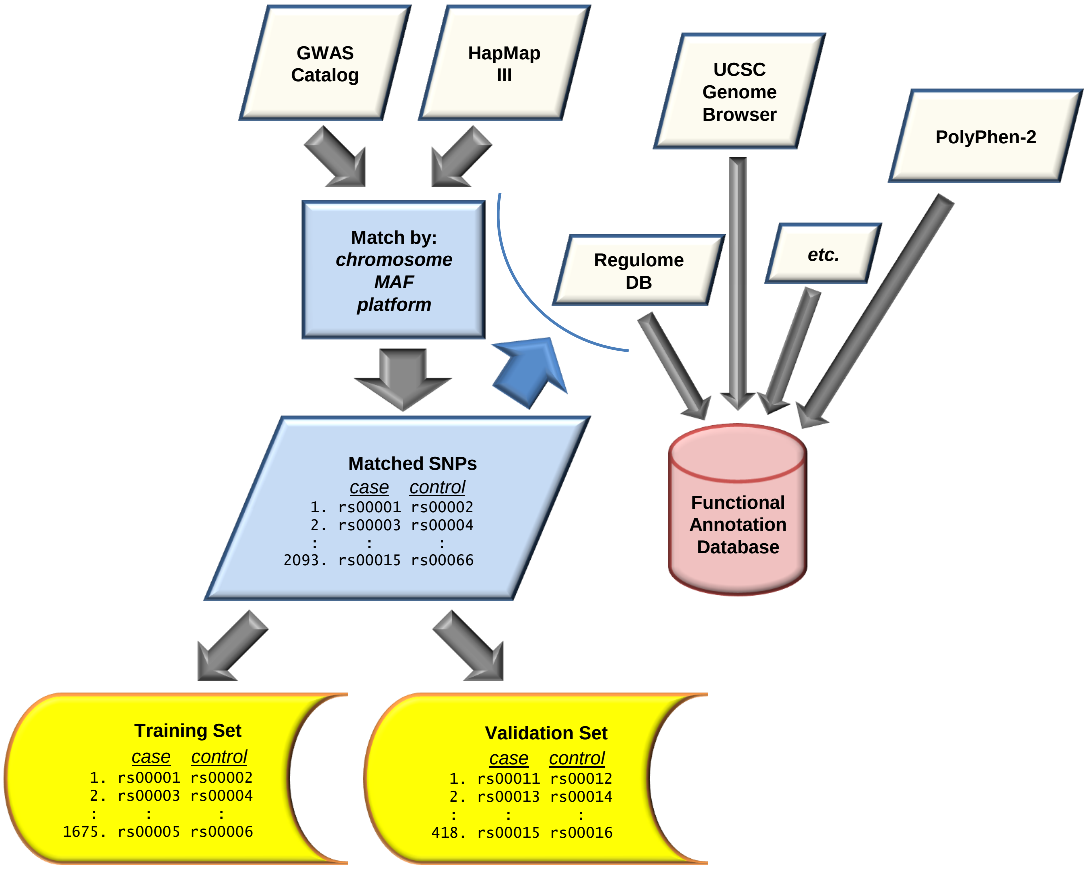


Fig 4

Figure 4

Tables

Table 1. Annotations used to construct the functional signatures. Definitions of the 54 variables appearing in the prior model for association status arranged by type/class of annotation.

Name	Annotation Class	Description
MAF*1..4*	Minor Allele Frequency	Natural spline basis for MAF
funcIntron	dbSNP Function Class	Indicator that variant is intronic.
funcNg3	dbSNP Function Class	Indicator that variant is near-gene-3.
funcNg5	dbSNP Function Class	Indicator that variant is near-gene-5.
funcNonsynon	dbSNP Function Class	Indicator that variant is missense or nonsense.
funcSynon	dbSNP Function Class	Indicator that variant is synonymous.
funcUTR	dbSNP Function Class	Indicator that variant is in the 3′ or 5′ UTR.
PhyPC*1..4*	phyloP Evol. Cons. Score	First 4 PCs for PhyloP data.
IndelInd	DGV Regions	Indicator that SNP is in the region of a known in–del.
CNVInd	DGV Regions	Indicator that SNP is in the region of a known CNV.
InvInd	DGV Regions	Indicator that SNP is in the region of a known inversion.
BrPC*1..18*	ENCODE Super Track	PCs of Broad promoter/enhancer ChIP–seq data.
CalPC*1..11*	ENCODE Super Track	PCs of CalTech transcription level RNA–seq data.
logDNase	ENCODE Regulatory Super Track	DNaseI hypersensitivity cluster log(score).
TFBSfreq	ENCODE Regulatory Super Track	SNP in ChIP–seq TFBS region(s) – count.
logTFBS	ENCODE Regulatory Super Track	SNP in ChIP–seq TFBS region(s) – log(TFBS score).
ORegInd	Open REGulatory ANNOtation DB	Indicator that SNP is in ORegAnno DB.
PPh2Prob	PolyPhen–2	Probability that SNP is damaging.
RegDBcat	RegulomeDB	RegulomeDB category.

Table 2. Summary of estimates for the model for association status given the functional annotation data. Estimates of the posterior mean and standard deviation are provided for each coefficient in the model along with the ratio of these quantities, a 'signal–to–noise' measure analogous to the Z statistic.

Coefficient	Mean	SD	Mean/SD	Coefficient	Mean	SD	Mean/SD
MAF1	0.029	0.0272	1.051	CalPC8	0.096	0.0608	1.584
MAF2	0.003	0.0185	0.151	CalPC9	0.009	0.0218	0.406
MAF3	0.018	0.0236	0.759	CalPC10	-0.019	0.0234	-0.809
MAF4	-0.008	0.0199	-0.425	CalPC11	-0.044	0.0468	-0.943
BrPC1	-0.348	0.0388	-8.983	PhyPC1	0.225	0.0452	4.982
BrPC2	0.174	0.0360	4.845	PhyPC2	-0.023	0.0308	-0.742
BrPC3	0.002	0.0172	0.099	PhyPC3	0.053	0.0395	1.336
BrPC4	-0.077	0.0301	-2.561	PhyPC4	0.024	0.0289	0.839
BrPC5	0.149	0.0301	4.932	funcIntron	-0.003	0.0199	-0.160
BrPC6	0.097	0.0329	2.961	funcNg3	0.002	0.0238	0.098
BrPC7	-0.019	0.0229	-0.825	funcNg5	0.003	0.0200	0.158
BrPC8	-0.078	0.0312	-2.498	funcNonsynon	0.089	0.0387	2.308
BrPC9	-0.012	0.0202	-0.573	funcSynon	-0.007	0.0236	-0.283
BrPC10	0.007	0.0182	0.407	funcUTR	0.002	0.0210	0.078
BrPC11	0.021	0.0239	0.887	logDNase	0.011	0.0253	0.419
BrPC12	-0.039	0.0290	-1.343	TFBSfreq	0.024	0.0267	0.885
BrPC13	-0.094	0.0318	-2.972	logTFBS	0.019	0.0299	0.641
BrPC14	-0.000	0.0175	-0.015	ORegInd	0.027	0.0236	1.163
BrPC15	-0.039	0.0286	-1.354	IndelInd	-0.023	0.0337	-0.693
BrPC16	0.009	0.0185	0.467	CNVInd	0.059	0.0321	1.842
BrPC17	-0.015	0.0213	-0.696	InvInd	0.090	0.0305	2.938
BrPC18	-0.014	0.0210	-0.688	rDBcat1	0.014	0.0257	0.550
CalPC1	-0.103	0.0492	-2.084	rDBcat2	0.066	0.0378	1.741
CalPC2	0.090	0.0441	2.030	rDBcat3	0.012	0.0264	0.467
CalPC3	-0.019	0.0270	-0.709	rDBcat4	0.116	0.0461	2.508
CalPC4	0.086	0.0457	1.889	rDBcat5	0.106	0.0527	2.003
CalPC5	0.053	0.0395	1.350	rDBcat6	-0.056	0.0552	-1.018
CalPC6	-0.012	0.0233	-0.496	pph2prob	0.078	0.0269	2.905
CalPC7	0.002	0.0207	0.078				

Table 3. Means and 95% interval estimates of the concordance indices for each of the four models.

		Concordance Index	
Label	**Prior**	**Mean**	**95% CI**
Normal	N(0, 1)	0.6348	(0.6112, 0.6555)
NEG1	NEG(0.834, 0.1610)	0.6397	(0.6148, 0.6615)
NEG2	NEG(0.950, 0.0588)	0.6433	(0.6208, 0.6675)
NEG3	NEG(0.978, 0.0245)	0.6487	(0.6244, 0.6675)

Table 4. Functional signatures improve inference for association status in a GWAS of ovarian cancer. Ranks of known associated variants (labeled ’true $+$’) tend to improve (i.e. are closer to one) when association and functional data are incorporated in the analysis (Rank$_{\mathrm{A+F}}$) relative to when only the association data are used (Rank$_{\mathrm{A}}$) and, hence, are more likely to be studied further. Conversely, ranks of (very likely) *unassociated* variants (labeled ’true $-$’) tend to fall with inclusion of the functional data. The functional data for a given variant is summarized by its ’functional signature,’ defined as the prior log–odds of its association given the functional data (LO$_{\mathrm{F}}$). Aggregate (mean and median) values are provided for the true $+$ set and the true $-$ set. Results are provided both for when the Bayes Factors in favor of genetic association (BF$_{\mathrm{A}}$) are estimated from a Bayesian analysis and for when they are approximated using a transformation of p–values. Ranks are out of approximately 2.5M variants.

				Bayesian Analysis			P–Value Approximation		
Variant	**Locus**	**MAF**	**LO$_{\mathrm{F}}$**	**log(BF$_{\mathrm{A}}$)**	**Rank$_{\mathrm{A}}$**	**Rank$_{\mathrm{A+F}}$**	**log(BF$_{\mathrm{A}}$)**	**Rank$_{\mathrm{A}}$**	**Rank$_{\mathrm{A+F}}$**
rs2072590	2q31	0.34	1.46	8.63	65	59	8.97	65	59
rs2665390	3q25	0.09	0.77	8.08	77	73	8.42	76	71
rs10069690	5p15	0.23	0.91	-1.38	1,549,122	651,710	0.00	1,716,235	378,319
rs11782652	8q21	0.08	0.22	2.98	5,272	6,843	3.23	6,602	9,476
rs7814937	8q24	0.12	1.54	14.61	21	16	15.11	21	17
rs3814113	9p22	0.30	-0.09	14.01	38	38	14.31	38	38
rs7084454	10p12	0.31	1.44	1.19	45,616	12,221	1.81	38,214	11,018
rs757210	17q12	0.37	1.74	2.31	11,630	2,411	2.86	10,245	2,177
rs2077606	17q21	0.18	0.70	-0.25	339,456	200,494	0.48	255,254	228,953
rs9303542	17q21	0.27	0.05	3.70	2,276	3,532	4.13	2,293	3,852
rs8170	19p13	0.19	0.82	2.72	7,133	4,179	3.13	7,479	4,846
Mean	True $+$	0.23	0.87	5.15	178,246	80,143	5.68	185,138	58,075
Median	True $+$	0.23	0.82	2.98	5,272	3,532	3.23	6,602	3,852
Mean	True $-$	0.35	0.11	0.37	438,664	517,810	1.21	310,830	554,051
Median	True $-$	0.36	0.06	0.14	181,116	244,393	0.91	129,608	267,892

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nature Methods* **7**: 248–249.

Aragaki CC, Greenland S, Probst-Hensch N, and Haile RW. 1997. Hierarchical modeling of gene-environment interactions: estimating NAT2 genotype–specific dietary effects on adenomatous polyps. *Cancer Epidemiology Biomarkers & Prevention* **6**: 307–314.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al.. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**: 315–326.

Bojesen SE, Pooley KA, Johnatty SE, Beesley J, Michailidou K, Tyrer JP, Edwards SL, Pickett HA, Shen HC, Smart CE, et al.. 2013. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nature Genetics* **45**: 371–384.

Bolton KL, Tyrer J, Song H, Ramus SJ, Notaridou M, Jones C, Sher T, Gentry-Maharaj A, Wozniak E, Tsai YY, et al.. 2010. Common variants at 19p13 are associated with susceptibility to ovarian cancer. *Nature Genetics* **42**: 880–884.

Boyle A, Hong E, Hariharan M, Cheng Y, Schaub M, Kasowski M, Karczewski K, Park J, Hitz B, Weng S, et al.. 2012. Annotation of functional variation in personal genomes using regulomedb. *Genome Research* **22**: 1790–1797.

Euskirchen G, Royce TE, Bertone P, Martone R, Rinn JL, Nelson FK, Sayward F, Luscombe NM, Miller P, Gerstein M, et al.. 2004. CREB binds to multiple loci on human chromosome 22. *Molecular Cell Biology* **24**: 3804–3814.

Euskirchen GM, Rozowsky JS, Wei CL, Lee WH, Zhang ZD, Hartman S, Emanuelsson O, Stolc V, Weissman S, Gerstein MB, et al.. 2007. Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array– and sequencing–based technologies. *Genome Research* **17**: 898–909.

Freedman ML, Monteiro ANA, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, Biasi MD, Carlson C, Duggan D, et al.. 2011. Principles for the post–GWAS functional characterization of cancer risk loci. *Nature Genetics* **43**: 513–518.

Gelman A and Rubin DB. 1992. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**: 457–511.

Geweke J. 1992. Evaluating the accuracy of sampling–based approaches to calculating posterior moments. In *Bayesian Statistics 4* (eds. J Bernado, J erger, D AP, and A Smith). Clarendon Press, Oxford, UK.

Gilks WR, Richardson S, and Spiegelhalter DJ. 1996. Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (eds. WR Gilks, S Richardson, and DJ Spiegelhalter). Chapman and Hall, London.

Goode EL, Chenevix-Trench G, Song H, Ramus SJ, Notaridou M, Lawrenson K, Widschwendter M, Vierkant RA, Larson MC, Kjaer SK, et al.. 2010. A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nature Genetics* **42**: 874–879.

Griffin J and Brown P. 2007. Bayesian adaptive lassos with non–convex penalization. Technical report, University of Kent.

Griffin J and Brown P. 2010. Inference with normal–gamma prior distributions in regression problems. *Bayesian Analysis* **5**: 171–188.

Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, et al.. 2008. ORegAnno: an open–access community–driven resource for regulatory annotation. *Nucleic Acids Research* **36**: D107–D113.

Hans CM. 2009. Bayesian lasso regression. *Biometrika* **96**: 835–845.

Heidelberger P and Welch P. 1983. Simulation run length control in the presence of an initial transient. *Operations Research* **31**: 1109–1144.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**: 9362–9367.

Hoggart C, Whittaker J, De Iorio M, and Balding D. 2008. Simultaneous analysis of all SNPs in genome–wide and re–sequencing association studies. *PLoS Genetics* **4**: e1000130.

Hung RJ, Baragatti M, Thomas D, McKay J, Szeszenia-Dabrowska N, Zaridze D, Lissowska J, Rudnai P, Fabianova E, Mates D, et al.. 2007. Inherited predisposition of lung cancer: a hierarchical modeling approach to DNA repair and cell cycle control pathways. *Cancer Epidemiology Biomarkers & Prevention* **16**: 2736–44.

Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Boffetta P, and Witte JS. 2004. Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. *Cancer Epidemiology Biomarkers & Prevention* **13**: 1013–21.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, and Lee C. 2004. Detection of large–scale variation in the human genome. *Nat Genet* **36**: 949–951.

Ihaka R and Gentleman R. 1996. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**: 299–314.

Jeffreys H. 1961. *Theory of Probability*. Oxford Univ. Press, 3rd edition.

Kass RE and Raftery AE. 1995. Bayes factors. *J. Amer. Statist. Assoc.* **90**: 773–795.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D. 2002. The human genome browser at UCSC. *Genome Research* **12**: 996–1006.

Langmead B, Trapnell C, Pop M, and Salzberg SL. 2009. Ultrafast and memory–efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**: R25.

Manolio TA. 2010. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine* **363**: 166–176. PMID: 20647212.

Marchini J, Howie B, Myers S, McVean G, and Donnelly P. 2007. A new multipoint method for genomewide association studies by imputation of genotypes. *Nature Genetics* **39**: 906–913.

Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, Luscombe NM, Rinn JL, Nelson FK, Miller P, Gerstein M, et al.. 2003. Distribution of nf–κb–binding sites across human chromosome 22. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 12247–12252.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, and Teller E. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**: 1087–1091.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al.. 2007. Genome–wide maps of chromatin state in pluripotent and lineage–committed cells. *Nature* **448**: 553–560. 10.1038/nature06008.

Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, and Jones SJM. 2006. ORegAnno: an open access database and curation system for literature–derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics* **22**: 637–640.

Mortazavi A, Williams BA, McCue K, Schaeffer L, and Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**: 621–628. 10.1038/nmeth.1226.

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, and Cox NJ. 2010. Trait–associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genetics* **6**: e1000888.

Permuth-Wey J, Kim D, Tsai YY, Lin HY, Chen YA, Barnholtz-Sloan J, Birrer MJ, Bloom G, Chanock SJ, Chen Z, et al.. 2011. LIN28B polymorphisms influence susceptibility to epithelial ovarian cancer. *Cancer Research* **71**: 3896–3903.

Permuth-Wey J, Lawrenson K, Shen HC, Velkova A, Tyrer JP, Chen Z, Lin HY, Ann Chen Y, Tsai YY, Qu X, et al.. 2013. Identification and molecular characterization of a new ovarian cancer susceptibility locus at 17q21.31. *Nature Communications* **4**: 1627.

Pharoah PDP, Tsai YY, Ramus SJ, Phelan CM, Goode EL, Lawrenson K, Buckley M, Fridley BL, Tyrer JP, Shen H, et al.. 2013. GWAS meta–analysis and replication identifies three novel susceptibility loci for ovarian cancer. *Nature Genetics* **45**: 362–370.

Plummer M, Best N, Cowles K, and Vines K. 2010. *CODA: Output analysis and diagnostics for MCMC*. R package version 0.13-5.

Raftery AE and Lewis SM. 1996. Implementing MCMC. In *Markov Chain Monte Carlo in Practice* (eds. WR Gilks, S Richardson, and DJ Spiegelhalter), pp. 115–127. Chapman and Hall, London.

Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, et al.. 2010. The UCSC genome browser database: update 2010. *Nucleic Acids Research* **38**: D613–D619.

Richardson S, Bottolo L, and Rosenthal JS. 2011. Bayesian models for sparse regression analysis of high dimensional data. In *Bayesian Statistics 9* (eds. JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, and AFM Smith). Oxford University Press, Oxford.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al.. 2007. Genome–wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**: 651–657. 10.1038/nmeth1068.

Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, and Gerstein MB. 2009. PeakSeq enables systematic scoring of ChIP–seq experiments relative to controls. *Nature Biotech* **27**: 66–75. 10.1038/nbt.1518.

Sabo PJ, Hawrylycz M, Wallace JC, Humbert R, Yu M, Shafer A, Kawamoto J, Hall R, Mack J, Dorschner MO, et al.. 2004. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proceedings of the National Academy of Sciences* **101**: 16837–16842.

Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, et al.. 2006. Genome–scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature Methods* **3**: 511–518. 10.1038/nmeth890.

Schaub MA, Boyle AP, Kundaje A, Batzoglou S, and Snyder M. 2012. Linking disease associations with regulatory information in the human genome. *Genome Research* **22**: 1748–1759.

Sellke T, Bayarri MJ, and Berger JO. 2001. Calibration of p values for testing precise null hypotheses. *The American Statistician* **55**: 62–71.

Servin B and Stephens M. 2007. Imputation–based analysis of association studies: Candidate regions and quantitative traits. *PLOS Genetics* **3**.

Sherry S, Ward M, Kholodov M, Baker J, Phan L, Smigielski E, and Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Reseach* **29**: 308–11.

Siepel A, Pollard K, and Haussler D. 2006. New methods for detecting lineage–specific selection. *Research in Computational Molecular Biology* **3909**: 190–205.

Song H, Ramus SJ, Tyrer J, Bolton KL, Gentry-Maharaj A, Wozniak E, Anton-Culver H, Chang-Claude J, Cramer DW, DiCioccio R, et al.. 2009. A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. *Nature Genetics* **42**: 996–1000.

Tavtigian S, Greenblatt M, Lesueur F, and Byrnes GB. 2008. *In silico* analysis of missense substitutions using sequence–alignment based methods. *Human Mutation* **29**: 1327–36.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.

The ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology* **9**: e1001046.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Wilson MA, Iversen ES, Clyde MA, Schmidler SC, and Schildkraut JM. 2010. Supplement to “Bayesian Model Search and Multilevel Inference for SNP Association Studies”.

Witte JS, Greenland S, Haile RW, and Bird CL. 1994. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology* **5**: 612–621.

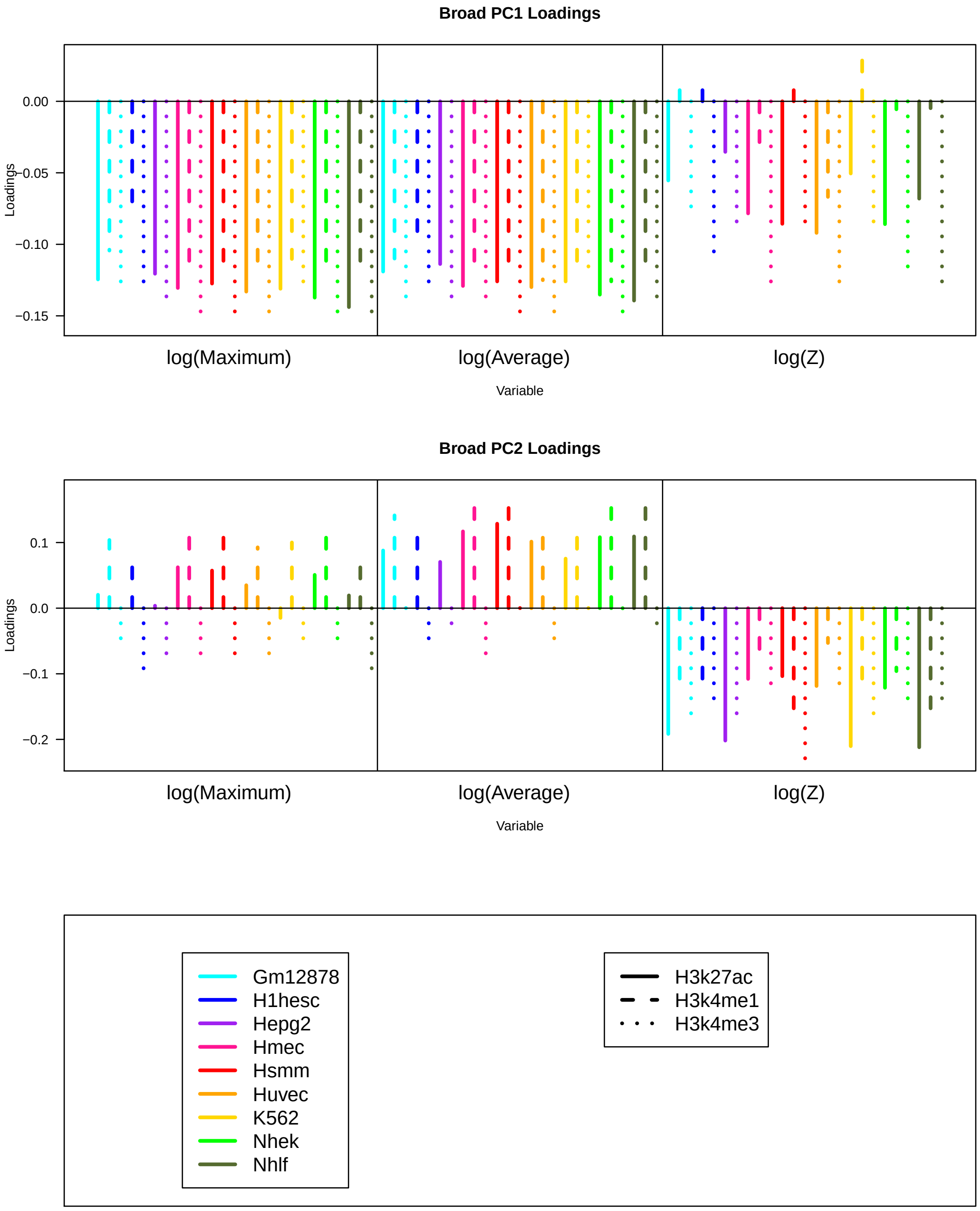
Zhang J, Feuk L, Duggan GE, Khaja R, and Scherer SW. 2006. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenetic & Genome Research* **115**: 205–214.

Supplement

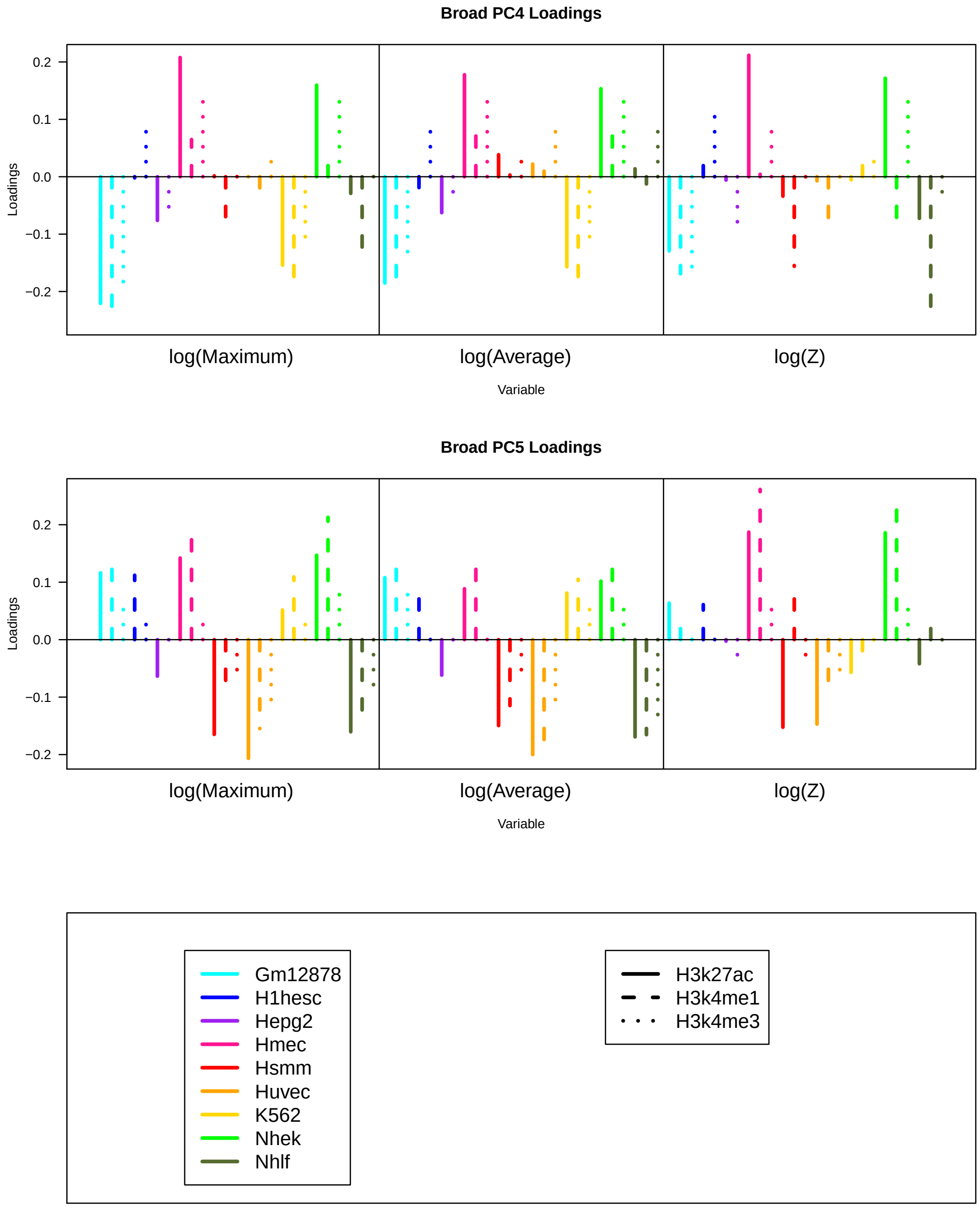
Annotation	Location	Date[◇]
GWAS Catalog	gwasCatalog.txt	11/15/10
Affymetrix 500K Set	snpArrayAffy250{Nsp, Sty}.txt	08/02/07
Illumina 550K Set	snpArrayIllumina550.txt	08/02/07
dbSNP 130	snp130.txt	09/20/09
HapMap Rel27	hapmapSnpsCEU.txt	07/11/07
HapMap Rel27 LD[†]	ld.chr{1..22}.CEU.txt	02/09
Broad GM12878 H3K27ac	wgEncodeBroadChipSeqSignalGm12878H3k27ac.txt.gz	06/22/10
Broad GM12878 H3K4me1	wgEncodeBroadChipSeqSignalGm12878H3k4me1.txt.gz	06/22/10
Broad GM12878 H3K4me3	wgEncodeBroadChipSeqSignalGm12878H3k4me3.txt.gz	06/22/10
Broad H1hesc H3K4me1	wgEncodeBroadChipSeqSignalH1hescH3k4me1.txt.gz	06/22/10
Broad H1hesc H3K4me3	wgEncodeBroadChipSeqSignalH1hescH3k4me3.txt.gz	06/22/10
Broad HepG2 H3K27ac	wgEncodeBroadChipSeqSignalHepg2H3k27ac.txt.gz	06/22/10
Broad HepG2 H3K4me3	wgEncodeBroadChipSeqSignalHepg2H3k4me3.txt.gz	06/22/10
Broad HMEC H3K27ac	wgEncodeBroadChipSeqSignalHmecH3k27ac.txt.gz	06/22/10
Broad HMEC H3K4me1	wgEncodeBroadChipSeqSignalHmecH3k4me1.txt.gz	06/22/10
Broad HMEC H3K4me3	wgEncodeBroadChipSeqSignalHmecH3k4me3.txt.gz	06/22/10
Broad HSMM H3K27ac	wgEncodeBroadChipSeqSignalHsmmH3k27ac.txt.gz	06/22/10
Broad HSMM H3K4me1	wgEncodeBroadChipSeqSignalHsmmH3k4me1.txt.gz	06/22/10
Broad HSMM H3K4me3	wgEncodeBroadChipSeqSignalHsmmH3k4me3.txt.gz	06/22/10
Broad HUVEC H3K27ac	wgEncodeBroadChipSeqSignalHuvecH3k27ac.txt.gz	06/22/10
Broad HUVEC H3K4me1	wgEncodeBroadChipSeqSignalHuvecH3k4me1.txt.gz	06/22/10
Broad HUVEC H3K4me3	wgEncodeBroadChipSeqSignalHuvecH3k4me3.txt.gz	06/22/10
Broad K562 H3K27ac	wgEncodeBroadChipSeqSignalK562H3k27ac.txt.gz	06/22/10
Broad K562 H3K4me1	wgEncodeBroadChipSeqSignalK562H3k4me1.txt.gz	06/22/10
Broad K562 H3K4me3	wgEncodeBroadChipSeqSignalK562H3k4me3.txt.gz	06/22/10
Broad NHEK H3K27ac	wgEncodeBroadChipSeqSignalNhekH3k27ac.txt.gz	06/22/10
Broad NHEK H3K4me1	wgEncodeBroadChipSeqSignalNhekH3k4me1.txt.gz	06/22/10
Broad NHEK H3K4me3	wgEncodeBroadChipSeqSignalNhekH3k4me3.txt.gz	06/22/10
Broad NHLF H3K27ac	wgEncodeBroadChipSeqSignalNhlfH3k27ac.txt.gz	06/22/10
Broad NHLF H3K4me1	wgEncodeBroadChipSeqSignalNhlfH3k4me1.txt.gz	06/22/10
Broad NHLF H3K4me3	wgEncodeBroadChipSeqSignalNhlfH3k4me3.txt.gz	06/22/10
Caltech Rep1 GM12878 Long PolyA BB1 2x75	wgEncodeCaltechRnaSeqRawSignalRep1Gm12878CellLongpolyaBb12x75.txt.gz	12/20/09
Caltech Rep1 H1hesc PAP BB2R 2x75	wgEncodeCaltechRnaSeqRawSignalRep1H1hescCellPapBb2R2x75.txt.gz	06/14/10
Caltech Rep1 HUVEC PAP BB2R 2x75	wgEncodeCaltechRnaSeqRawSignalRep1HuvecCellPapBb2R2x75.txt.gz	06/14/10
Caltech Rep1 K562 Long PolyA BB1 2x75	wgEncodeCaltechRnaSeqRawSignalRep1K562CellLongpolyaBb12x75.txt.gz	12/20/09
Caltech Rep1 NHEK PAP BB2R 2x75	wgEncodeCaltechRnaSeqRawSignalRep1NhekCellPapBb2R2x75.txt.gz	06/15/10
Caltech Rep2 HepG2 PAP BB2R 2x75	wgEncodeCaltechRnaSeqRawSignalRep2Hepg2CellPapBb2R2x75.txt.gz	06/14/10
DNase I Hypersensitivity Clusters	wgEncodeRegDnaseClustered.txt.gz	08/15/10
TFBS Clustered ChIP-seq	wgEncodeRegTfbsClustered.txt.gz	08/15/10
PhyloP 28–Way Base Cons	phyloP28way.txt.gz	11/30/08
PhyloP 28-Way Base Cons Plac Mammal	phyloP28wayPlacMammal.txt.gz	11/30/08
PhyloP 44-Way Base Cons	phyloP44wayAll.txt.gz	02/02/09
PhyloP 44-Way Base Cons Plac Mammal	phyloP44wayPlacMammal.txt.gz	02/02/09
PhyloP 44-Way Base Cons Primate	phyloP44wayPrimate.txt.gz	02/02/09
ORegAnno	oreganno.txt.gz	07/31/08
DGV Indel[‡]	indel.hg18.v10.nov.2010.txt	11/10
DGV Variation[‡]	variation.hg18.v10.nov.2010.txt	11/10
PolyPhen–2	pph2-full.hg18.txt	06/26/13
RegulomeDB	RegulomeDB.dbSNP132.Category[1-7].txt	06/25/13

All files downloaded from `http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database`, except as noted. [†]`http://hapmap.ncbi.nlm.nih.gov/downloads/ld_data/2009-04_rel27` [‡]`http://projects.tcag.ca/variation/downloads` [◇]For UCSC, date stamp is for associated SQL file, e.g. gwasCatalog.sql.

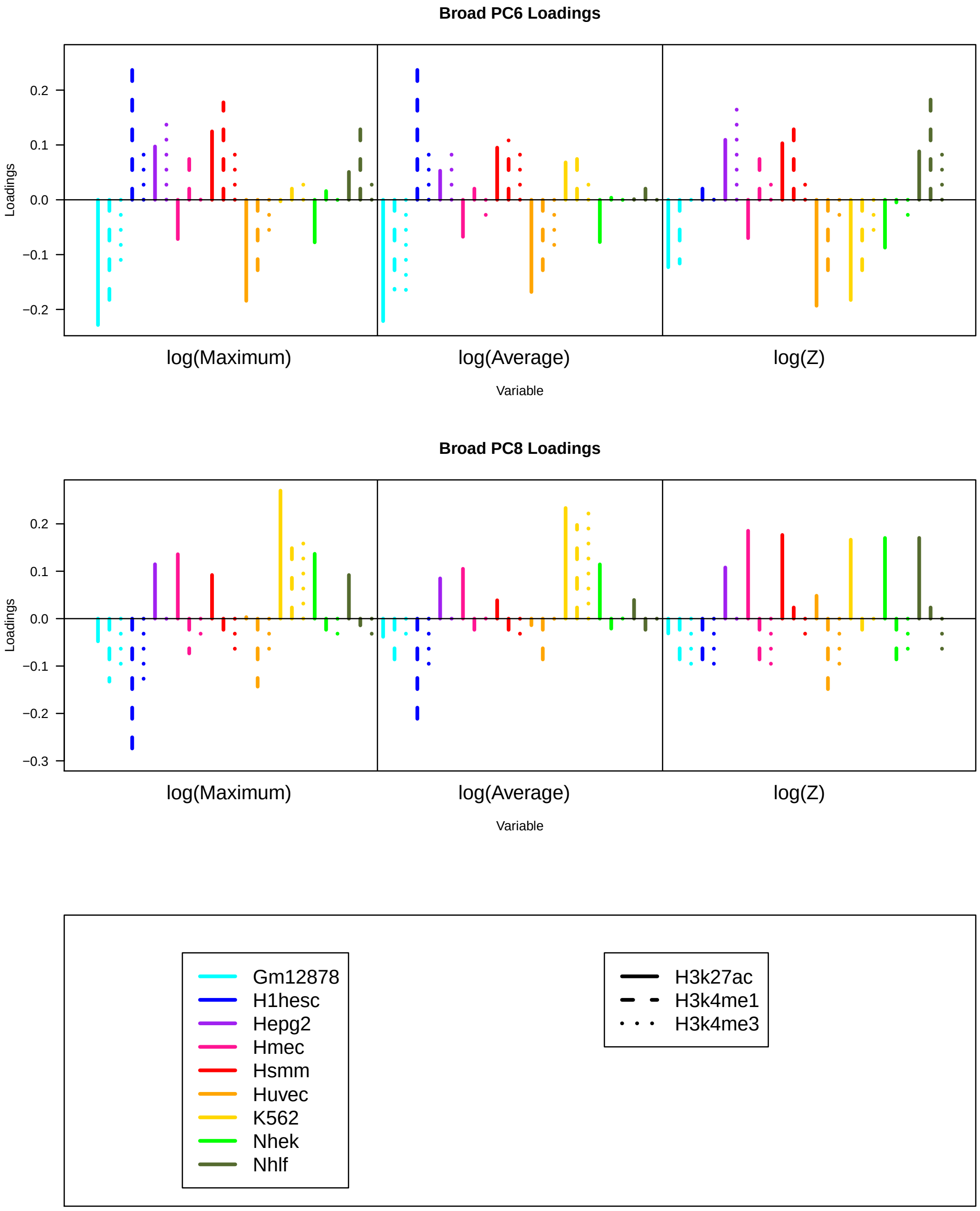
Supplemental Table 1: Summary of data sources.



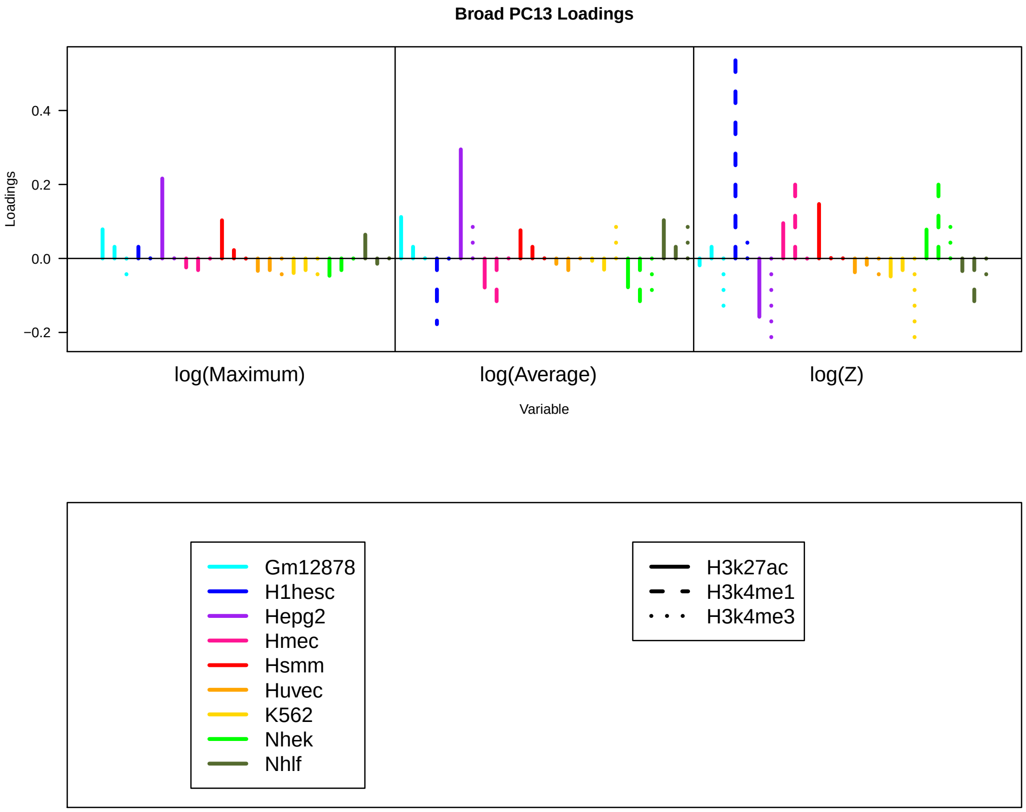
Supplemental Figure 1: Plot of the loadings (weights in the linear combinations) for the most highly associated Broad promoter/enhancer ChIP–seq principal components (PCs) as they depend on histone modification, cell line and summary statistic.



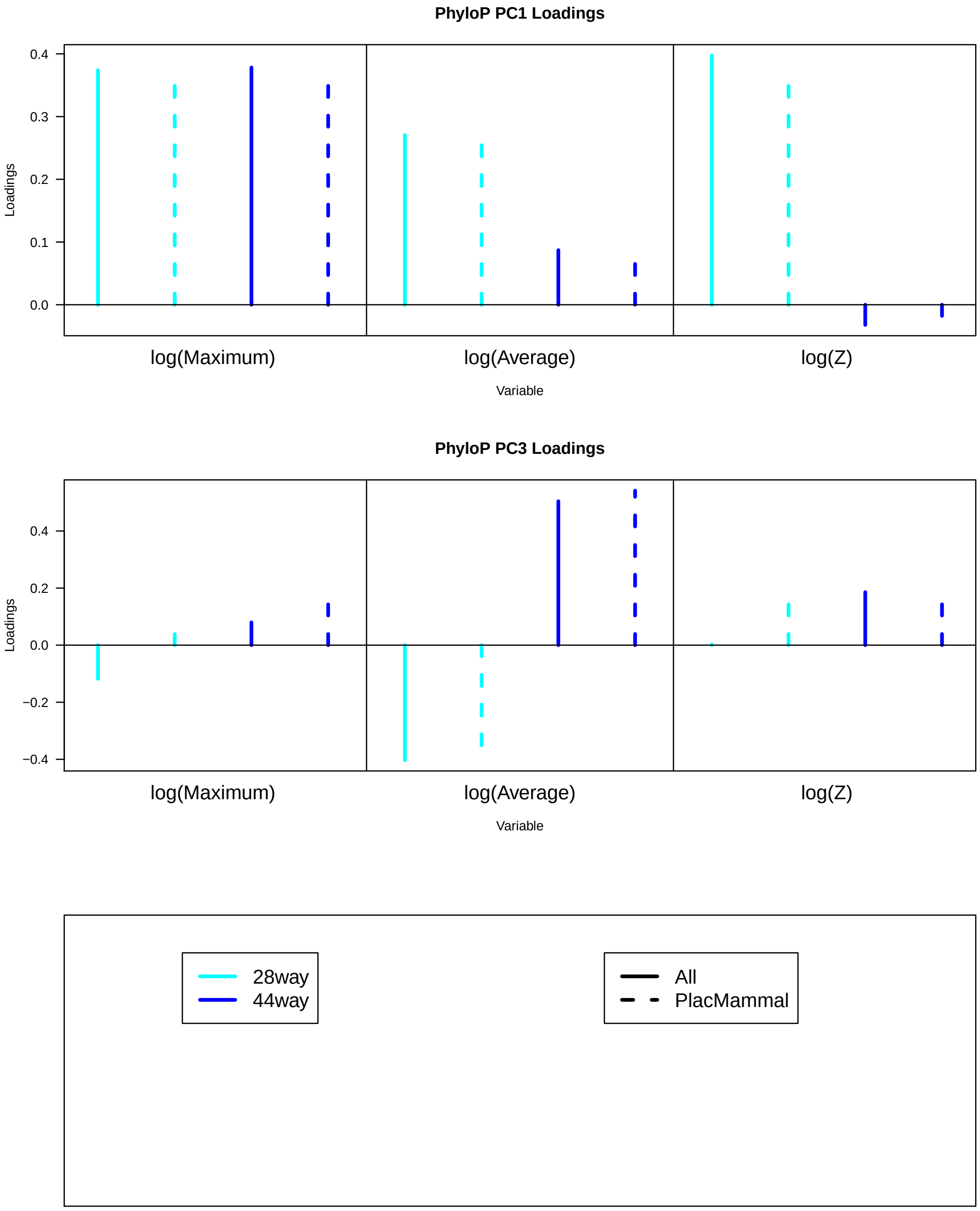
Supplemental Figure 1, Continued: Plot of the loadings (weights in the linear combinations) for the most highly associated Broad promoter/enhancer ChIP–seq principal components (PCs) as they depend on histone modification, cell line and summary statistic.



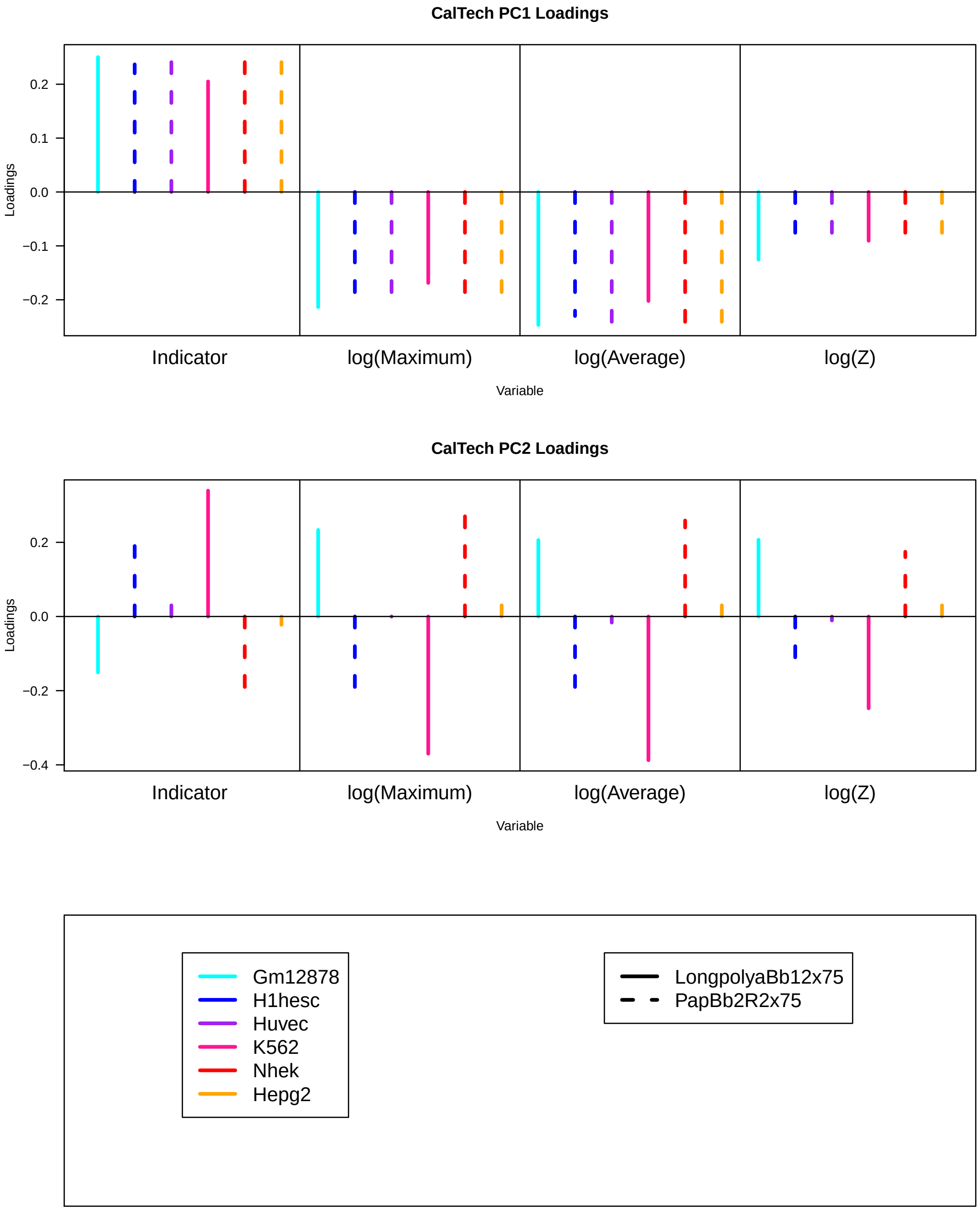
Supplemental Figure 1, Continued: Plot of the loadings (weights in the linear combinations) for the most highly associated Broad promoter/enhancer ChIP–seq principal components (PCs) as they depend on histone modification, cell line and summary statistic.



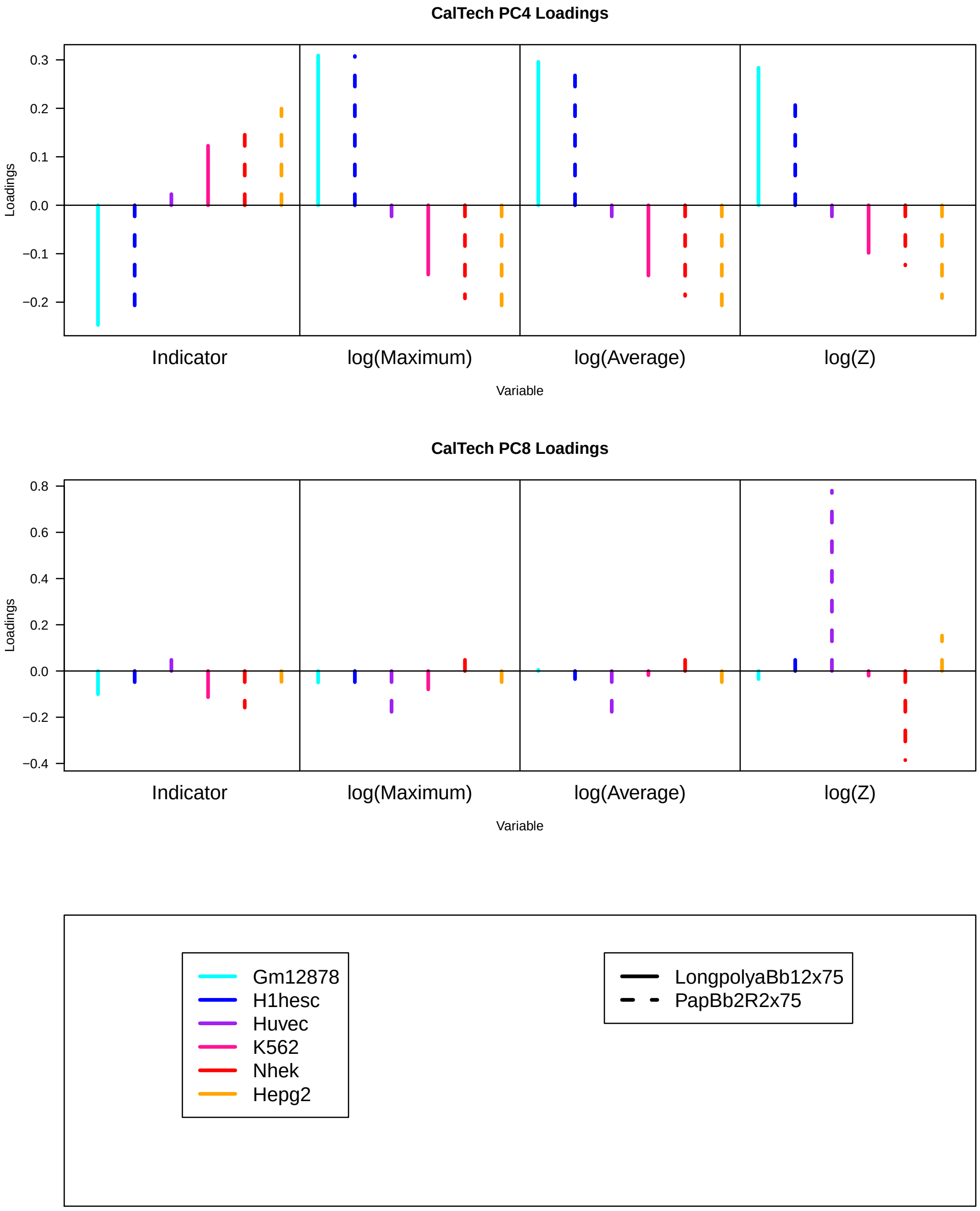
Supplemental Figure 1, Continued: Plot of the loadings (weights in the linear combinations) for the most highly associated Broad promoter/enhancer ChIP–seq principal components (PCs) as they depend on histone modification, cell line and summary statistic.



Supplemental Figure 2: Plot of the loadings (weights in the linear combinations) for the most highly associated sequence conservation principal components (PCs) as they depend on number of species, depth of alignment and summary statistica.



Supplemental Figure 3: Plot of the loadings (weights in the linear combinations) for the most highly associated CalTech RNA–seq principal components (PCs) as they depend on cell line and summary statistic.



Supplemental Figure 3, Continued: Plot of the loadings (weights in the linear combinations) for the most highly associated CalTech RNA–seq principal components (PCs) as they depend on cell line and summary statistic.